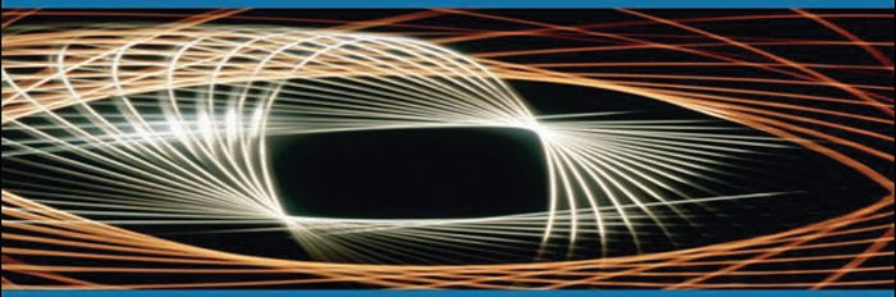
DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES



Visual Perception Through Video Imagery

Edited by Michel Dhome

ISTE

WILEY

This page intentionally left blank

Visual Perception through Video Imagery

This page intentionally left blank

Visual Perception through Video Imagery

Edited by
Michel Dhome

ISTE
WILEY

First published in France in 2003 by Hermes Science/Lavoisier entitled *Perception visuelle par imagerie video* © LAVOISIER, 2003

First published in Great Britain and the United States in 2009 by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Cataloging-in-Publication Data

[Perception visuelle par imagerie video. English] Visual perception through video imagery / edited by Michel Dhome.
p. cm.
Includes index.
ISBN 978-1-84821-016-5
1. Computer vision. 2. Visual perception. 3. Vision. I. Dhome, Michel.
TA1634.P4513 2007
006.3'7--dc22

2007028789

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN: 978-1-84821-016-5

Printed and bound in Great Britain by CPI Antony Rowe Ltd, Chippenham and Eastbourne.

Table of Contents

This page intentionally left blank

Introduction

Artificial vision with a main objective of automatic perception and interpretation of the universe observed by a system containing one or several cameras is a relatively new field of investigation. It leads to a surprisingly large range of problems, and most of these are not currently resolved in a reliable way. Although a general theory is not close to being reached, significant progress has been made recently, theoretically as well as methodologically.

In the visible world, images of luminance are the result of two physical processes: the first one is linked to reflectance properties of the surface of observed objects, while the second one corresponds to the projection of these same objects on the light sensitive plate of the sensor used. From a mathematical standpoint, in order to interpret the observed scene, we must solve an inverse problem, i.e. infer the surface geometry (3D) of objects present, from the purely 2D content of the image or from logged images.

This reputedly complex problem in the context of computer vision is solved by man with surprising ease. However, the human vision system operation is clearly not founded on a single concept. Examining the implemented processes during short or long distance vision is sufficient proof. In the first case, the existing disparity between left and right retinal images makes it possible for man to obtain indepth information by triangulation (*stereoscopy*) relating to its close environment which is vital in particular to manually capture objects. In the second case, when looking at long distance, or even more so when contemplating a picture, stereoscopy is obviously no help in interpreting the observed scene. Even under these conditions, however (*total lack of direct 3D information*), man is able to estimate the form and spatial position of objects he observes in the vast majority of cases. This requires mental processes, from 2D

information extracted from a luminance image, able to infer 3D information. These are based on the unconscious use of prior knowledge relating to the principle of retinal image composition and the form of 3D objects surrounding us. The surprising capabilities of the human vision system are because this knowledge is continuously enhanced from early childhood.

In the last few decades, researchers in the artificial vision community have attempted to develop perception systems that would work from data emanating from video cameras. This book presents a few tools emerging from recent advances in the field.

In Part 1, the reader will find three chapters dedicated to *calibration* or *self-calibration* of video sensors. Chapter 1 presents a finite estimation approach of the intrinsic parameters of a video camera, greatly inspired by the world of photogrammetry. It is based on the interpretation of images from a calibration test chart which is not generally known with great precision. Chapter 2 addresses the complex self-calibration problem from a series of matching points between different images from a single scene. The recommended method is based on an elegant and simple decomposition of Kruppa equations. Chapter 3 explores the self-calibration problem of cameras with *specific movements* making the implementation of simplified development of the main matrix possible.

Part 2 mainly involves the estimation of the *relative object/sensor position* by introducing prior knowledge (CAD object model). The reader will discover how to treat the localization problem of a rigid object observed by a monocular system. The formalism presented is then extended to understand cases as different as multi-ocular localization and hand-eye calibration, and research on the posture of articulated objects such as robotic arms.

Part 3 addresses volume information inference in two chapters. Chapter 5 discusses the *reconstruction* problem of a fixed scene observed by a multi-ocular system. The notions homo-log points, epipolar geometry, fundamental matrix, essential matrix and trifocal tensor are first introduced as well as different approaches for obtaining these entities. The problem of dense matching between image pairs is addressed before a few reconstruction examples are shown. Chapter 6 discusses the notion of *active dynamic vision*. It is the control of the path of a camera embedded in a robotic arm in order to reconstruct the surrounding scene. An underlying problem involves the definition of optimal movements necessary for the reconstruction of different primitives (points, straight lines, cylinders). Finally, perception strategies are

proposed in order to ensure a complete reconstruction taking inter-object blanking into consideration.

The *recognition of forms in images* is the heart of Part 4. Chapter 7 is dedicated to proposing tools for the identification in a database of the images containing visual elements identical to those contained in a request image; the differences of acquisition between images which could involve the point of view, conditions of illumination and global composition of the scene observed. The methods presented are based on a common principle: the use of quasi-invariants associated with local descriptors. These are tested against large image databases.

This page intentionally left blank

Part 1

This page intentionally left blank

Chapter 1

Calibration of Vision Sensors

1.1. Introduction

Calibration of a vision system consists of determining a mathematical relation existing between the three-dimensional (3D) coordinates of the reference markers of a scene and the two-dimensional (2D) coordinates of these same reference markers projected and detected in an image.

Determining this relation is an uphill task in vision, particularly for reconstruction, where it is necessary to infer 3D information from data extracted from the 2D image. In reality, the field of application is broader and calibration proves to be essential since it is necessary to establish a relation between an image and the 3D world: recognition and localization of objects, dimensional control of parts, and reconstruction of the environment for the navigation of a robot.

A complete analysis of the calibration of a vision system must take into account all the photometric, optic and electronic phenomena present in the image acquisition chain.

In general, a system that enables the calibration of a camera is made up of:

– a calibration test card (grid or standard object), generally consisting of reference markers from which 3D coordinates are accurately inferred in its local reference coordinate system;

Chapter written by Jean-Marc LAVEST and Gérard RIVES.

– an image acquisition system for the digitization and storage of test card images;

– an algorithm of correlating 2D reference marker detected in the images with their counterparts of the test card;

– a calculation algorithm of transformation matrix perspective of the camera, relating the calibration test card mark with that of the image.

Generally, the problem of calibrating CCD cameras unfolds in two ways, namely, the *geometric* study of calibration (calculation of the projection matrix) and a *radiometric* calibration (uniformity of brightness in an image). The first problem is widely covered, whereas the latter is less studied.

In this chapter, we will approach the problem of geometric calibration of a video sensor in a didactic way and will solve it first using linear methods and then using non-linear methods. Often, written works talk about *weak* or *strong* calibration; the distinction is at the level of an overall estimate of the projection matrix (weak case) or in each parameter estimation which forms this matrix (strong case). The continuation of this work will deal with the problem of *strong* calibration of vision sensors.

1.2. General formulation of the problem of calibration

The process of calibration is a monitored process often requiring the operator's attention. This is a static process, which is carried out offline, before using the camera for a precise visual task. Once the camera is calibrated, its parameters must remain fixed throughout its use. Each time we wish to modify the focus, the focal distance, and even the opening of lens, the camera will have to be recalibrated. Many works in the last 20 years have made it possible to obtain fairly complex and precise methods for assimilating a vision system in a metrological collection. The reader can refer to the following link for an exhaustive view of the major publications in the field: http://iris.usc.edu/Vision-Notes/bibliography.

1.2.1. *Formulation of the problem*

Let us consider an image acquisition system (Figure 1.1).

Calibration of the system consists of determining the transformation (R^3, R^2), which makes it possible to analytically express the process of image formation.

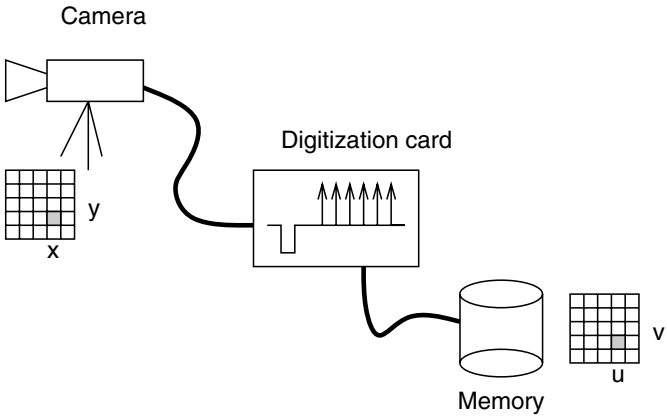


Figure 1.1. *System of acquiring images*

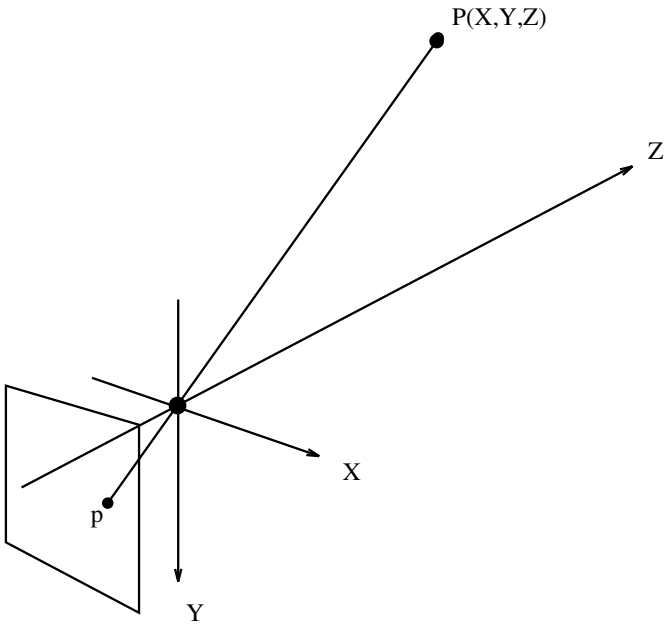


Figure 1.2. *Geometry of the image formation system*

1.2.1.1. *Modeling the camera and lens: pin-hole model*

Traditional geometric optics prefers the use of models with thick or thin lenses. The difficulty of expressing vergency constraints in a simpler way compels us to resort to the *pin-hole model* in which all rays pass through the same point (optical center). The photosensitive cell (image plane) is located at a distance f from this center and represents the focal distance or selection of the objective.

Let us note that the image obtained is normally inverted when compared to a naked eye view. To overcome this problem, we artificially place the image plane in front of the optical center (from the physics point of view, this artifice is carried out by reading the CCD matrix in such a manner as to obtain the inversion of the image).

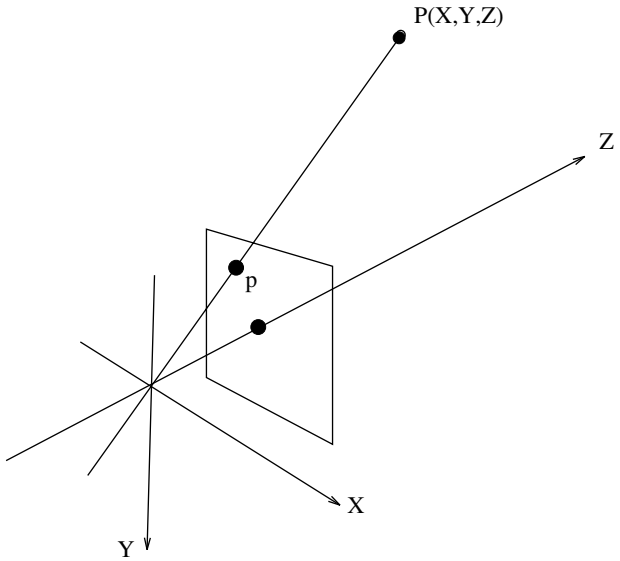


Figure 1.3. *Image plane defined in front of the optic center*

1.2.1.2. *Formation of images: perspective projection*

In works on this subject, we find several types of image projection: orthographic, scale orthographic, para-perspective and perspective. It is this last family of projection that will capture our attention due to it being the best suited in the physical reality of vision sensors.

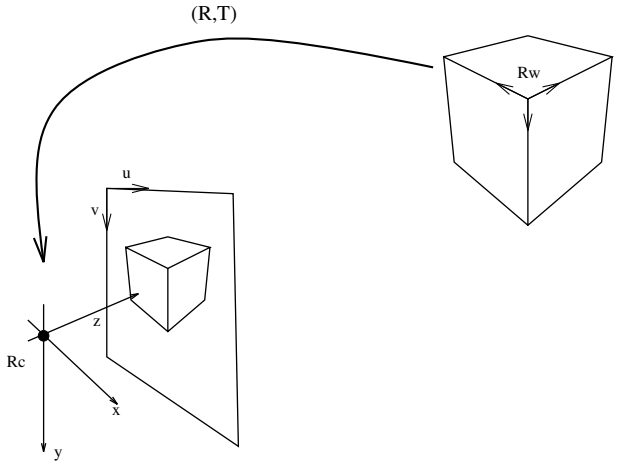


Figure 1.4. *Notation of the different reference marks*

In this section, we will use the following notations:

– R_c: camera reference ($-z$ axis meeting the optical axis);

– R_w: reference marker related to object modeling;

– $(0, u, v)$: image reference (considering the effect of digitization).

1.2.1.3. *Changing lens/camera reference point*

This first point change makes it possible to express the coordinates of the test card or the reference instrument (positioned in the surroundings) in the reference point relating to the camera. As we can see, this point change is expressed by a transformation made up of a rotation and a translation.

$$P_c = (M_1) \cdot P_w \tag{1.1}$$

$$P_c = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{1.2}$$

with

– P_w 3D coordinates of a test card reference point, indicated in the modeling reference;

– P_c coordinates of the same 3D point, indicated in the camera reference point.

The rotation and translation matrices ($R_{(3\times3)}$ and $T_{(3\times1)}$) are defined in the *camera* reference marker. Coupling of (R, T) in the same matrix notation makes it necessary to use a notation in homogenous coordinates.

1.2.1.4. *Changing of the camera/image point*

Changing from the camera point to the image point is related to perspective projection equations.

Traditionally these equations are of the form:

$$\begin{aligned} x &= X_c f/Z_c \\ y &= Y_c f/Z_c \\ z &= f \end{aligned} \tag{1.3}$$

In homogenous coordinates, the system is written as:

$$\begin{bmatrix} sx \\ sy \\ s \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} P_c \tag{1.4}$$

It must be noted that while changing (R^3, R^2), homogenous notations introduce a multiplicative factor s.

Demonstration

$$\begin{cases} sx = fX_c \\ sy = fY_c \\ s = Z_c \end{cases} \tag{1.5}$$

by substituting s:

$$\begin{cases} x = X_c f/Z_c \\ y = Y_c f/Z_c \end{cases} \tag{1.6}$$

1.2.1.5. *Changing of coordinates in the image plane*

Remark. It is necessary to include the step difference, according to coordinates x and y, relative on the one hand to the elementary pixel form on the CCD

matrix and on the other to the rhythmic sampling of the video signal.

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} 1/dx & 0 & u_0 \\ 0 & 1/dy & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} sx \\ sy \\ s \end{bmatrix} \tag{1.7}$$

This leads to traditional equations for changing the point (camera coordinate pixel to pixel coordinate):

$$\begin{cases} u = x/dx + u_0 \\ v = y/dy + v_0 \end{cases} \tag{1.8}$$

– (u_0, v_0) represent coordinates (in pixel) in the image, intersection of the optical axis and the image plane (origin of reference point change);

– (dx, dy) are respectively the dimensions according to x and y of an elementary pixel of the CCD matrix (see Figure 1.5).

1.2.2. *General expression*

The complete system of image formation is thus expressed by the following relation:

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} 1/dx & 0 & u_0 \\ 0 & 1/dy & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_x \\ R_{21} & R_{22} & R_{23} & T_y \\ R_{31} & R_{32} & R_{33} & T_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1.9}$$

where:

– (X_w, Y_w, Z_w) are the 3D coordinates of the calibration point pertaining to the reference instrument;

– (u, v) are the 2D pixel coordinates in the projection image of this point;

– (u_0, v_0, f, dx, dy) are called the *intrinsic parameters* of calibration. They belong to the system of acquisition;

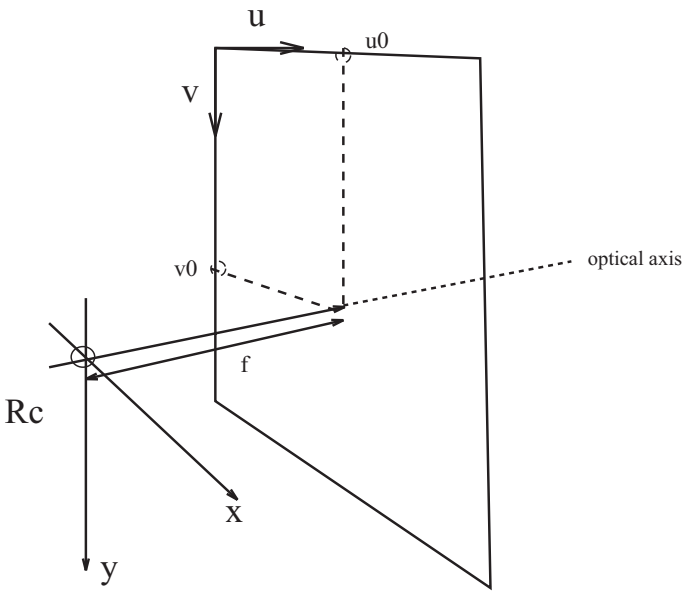


Figure 1.5. *Coordinate system on the CCD matrix*

– $(R_{(11,\ldots,33)}, T_{(x,y,z)})$ are called the *extrinsic parameters* of calibration. They provide the localization of the reference instrument in the camera reference point, while shooting a picture.

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = M_{\text{int}} M_{\text{ext}} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_{(3\times 4)} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1.10}$$

M is called the *calibrating matrix* of the system.

It contains 12 elements, which can be divided into:

– *5 intrinsic parameters* pertaining to the camera. Generally, we use the M_{int} matrix in the form:

$$\begin{bmatrix} f/dx & 0 & u_0 & 0 \\ 0 & f/dy & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{1.11}$$

By showing:

$$\left.\begin{aligned} f_x &= f/dx \\ f_y &= f/dy \end{aligned}\right\} \longrightarrow \left(f_x/f_y = dy/dx\right) \tag{1.12}$$

The dx/dy ratio represents the pixel ratio; at known dx (supplied by the camera maker), the intrinsic parameter estimation is reduced to the calculation of *4 parameters* (f_x, f_y, u_0, v_0).

– *12 independent extrinsic parameters* (9 for rotation $(R_{11} \ldots R_{33})$ and 3 for translation $(T_{(x,y,z)})$) which are independent of the camera.

In other words, there are a total of 16 parameters.

1.2.2.1. *General formulation of the problem of calibration*

To calibrate a vision system is to be capable of determining all the parameters intervening in the analytical expression of the image formation illustrated in (1.9).

1.3. Linear approach

Problem. Given n number of 3D-2D pairs (X_w, Y_w, Z_w; u, v) (expression (1.10)) between a test card and its image, determine the 16 parameters of image formation.

Resolving the problem of calibration by the linear method has the advantage of not requiring initial values of calibration parameters to find a solution. We will see that this is expensive in practice and that the performances of such an approach are limited. Faugeras and Toscani [FAU 87] proposed this method at the beginning of the 1980s. In the 1940s, a similar approach known as DLT (Direct Linear Transform) was introduced by the photogrammetric community.

1.3.1. *Principle*

Let the 3D-2D projection between a reference mark and its image be:

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1.13}$$

where $(m_{11} \ldots m_{34})$ are the 12 unknown elements of the system to be solved.

By substituting s, we obtain:

$$\begin{cases} u = \dfrac{m_{11}X_w + m_{12}Y_w + m_{13}Z_w + m_{14}}{m_{31}X_w + m_{32}Y_w + m_{33}Z_w + m_{34}} \\ v = \dfrac{m_{21}X_w + m_{22}Y_w + m_{23}Z_w + m_{24}}{m_{31}X_w + m_{32}Y_w + m_{33}Z_w + m_{34}} \end{cases} \tag{1.14}$$

using the following equations drawn from matrix expression (1.9):

$$\begin{aligned} m_{1(1,2,3)} &= \frac{f}{dx}R_{1(1,2,3)} + u_0 R_{3(1,2,3)} \\ m_{14} &= \frac{f}{dx}T_x + u_0 T_z \\ m_{2(1,2,3)} &= \frac{f}{dy}R_{2(1,2,3)} + v_0 R_{3(1,2,3)} \\ m_{24} &= \frac{f}{dy}T_y + v_0 T_z \\ m_{3(1,2,3)} &= R_{3(1,2,3)} \\ m_{34} &= T_z. \end{aligned}$$

It is possible to rewrite the system to solve (1.14) as:

$$\begin{bmatrix} & & & & & & & \cdot & & & & \\ & & & & & & & \cdot & & & & \\ & & & & & & & \cdot & & & & \\ X_w^i & Y_w^i & Z_w^i & 1 & 0 & 0 & 0 & 0 & -u^i X_w^i & -u^i Y_w^i & -u^i Z_w^i & -u^i \\ 0 & 0 & 0 & 0 & X_w^i & Y_w^i & Z_w^i & 1 & -v^i X_w^i & -v^i Y_w^i & -v^i Z_w^i & -v^i \\ & & & & & & & \cdot & & & & \\ & & & & & & & \cdot & & & & \\ & & & & & & & \cdot & & & & \end{bmatrix}$$

$$\times \begin{bmatrix} m_{11} \\ m_{12} \\ \cdot \\ \cdot \\ \cdot \\ m_{32} \\ m_{33} \end{bmatrix} = 0 \tag{1.15}$$

Index i represents the pairing between the 3D standard reference marker and the 2D reference marker detected in the image. Each pair paves the way to write 2 equations, the minimum number necessary for solving the problem thus being *6 pairs*.

Solving an overdetermined system $(AX = 0)$ amounts to searching for the eigenvector associated with the smallest eigenvalue of A. Generally, we use the SVD (Singular Value Decomposition [PRE 92]) algorithm.

By normalizing $m_{3(1,2,3)}$, we obtain the third vector of rotation matrix $R_{3(1,2,3)}$:

$$T_z = m_{34} \tag{1.16}$$

$$\begin{pmatrix} m_{21} \\ m_{22} \\ m_{23} \end{pmatrix} \cdot \begin{pmatrix} m_{31} \\ m_{32} \\ m_{33} \end{pmatrix} = v_0 \tag{1.17}$$

$$\begin{pmatrix} m_{11} \\ m_{12} \\ m_{13} \end{pmatrix} \cdot \begin{pmatrix} m_{31} \\ m_{32} \\ m_{33} \end{pmatrix} = u_0 \tag{1.18}$$

$$\left\| m_{2(1,2,3)} - v_0 R_{3(1,2,3)} \right\| = f/dy \tag{1.19}$$

$$\left\| m_{1(1,2,3)} - u_0 R_{3(1,2,3)} \right\| = f/dx \tag{1.20}$$

$$\left(m_{2(1,2,3)} - v_0 R_{3(1,2,3)} \right) / (f/dy) = R_{2(1,2,3)} \tag{1.21}$$

$$\left(m_{1(1,2,3)} - u_0 R_{3(1,2,3)} \right) / (f/dx) = R_{1(1,2,3)} \tag{1.22}$$

Thus, the 16 parameters of image formation are completely determined.

1.3.2. *Notes and comments*

This method is very easy to implement. Solving a linear system is inexpensive in computing times. However, the results obtained are not very stable. The stability of a method refers to its aptitude to give similar results (on the intrinsic parameters) for different behaviors of the test card.

Let us note that the preceding division allows us to foresee a problem on the estimation level of rotation R. Indeed, we are never certain that:

$$R_{1(1,2,3)} \wedge R_{2(1,2,3)} = R_{3(1,2,3)} \tag{1.23}$$

In other words, it is never certain that the rotation matrix R is really conditioned like a rotation matrix. Therefore, a palliative solution consists of modifying the perspective projection matrix to add an extra term to it, which can be interpreted like the non-orthogonality of the optical axis as compared to the CCD sensor, which makes it possible to ensure the orthogonality of the three rotation vectors.

In fact, in the linear approach, the rotation matrix is not parametrized in a satisfactory manner. It will be advisable to use its division according to Euler's angles (α, β, γ), respectively, around the axes (x, y, z) of the camera point. In these conditions, the results are that the system to be solved is no longer *linear*. Moreover, no phenomenon of optical distortion is taken into account in the process of image formation.

Nevertheless, this method obtains a very good initial estimation in the process of non-linear optimization.

1.4. Non-linear photogrammetric approach

In this section, we approach the problem of calibration of CCD cameras based on the formalism used in photogrammetry. The projection model used for the process of image formation refers to the pin-hole optical model, which is the approximation of the optical model of thin lens. This approach is different from the preceding one due to a precise modeling of the optical distortion phenomena caused on the surface of the lenses and also due to the implementation of a non-linear optimization process minimizing a criterion of reprojection of the reference marker in images which is expressed in pixels.

The notation conventions in photogrammetry are slightly different from those used in artificial vision. Also, traditionally, the data are no longer expressed in the camera reference marker but in the reference frame of the world, which remains fixed irrespective of the position of the camera. We will thus transform the usual photo-grammetric notations in such a manner to determine the formalism belonging to our community.

Let us consider the following notations:

– $R_W - XYZ$ is a direct 3D point. It is the reference frame of the world, which will also be used as the modeling base of the object.

– $o - uv$ is the 2D image reference frame, whose origin is located at the top left corner of the image.

– $R_C - xyz$ is the 3D reference frame of the camera, whose origin is at the optic center c and whose z axis is confused with the optic axis. x, y are respectively parallel to $o - uv$.

The intrinsic parameters in the sensor to be determined are: the principal point $o - u_0v_0$, the focal distance f, the pixel size of the CCD matrix (dx, dy) or their ratio and, finally, the optical distortion parameters introduced by the camera lens.

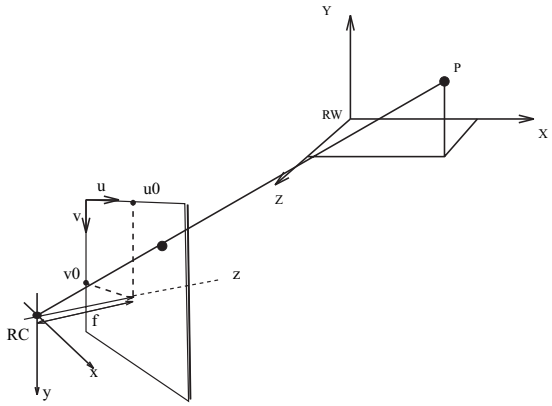


Figure 1.6. *Pin-hole model, image geometry and coordinate systems*

The extrinsic parameters are the rotation matrix $\mathbf{R}$ as well as the translation vector $\mathbf{T}$ between $R_c - xyz$ and $R_w - XYZ$.

The method described in this chapter strictly follows the least squares approach; we will try to minimize the *measuring errors* in the image, representing the difference between a point detected in the image and a projected 3D point in the corresponding test card.

1.4.1. *Mathematic model*

Let there be a perspective projection between a 2D image and a 3D object (assumed to be *pin-hole*). The relation between a point and its projection in the image is described by the following expression:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \lambda_i \left[\mathbf{R} \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} + \mathbf{T} \right] \tag{1.24}$$

where:

– (x_i, y_i, z_i) is an image point defined in the camera reference point (see Figure 1.6 with $\bar{z}_i \equiv f$, i.e., the focal distance of the camera);

– λ_i is a scaling factor introduced while changing from R^3 to R^2;

– (X_i, Y_i, Z_i) are the coordinates of the test card point defined in the reference frame of the surrounding $W - XYZ$;

– (T_x, T_y, T_z) is the translation vector;

– $\mathbf{R}$ is the rotation matrix expressed in the camera point and parametrized according to Euler's three angles: α rotation around x axis, β around y axis, and γ around z axis:

$$\mathbf{R} = \begin{pmatrix} \cos\gamma\cos\beta & \cos\gamma\sin\beta\sin\alpha - \sin\gamma\cos\alpha & \cos\gamma\sin\beta\cos\alpha + \sin\gamma\sin\alpha \\ \sin\gamma\cos\beta & \cos\gamma\cos\alpha + \sin\gamma\sin\beta\sin\alpha & \sin\gamma\sin\beta\cos\alpha - \cos\gamma\sin\alpha \\ -\sin\beta & \cos\beta\sin\alpha & \cos\beta\cos\alpha \end{pmatrix} \tag{1.25}$$

By eliminating λ_i in (1.24) and by removing the index i, we obtain the following expressions called the *colinearity equations* in photogrammetry:

$$\left.\begin{aligned} x &= f\frac{r_{11}X + r_{12}Y + r_{13}Z + T_x}{r_{31}X + r_{32}Y + r_{33}Z + T_z} \\ y &= f\frac{r_{21}X + r_{22}Y + r_{23}Z + T_y}{r_{31}X + r_{32}Y + r_{33}Z + T_z} \end{aligned}\right\} \tag{1.26}$$

If we express (x, y) in the pixel coordinate system of the image, we obtain:

$$\left.\begin{aligned} x &= (u + e_x - u_0)dx - do_x \\ y &= (v + e_y - v_0)dy - do_y \end{aligned}\right\} \tag{1.27}$$

In this expression, e_x and e_y are respectively the measuring errors according to the x and y coordinates, (i.e., the corrections to be made in the measurements to ensure that there is a perfect correspondence with the data resulting from the projection function). do_x, do_y are the components of optical distortion, which are divided into two parts: *radial and tangential* distortions (i.e., $do_x = do_{xr} + do_{xt}$ and $do_y = do_{yr} + do_{yt}$). Here, we

introduce two forms that are commonly used in photogrammetry [AME 84]:

$$\left.\begin{aligned} do_{xr} &= \big(u-u_0\big)dx\big(a_1r^2+a_2r^4+a_3r^6\big)\\ do_{yr} &= \big(v-v_0\big)dy\big(a_1r^2+a_2r^4+a_3r^6\big)\end{aligned}\right\} \tag{1.28}$$

$$\left.\begin{aligned} do_{xt} &= p_1\big[r^2+2\big(u-u_0\big)^2dx^2\big]+2p_2\big(u-u_0\big)dx\big(v-v_0\big)dy\\ do_{yt} &= p_2\big[r^2+2\big(v-v_0\big)^2dy^2\big]+2p_1\big(u-u_0\big)dx\big(v-v_0\big)dy\end{aligned}\right\} \tag{1.29}$$

where in expressions (1.27), (1.28) and (1.29):

– u, v are the image coordinates in the image reference point;

– u_0, v_0 are the coordinates of the principal point in the referential image;

– a_1, a_2, a_3 are the polynomial coefficients that model the radial distortion;

– p_1, p_2, p_3 are the polynomial coefficients that model the tangential distortion;

– dx, dy represent the scale factors of the elementary pixel form;

– parameter $r = \sqrt{(u-u_0)^2dx^2+(v-v_0)^2dy^2}$, is the radial distance from the principal point. Since r can take significant values (based on the size of the image), r^4 and r^6 sometimes become enormous; expression (1.28) can then lead to a numerical instability during the estimation of the different parameters. A means of circumventing this difficulty is to rewrite this expression in the following way:

$$\left.\begin{aligned} do_{xr} &= \big(u-u_0\big)dx\big(a_1\big(r^2-r_0^2\big)+a_2\big(r^4-r_0^4\big)+a_3\big(r^6-r_0^6\big)\big)\\ do_{yr} &= \big(v-v_0\big)dy\big(a_1\big(r^2-r_0^2\big)+a_2\big(r^4-r_0^4\big)+a_3\big(r^6-r_0^6\big)\big)\end{aligned}\right\} \tag{1.30}$$

which assumes that the distortion is zero for a radial distance r_0.

By substituting (1.27), (1.28) and (1.29) in (1.26), and by showing $f_x = \frac{f}{d_x}$ and $f_y = \frac{f}{d_y}$, we obtain the following system:

$$\left.\begin{aligned} u+e_x &= u_0+do_{xr}+do_{xt}+f_x\frac{r_{11}X+r_{12}Y+r_{13}Z+T_x}{r_{31}X+r_{32}Y+r_{33}Z+T_z} = P(\mathbf{\Phi})\\ v+e_y &= v_0+\big(do_{yr}+do_{yt}\big)\frac{f_x}{f_y}+f_y\frac{r_{21}X+r_{22}Y+r_{23}Z+T_y}{r_{31}X+r_{32}Y+r_{33}Z+T_z} = Q(\mathbf{\Phi})\end{aligned}\right\} \tag{1.31}$$

where in (1.31), $\mathbf{\Phi}$ is a vector of 15 parameters, i.e.:

$$\mathbf{\Phi} = \left[u_0, v_0, a_1, a_2, a_3, p_1, p_2, f_x, f_y, T_x, T_y, T_z, \alpha, \beta, \gamma\right]^T$$

1.4.2. *Solving the problem*

Let us again consider the colinearity equations defined in expression (1.31):

$$\begin{pmatrix} e_x \\ e_y \end{pmatrix} = \begin{pmatrix} P(\Phi) - u \\ Q(\Phi) - v \end{pmatrix} = V(\Phi) \tag{1.32}$$

Presently, the problem is to determine the value of $\mathbf{\Phi}$, which reduces:

$$S = \sum_{i=1}^{n} \left(e_{x_i}^2 + e_{y_i}^2\right).$$

In (1.32), $P(\mathbf{\Phi})$ and $Q(\mathbf{\Phi})$ are non-linear functions of $\mathbf{\Phi}$ and therefore minimization is a non-linear optimization problem.

A method to solve this problem is to make a linearization of (1.32) from an initial value $\mathbf{\Phi}_0$ (generally provided by the results of the linear resolution of the problem of calibration described in section 1.3) and calculate a correction $\mathbf{\Delta\Phi}$. Then, we add $\mathbf{\Delta\Phi}$ to $\mathbf{\Phi}_0$, which becomes the new initial value: the process must be repeated until the convergence of the system is obtained.

Let there be n 3D reference markers and their corresponding reference markers in the image; we can write the $2 \times n$ linearized equation system in matrix form:

$$\mathbf{V}(\mathbf{\Phi}) = \mathbf{V}(\mathbf{\Phi_0}) - \sum_{i=1}^{15} \frac{\partial \mathbf{V}}{\partial \mathbf{\Phi_i}} \mathbf{\Delta\Phi}_i \tag{1.33}$$

$$\mathbf{V} = \mathbf{L} - \mathbf{A\Delta\Phi} \tag{1.34}$$

Thus, $\mathbf{L}$ represents the common value of the criterion and $\mathbf{A}$ represents the Jacobian matrix of the system, around the current vector $\mathbf{\Phi_0}$.

Let the weighting matrix of measures be[1] $\mathbf{W}$ then the resolution in the context of least squares of (1.34) amounts to estimating:

$$\min_{\mathbf{\Delta\Phi}\in\Re^{15}} \left(\mathbf{V}^T\mathbf{W}\mathbf{V}\right) \tag{1.35}$$

The solution of (1.35) is given by:

$$\mathbf{\Delta\Phi} = \left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{W}\mathbf{L}\right) \tag{1.36}$$

Demonstration. Let $\mathbf{\Omega} = \mathbf{V}^T\mathbf{W}\mathbf{V}$. In the solution of the system, we have, $\frac{\partial\mathbf{\Omega}}{\partial\mathbf{\Phi}} = 0$, i.e.:

$$\frac{\partial\mathbf{\Omega}}{\partial\mathbf{\Phi}} = 2\mathbf{V}^T\mathbf{W}\frac{\partial\mathbf{V}}{\partial\mathbf{\Phi}} = -2\mathbf{V}^T\mathbf{W}\mathbf{A} = 0, \quad \Longrightarrow \mathbf{A}^T\mathbf{W}\mathbf{V} = 0.$$

By replacing $\mathbf{V}$ from its expression in (1.34), the above equation becomes:

$$\mathbf{A}^T\mathbf{W}\left(\mathbf{L} - \mathbf{A}\mathbf{\Delta\Phi}\right) = \mathbf{A}^T\mathbf{W}\mathbf{L} - \mathbf{A}^T\mathbf{W}\mathbf{A}\mathbf{\Delta\Phi} = 0$$

which leads to the solution of $\mathbf{\Delta\Phi}$:

$$\mathbf{\Delta\Phi} = \left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{W}\mathbf{L}\right).$$

1.4.3. *Multi-image calibration*

Error in measurements is one of the main causes for obtaining bad results in calibration (which are in the image as well as on the test card). To mitigate this problem, it is possible to combine, in the same system, several images coming from the same camera but for different spatial positions (rotation and/or translation). In this case, the intrinsic parameters of the sensor are the same for all images and calibration estimates the following vector of parameters:

$$\begin{aligned}\mathbf{\Phi}_{9+6m} = \big[&x_0, y_0, a_1, a_2, a_3, p_1, p_2, f_x, f_y, T_x^1, T_y^1, T_z^1,\\ &\alpha^1, \beta^1, \gamma^1, \ldots, T_y^m, T_z^m, \alpha^m, \beta^m, \gamma^m\big]^T\end{aligned}$$

1. Generally, $\mathbf{W}$ is a diagonal matrix $2n \times 2n$. If all measurements are made with the same precision and if there is no correlation between the parameters, then $\mathbf{W}$ is the identity matrix: i.e., $\mathbf{W} = \mathbf{I}$.

The matrix $\mathbf{A}$ of (1.34) is therefore of the form:

$$\mathbf{A}_{2mn\times(9+6m)} = \left[\begin{array}{c|ccc} & \mathbf{A}^{11}_{2n\times 6} & & 0 \\ & & \ddots & \\ \mathbf{A}^{1}_{2mn\times 9} & & \mathbf{A}^{ii}_{2n\times 6} & \\ & & & \ddots \\ & 0 & & \mathbf{A}^{mm}_{2n\times 6} \end{array}\right]$$

where m is the number of images and n is the number of reference markers per image. The total number of equations is $(2mn)$ and the total number of parameters is $(9 + 6m)$. Redundancy r is given by: $r = 2mn - 9 - 6m$ and it is much more significant than in the case of calibration with a single image. The multi-image approach, from an experimental point of view, leads to quality results and ensures a greater reproducibility of the estimation of the internal parameters of the sensor.

1.4.4. *Self-calibration by bundle adjustment*

1.4.4.1. *Redefinition of the problem*

The main idea comes from the following observation: quality test cards of calibration are difficult to achieve: the mechanical stability of the set for a precision < 0.05 mm is obtained only with specific materials and a precise measurement of 3D reference markers used is expensive. Moreover, the variability of the angular fields based on applications leads to the usage of different test cards adapted to experimental conditions.

Self-calibration by bundle adjustment is a multi-image approach, which jointly allows a re-estimation of the three-dimensional structure of the test card and an estimation of the intrinsic and extrinsic traditional parameters of the sensor. In other words, we calibrate the camera and reconstruct the test card simultaneously.

Let the colinearity equations be:

$$\left.\begin{aligned} u + e_x &= u_0 + \big(do_{xr} + do_{xt}\big)/dx \\ &\quad + \left(\frac{f}{dx}\right)\frac{r_{11}X + r_{12}Y + r_{13}Z + T_x}{r_{31}X + r_{32}Y + r_{33}Z + T_z} = P(\mathbf{\Phi}) \\ v + e_y &= v_0 + \big(do_{yr} + do_{yt}\big)/dy \\ &\quad + \left(\frac{f}{dy}\right)\frac{r_{21}X + r_{22}Y + r_{23}Z + T_y}{r_{31}X + r_{32}Y + r_{33}Z + T_z} = Q(\mathbf{\Phi}) \end{aligned}\right\} \tag{1.37}$$

To simplify writing, we removed the indices, without loss of generality, but the colinearity expressions go well with all the reference markers (X_i, Y_i, Z_i) of n reference markers of the test card projecting itself in (v_j, v_j) in m images.

If we wish to calibrate the sensor and to calculate the coordinates of the reference marker of the test card, then, the parameter vectors to be estimated take the following form:

$$\begin{aligned}\mathbf{\Phi}_{9+6m+3*n} = \big[&x_0, y_0, a_1, a_2, a_3, p_1, p_2, f_x, f_y,\\ &\mathbf{X^1}, \mathbf{Y^1}, \mathbf{Z^1}, \ldots, \mathbf{X^n}, \mathbf{Y^n}, \mathbf{Z^n},\\ &T_x^1, T_y^1, T_z^1, \alpha^1, \beta^1, \gamma^1, \ldots, T_y^m, T_z^m, \alpha^m, \beta^m, \gamma^m\big]^T\end{aligned}$$

Problem. Find $\mathbf{\Phi}$, which reduces $\sum_{j=1}^{m}\sum_{i=1}^{n}(e_{x_{ij}}^2 + e_{y_{ij}}^2)$.

1.4.4.2. *Estimation of redundancy*

Number of unknown elements

$$\begin{aligned}9 \text{ (intrinsic parameters)} + 3^*n \text{ (test card reference marker)}\\ + 6^*m \text{ (extrinsic parameters)}.\end{aligned}$$

Number of equations

$$2^*m^*n.$$

The redundancy of the system $r = 2^*m^*n - (9 + 3^*n + 6^*m)$ is ensured without much difficulty. For a test card of 11 reference markers, for example, a minimum of 3 images allows overdetermination of the system.

1.4.4.3. *Solution for a near scale factor*

Intrinsic parameters. In a traditional way, the matrix of the intrinsic parameters is always determined in a near scale factor. The usual introduction of the factor $(dx = 1)$ makes it possible to fix the set of parameters in an arbitrary unit of pixels.

Extrinsic parameters. From the time we estimate, within the process, the three-dimensional coordinates of the reference marker of the test card, the extrinsic geometry of the system is also fixed to a near scale factor. Indeed,

it is always possible to find a more voluminous test card observed from afar, which would strictly give the same image.

This metric loss is not of much importance for the calibration of a simple camera where only the intrinsic parameters represent an interest for the user. Nevertheless, to facilitate convergence, two reference markers of the test card will not be reconstructed and will impose the metric system. We will also impose that one of the coordinates of an unspecified point among n remains fixed to solidify the total extrinsic geometry of the reconstructed scene.

1.4.4.4. *Initial conditions*

It is obvious that the optimization of such a non-linear system requires initial conditions in the field of convergence. In the experimental part, we will show that this constraint is resolved without difficulties from the time of observing the test card under different orientation behaviors. This amounts to ensuring the significant angles of triangulation for the estimation of test card reference marker.

The values of each term of the initial vector of calibration can be provided with the help of the linear approach of calibration or more simply from the data creator for intrinsic parameters and from a localization algorithm such as that of Dementhon [DEM 95] for extrinsic parameters.

1.4.5. *Precision calculation*

From an estimate in the context of least squares (1.34) and (1.35), it is possible to calculate an estimate of the residual vector $\mathbf{V}$:

$$\hat{\mathbf{V}} = \left[\mathbf{A}\left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{W} - \mathbf{I}\right]\mathbf{L} \tag{1.38}$$

as well as the estimate of the *standard error of unit weight*, which represents an estimate *a posteriori* of σ_0 (scalar) noise on the reference marker detected in the image.

$$\hat{\sigma}_0^2 = \frac{\mathbf{V}^T\mathbf{W}\mathbf{V}}{N - P} \tag{1.39}$$

where P is the total number of estimated parameters. The value of the *covariance matrix* associated with $\mathbf{\Phi}$ parameters is given by:

$$\mathbf{C}_\Phi = \left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1} \tag{1.40}$$

Thus, for each parameter ϕ_i, it is possible to calculate its detection precision (standard deviation):

$$\hat{\sigma}_{\phi_i}^2 = \hat{\sigma}_0^2 c_{ii} \tag{1.41}$$

A method for measuring the reliability (quality) of an adjustment in the context of least squares is to calculate the relative redundancy of the system [TOR 81], i.e.:

$$q = \mathrm{tr}\left[\mathbf{I} - \mathbf{A}\left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^T\right] = \frac{r}{N} \tag{1.42}$$

where in (1.39) and (1.41), r represents redundancy, N is the total number of measurement equations and c_{ii} is the ith diagonal element of covariance matrix $\mathbf{C}_\Phi$. As we can observe, the relative redundancy for a multi-image calibration is much more significant than that for a simple calibration. It results in a greater reliability in the multi-image calibration results, $q_{\text{multi}} = \frac{r}{2mn} = 1 - \frac{9+6m}{2mn}$, when compared to simple-image resolution, $q_{\text{single}} = 1 - \frac{15}{2n}$. If $m > 1$, then $q_{\text{multi}} > q_{\text{single}}$.

1.5. Results of experimentation

1.5.1. *Bundle adjustment for a traditional lens*

This first part presents an example of calibration on a 1/2" camera equipped with a 10 mm lens.

1.5.1.1. *Initial and experimental conditions*

For this experiment, the polynomial distortion coefficients $(a_1, a_2, a_3, p_1, p_2)$ are initialized at zero. The test card $(X_i, Y_i, Z_i,\ i \in [1, n])$ is roughly measured and point coordinates are truncated to several centimeters; let us note that the multi-image approach makes it possible to use planar test cards, which largely facilitates their manufacture and in obtaining a rough model. The relative initial camera/test card positions $(R_j, T_j,\ j \in [1, m])$ are estimated by Dementhon's algorithm [DEM 95] applied to plane objects.

The position of the principal point (u_0, v_0) is placed in the middle of the image and the focal distance (f_x, f_y) is roughly estimated by the knowledge of the focal distance of the lens (10 mm, for example) and the elementary pixel size of the CCD matrix (ranging from 9 to 15 μm according to sensors).

The series of analyzed images must obligatorily form a beam of converging views in order to enforce the angles of triangulation to obtain reconstruction of the test card. Finally, the approach is based on an accurate detection of test card reference marker (see [LAV 98]).

1.5.1.2. *Sequence of classic images*

The images presented in Figure 1.7 show a sample of 15 photoshots taken to calibrate a camera equipped with a 10 mm lens. The test card is obtained from a plane plate and is equipped with retroreflective chips. The photographic device integrates annular high frequency lighting, which makes it possible to obtain good quality images of the chips used, irrespective of the observation angle of the object.

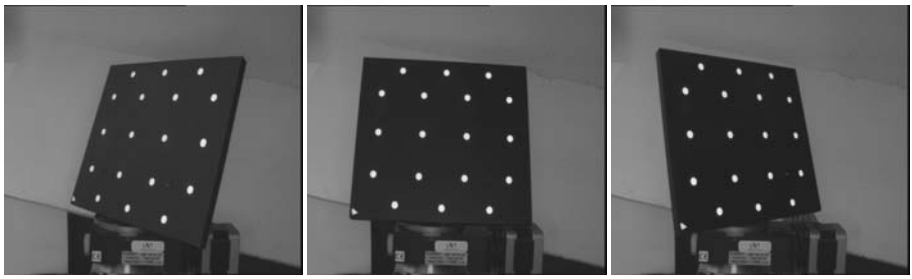


Figure 1.7. *Partial sequence of 15 photos taken for self-calibration (*768×576 *pixels)*

In order to highlight the convergence of the algorithm, we deliberately chose to leave the solution for the internal parameters of focal distance (1,500 pixels instead of 1,000), i.e., with an error of 50%. In some iterations, the algorithm is stabilized around a minimum, which leads to residues of about 0.025 pixels on each coordinate (see Table 1.1).

The curves in Figure 1.8 show the convergence of fx, u_0 and v_0; the minimum expression (1.35) will be attained at the end of 12 iterations.

Remarks:

– The convergence strategy does not pose a particular problem for this series of data, but we will see that this is not always the case for lenses displaying a strong radial distortion. For these images, all parameters (intrinsic, extrinsic and test card coordinates) are estimated from the first iteration, i.e., on the whole for 15 images and a test card with 18 reference markers ($11 + 15 * 6 + 3 * 18 = 155$ parameters):

Camera	: JAI M10
Lens	: 10 mm Ernitec
Digitization card	: Silicon Graphics
Algorithm	: Bundle adjustment

Number of images	: 15
Number of measurements	: 270

Residues ex ave and e-type (pix)	9.711e-05	2.555e-02
Residues ey ave and e-type (pix) (pix)	2.480e-06	2.367e-02
σ_0 (pix)		2.917e-02

	value	σ
fx(pix)	977.11	2.94e-01
fy(pix)	977.50	2.86e-01
u0(pix)	365.01	2.54e-01
v0(pix)	267.64	3.75e-01
a1	$+1.955$e-07	3.11e-09
a2	$+6.427$e-14	4.82e-14
a3	-1.437e-18	2.17e-19
p1	$+1.154$e-06	5.42e-08
p2	-6.077e-07	9.04e-08

Table 1.1. *Bundle adjustment: Jai M10 camera and 10 mm lens*

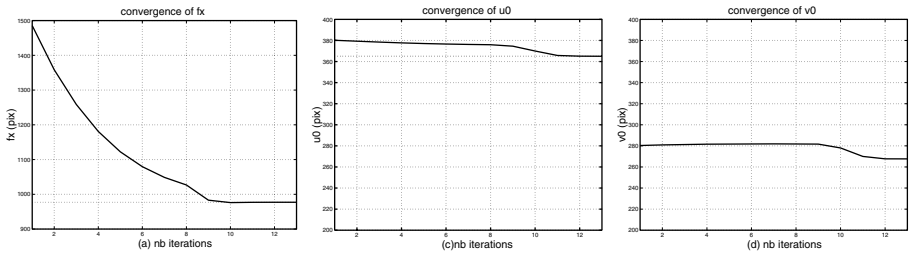


Figure 1.8. *Convergence of* fx, $u0$, $v0$

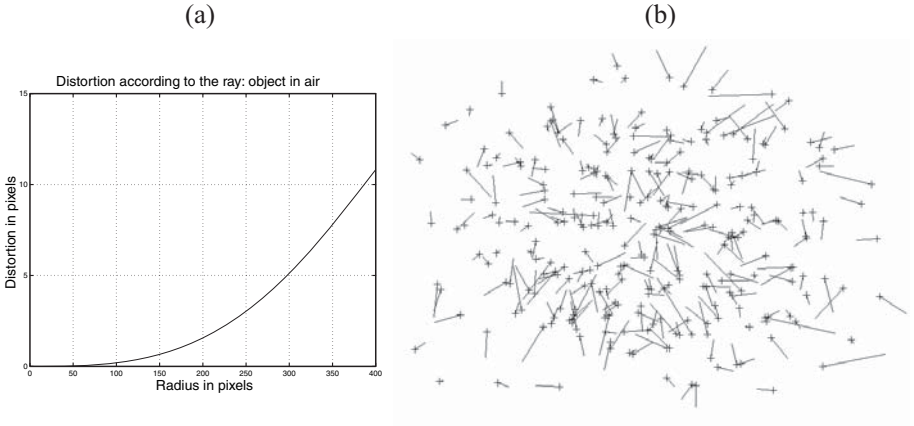


Figure 1.9. *Distortion according to radius (a) and residues in convergence (b) (enlarged 1,000 times) of a set of measuring reference markers (768 × 576 pixel format)*

– In the solution, we recalled the residues of convergence (e_x, e_y) corresponding to expression (1.36). Each cross represents a point of measurement resulting from one of the m images and the associated residue vector undergoes a multiplicative factor of 1,000 in Figure 1.9b. Homogenous distribution of residue vectors and their random orientation translates the non-correlation of errors in solution.

Flexibility in use. The use of the approach described in this chapter proves very flexible and enables calibrations with flexible test cards based on the focal distance of the lens used. The results of the residues obtained in comparison with traditional approaches using a test card measured with precision testify the reliability of the algorithm.

1.5.2. *Specific case of fish-eye lenses*

This section deals with the specific case of fish-eye lenses. The principal deformation generated by a lens with short focal length is a radial deformation. The more the lens presents a small focal distance, the broader its angular field of observation. Rays converging on the CCD matrix have an increasingly significant incidental angle on the very important front diopter of the lens and consequently move away from the assumption of paraxial optics. Therefore, the curvature radii of the lenses induce a dominating radial phenomenon.

The strategy to model the strong deformations is intuitive; it consists of increasing the order of polynomial distortion. For considering strong radial

distortions, it is necessary to have the polynomial order at 5. On the other hand, several writings of the criterion can be considered.

1.5.2.1. *Traditional criterion*

Formula (1.28), which was previously explained, changes into:

$$\left.\begin{aligned} do_{xr} &= \big(u-u_0\big)dx\big(a_1r^2+a_2r^4+a_3r^6+a_4r^8+a_5r^{10}\big)\\ do_{yr} &= \big(v-v_0\big)dy\big(a_1r^2+a_2r^4+a_3r^6+a_4r^8+a_5r^{10}\big)\end{aligned}\right\} \tag{1.43}$$

The distortion polynomial will be composed of five terms instead of three. As Li highlights it in [LI 94], since r can take a maximum value of $\frac{\sqrt{2}}{2}L$ (where L represents the size of the image), $r^2 \ldots r^{10}$ can take very high values and expression (1.43) can then become numerically unstable.

1.5.2.2. *Zero distortion at* r_0

To resolve this disadvantage, it is possible to rewrite expression (1.43) in the following way:

$$\left.\begin{aligned} do_{xr} &= \big(u-u_0\big)dx\Big(a_1\big(r^2-r_0^2\big)+a_2\big(r^4-r_0^4\big)\\ &\quad +a_3\big(r^6-r_0^6\big)+a_4\big(r^8-r_0^8\big)+a_5\big(r^{10}-r_0^{10}\big)\Big)\\ do_{yr} &= \big(v-v_0\big)dy\Big(a_1\big(r^2-r_0^2\big)+a_2\big(r^4-r_0^4\big)\\ &\quad +a_3\big(r^6-r_0^6\big)+a_4\big(r^8-r_0^8\big)+a_5\big(r^{10}-r_0^{10}\big)\Big)\end{aligned}\right\} \tag{1.44}$$

or further:

$$\left.\begin{aligned} do_{xr} &= \big(u-u_0\big)dx\sum_{i=1}^{5}\big(a_i\big(r^{2i}-r_0^{2i}\big)\big)\\ do_{yr} &= \big(v-v_0\big)dy\sum_{i=1}^{5}\big(a_i\big(r^{2i}-r_0^{2i}\big)\big)\end{aligned}\right\} \tag{1.45}$$

Therefore, expression (1.45) forces the distortion to take a zero value for a fixed radial distance r_0. In this equation, there is a variation of the focal distance in solution, which, in the first approximation, will be equal to: $\Delta f = \frac{f}{r_0}dr_0$.

Nevertheless, [LAV 00a] shows that this does not solve any aspect of the fact that the coefficients have numerically very low values when compared to parameters of focal distance, for example. This aspect can, however, be taken into account by a third expression of the criterion.

1.5.2.3. *Normalization of distortion coefficients*

Normalization consists of compensating the significant values of r^n by normalizing all distances as compared to the focal length f:

$$\left.\begin{aligned}(u-u_0)dx &\longrightarrow (u-u_0)dx/f\\ (v-v_0)dy &\longrightarrow (v-v_0)dy/f\end{aligned}\right\}$$

It becomes:

$$r'^2 = \left((u-u_0)\frac{dx}{f}\right)^2 + \left((v-v_0)\frac{dy}{f}\right)^2 = \frac{r^2}{f^2}$$

or even, in distortion expression:

$$\left.\begin{aligned}do'_{xr} &= (u-u_0)\frac{dx}{f}\left(a_1 r'^2 + \cdots + a_5 r'^{10}\right)\\ do'_{xt} &= p_1\left[r'^2 + 2(u-u_0)^2\frac{dx^2}{f^2}\right] + 2p_2(u-u_0)\frac{dx}{f}(v-v_0)\frac{dy}{f}\end{aligned}\right\} \quad (1.46)$$

which we can rewrite by putting f in factor:

$$\left.\begin{aligned}do'_{xr} &= \frac{1}{f}(u-u_0)dx\sum_{i=1}^{5}\left(\frac{a_i}{f^{2i}}r^{2i}\right)\\ do'_{xt} &= \frac{1}{f}\left[\frac{p_1}{f}\left(r^2 + 2(u-u_0)^2\frac{dx^2}{f^2}\right) + 2\frac{p_2}{f}(u-u_0)dx(v-v_0)dy\right]\end{aligned}\right\} \quad (1.47)$$

This new expression shows a form, which is entirely similar to that defined in (1.28), as compared to the near focal length. Then, the total criterion takes the following form:

$$\left.\begin{aligned}e'_x &= \frac{X}{Z} + do_{\prime xr} + do_{\prime xt} - (u-u_0)\frac{dx}{f}\\ e'_y &= \frac{Y}{Z} + do_{\prime yr} + do_{\prime yt} - (v-v_0)\frac{dy}{f}\end{aligned}\right\} \quad (1.48)$$

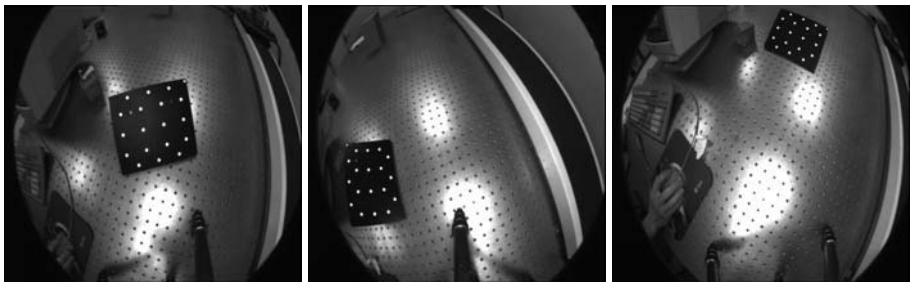


Figure 1.10. *Example of photoshots (format: 768 × 576 pixels)*

The interest of this rewriting lies in the fact that any significant value of (r^{2i}) in the distortion polynomial is compensated by (f^{2i}). We will see in the experimental part that the values of (a_i) will remain close to the unit.

1.5.2.4. *Experiments*

Part of the test sequence to calibrate is presented in Figure 1.10. We can note that we are in the presence of a very strong radial distortion. An exhaustive comparison of the behavior of these three criteria is analyzed in [LAV 00a] and the results presented in Table 1.2 implement the normalized equation of colinearity expressions.

General notes

– The convergence of the calibration algorithm requires some precautions to calibrate a sensor displaying such a distortion. By releasing a set of parameters from the first iteration, the algorithm systematically diverges choosing a solution, which consists of sending the object to infinity. This is particularly true as we move far from the final solution since the coefficients of radial and tangential distortion are initialized at zero.

– To enforce the parameters, we adopt an approach, which blocks the focal distance and the principal point $(fx, fy, u0, v0)$ as long as the average criterion is not passed under a threshold, which is fixed here at 0.6 pixels. Thus, the algorithm will initially estimate distortions, localizations and the geometry of the test card, and then optimize all the parameters as soon as the threshold is crossed. This parameter blocking comes within a Levenberg-Marquardt procedure of optimization while acting on the derivatives of the system (starting value of fx, fy fixed at 400 pixels or 40% of the solution).

Camera	: JAI M300
Lens	: 3.5mm TV-lens
Digitization card	: Silicon Graphics
Algorithm	: Bundle adjustment

Number of images	: 20
Number of measurements	: 360

Residues ex ave and e-type (pix)	9.631e-05	4.98e-02
Residues ey ave and e-type (pix)	−1.309e-05	3.78e-02
Std error of unit weight (pix)		5.15e-02

	value	σ
fx(pix)	326.89	1.82e-01
fy(pix)	327.29	2.00e-01
u0(pix)	396.67	1.78e-01
v0(pix)	258.22	1.81e-01
a1	3.985e-01	4.51e-03
a2	1.892e-02	2.16e-02
a3	4.557e-01	4.54e-02
a4	−3.921e-01	4.22e-02
a5	2.261e-01	1.46e-02
p1	7.304e-06	1.27e-04
p2	1.253e-03	1.97e-04

Table 1.2. *Fish-eye calibration. Normalized criterion*

Comments

– We observe the harmonization of the order of magnitude of radial distortion coefficients in Table 1.2. The variation between the terms does not exceed a factor of 10.

– The distortion in the image edge is higher than 200 pixels (Figure 1.11).

– The decrease of the criterion is carried out without much difficulty. From 0.6 pixels, the 185 parameters are estimated at the same time. All decreasing curves present the same "break" in the vicinity of the solution (0.25 pix) before

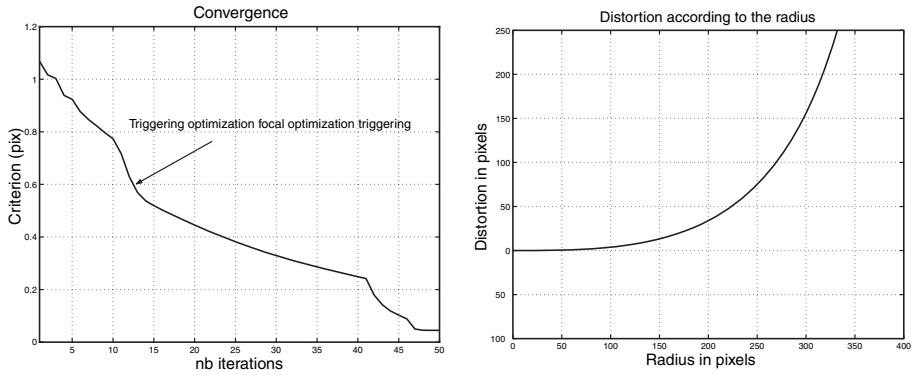


Figure 1.11. *Convergence and distortion of the normalized criterion*

dropping towards the final solution; this rupture is not because of the release of parameter in our part, but due to the choice of the Levenberg-Marquardt strategy of convergence.

– The residues in solution are about 0.05 pixels.

– In Figure 1.12, let us note the presence of some *outliers* (important standard of vector), which can be removed by simple filtering on the values of residues.

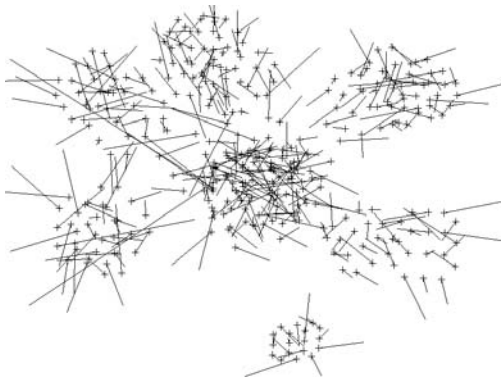


Figure 1.12. *Residues in convergence (×1,000)*

The calibration of lenses with short focal length is based on three major reference markers: a precise detection of markings in images, an adapted optimization criterion and a convergence strategy, which will gradually release the parameters to avoid the local minimum pitfall.

The approach presented on self-calibration by bundle adjustment offers the advantage of quickly obtaining a test card adapted to these specific lenses. Of the three criteria given, two hold our attention more particularly: the second, which consists of fixing a zero distortion at a distance r_0 from the principal point, and the third, which introduces a normalization of the distortion coefficients. In terms of residues in solution and for a constant size of the image, the two criteria are appreciably equivalent; the second offers the advantage of maintaining a practically constant size of the image before and after correction of distortion. As for the third, it offers a more favorable numerical conditioning.

The writing of a new hybrid criterion considering the advantages of the last two must bring a satisfactory solution to this problem. Figure 1.13 shows a sensor view before and after compensation of the distortion phenomena. The size of the final image is increased by 400 pixels in line and column, and leads to a real observation field of the lens of 120 degrees.

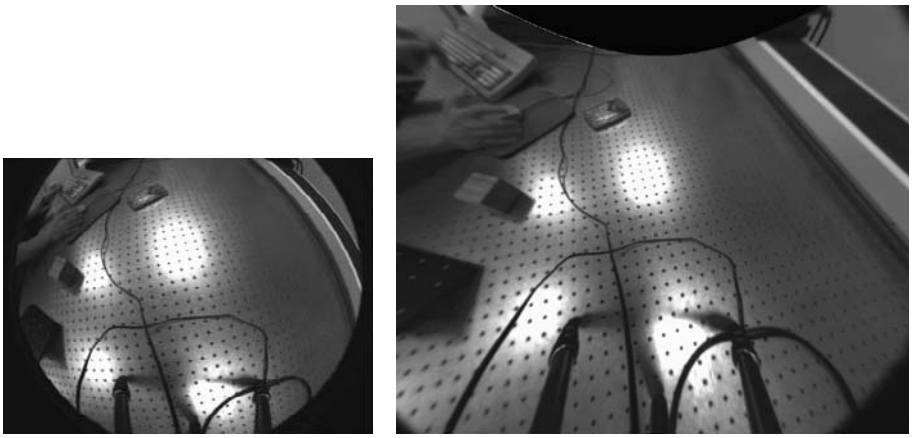


Figure 1.13. *Initial image (format: 768 × 576 pixels) and corrected image (format: 1,168 × 976 pixels)*

1.5.3. *Calibration of underwater cameras*

1.5.3.1. *Theoretical notes*

This section deals with certain concepts related to the calibration of underwater cameras. [LAV 00b] shows the passage of links on the change of focal distance and the modification of distortion curves between the use of

video sensor in air and in water. It is then possible to calibrate the sensor in air and to foresee its operation in a medium of unspecified index.

The following laws must be verified for water:

1) When the camera is plunged in water, we must observe a multiplicative factor of 1.333 on the focal distance value measured in air:

$$f_{\text{water}} = n_{\text{water}} * f_{\text{air}} \tag{1.49}$$

2) Let u be the distorted image of a point in air and du be the distortion correction made to obtain a perfect projection point.

Let u' be the distorted image of the same point in water and du' the distortion correction made to obtain a perfect projection point, then:

$$1.333(u + du) = u' + du' \tag{1.50}$$

1.5.3.2. *Experiments*

The experimental part consists of calibrating an underwater camera in air and then in water and of analyzing the calibration results taking into account the theoretical relations expressed in the preceding section.

1.5.3.3. *The material*

We use an experimental underwater camera made with a Sony CCD sensor. The entire optical device is known (index, dimension and localization of each diopter).

1.5.3.4. *Results in air*

We calibrated the sensor from 12 images presented in Figure 1.14. The effect of radial distortion is significant and the black circle visible at the edge of the image comes from the increase in the angular field of the camera in air. It will disappear when the latter is immersed in water. From these experiments we come to the following conclusions:

– We used an expression of the radial distortion in normalized writing, for a polynomial of order 5 (see section 1.5.2.3).

– Table 1.3 shows the results in the convergence of the system. The residues are about 0.04 pixels. The algorithm is stabilized for $fx = 376$ pixels and

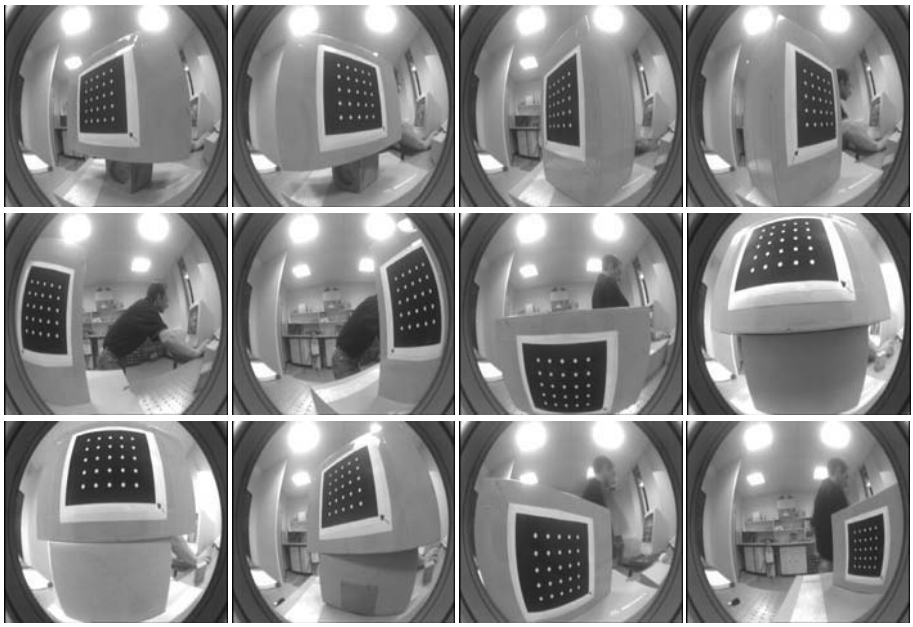


Figure 1.14. *Calibration in air (768 × 576 pixels)*

$fy = 376$ pixels, which corresponds to an angular field of observation of 110 degrees when the distortion is compensated.

– Radial distortion is represented in Figure 1.15a. It takes huge values and beyond 320 pixels (black circle) it no longer corresponds to the physical measurements taken in the images. To represent the compensated view of distortion (Figure 1.15b), we increased the size of the image of 400 pixels (rows and columns). This image corresponds to the sixth view of the sequence (Figure 1.14).

1.5.3.5. *Calibration in water*

Similarly, we calibrated the underwater camera in water with the help of an analysis of 12 views represented in Figure 1.16. As we emphasized before, the black circle at the edge of the image disappeared and this tends to show that there is a modification of the internal behavior of the sensor during index change. Let us finally note that the angular field was strongly reduced and that the distortion at the edge of the image seems less significant than in air.

Results: Calibration in air (index = 1)

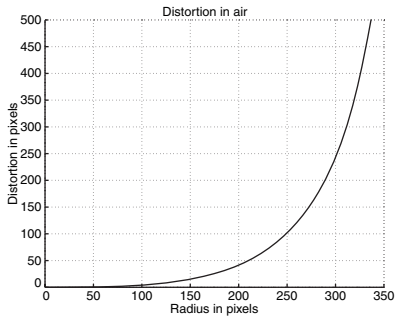
Camera	: Sony underwater
Lens	: 4 mm
Sampling card	: Silicon Graphics
Algorithm	: Bundle adjustment

Number of images	: 12
Number of measurements	: 283

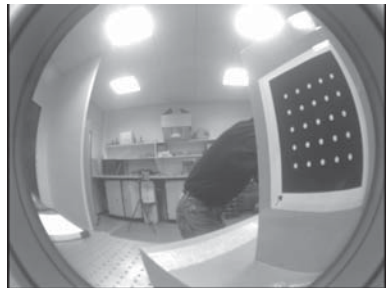
Ave residues e_x and σ (pixel)	2.69e-05	4.15e-02
Ave residues e_y and σ (pixel)	1.49e-04	4.45e-02

	value	σ
fx(pix)	375.65	3.39e-01
fy(pix)	375.81	3.06e-01
u0(pix)	390.87	5.59e-02
v0(pix)	291.75	7.31e-02
a1	6.63e-01	7.52e-03
a2	−1.15e-00	5.02e-02
a3	6.83e+00	1.73e-01
a4	−1.20e+01	2.65e-01
a5	8.95e+00	1.56e-01
p1	−1.02e-03	1.98e-04
p2	1.22e-04	1.91e-04

Table 1.3. *Bundle adjustment in the air on an underwater camera*



(a)



(b)

Figure 1.15. *Radial distortion and corrected image (1,168 × 976 pixels)*

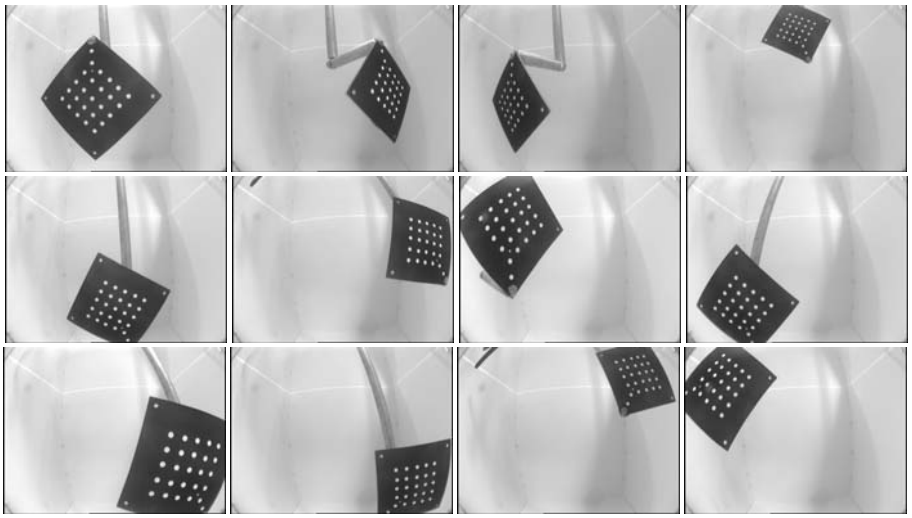


Figure 1.16. *Sequence of photoshots for calibration in water (768 × 576 pixels)*

Notes

– At convergence, the focal distance of the underwater device was stabilized around 500 pixels (instead of 376), which leads to an angular field in water nearing 90 degrees.

– The curve of radial distortion is plotted in Figure 1.17a. It is considerably less significant than for the experiment in air.

Figure 1.17b presents one of the calibration views after compensation of distortions.

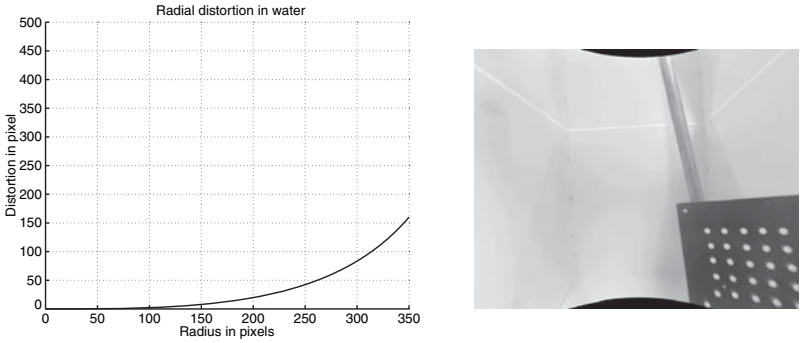


Figure 1.17. *Radial distortion and corrected underwater image (*968 × 776 *pixels)*

Calibration in water (media index = 1.333)
Underwater camera

Camera	: Sony underwater
Lens	: 4 mm
Digitization card	: Silicon Graphics
Algorithm	: Bundle adjustment

Number of images	: 13
Number of measurement	: 325

Ave. residues E_x and σ (pixel)	1.36e-05	4.641e-02
Ave. residues E_y and σ (pixel)	−3.17e-05	5.164e-02

	value	σ
fx(pix)	499.12	5.51e-01
fy(pix)	501.97	5.22e-01
u0(pix)	391.81	9.27e-02
v0(pix)	292.30	1.25e-01
a1	7.52e-01	1.06e-02
a2	−1.84e-00	7.22e-02
a3	9.88e-00	2.54e-01
a4	−1.73e+01	4.11e-01
a5	1.30e+01	2.57e-01
p1	−1.66e-03	2.32e-04
p2	1.62e-03	1.76e-04

Table 1.4. *Bundle adjustment. Underwater camera*

1.5.3.6. *Relation between the calibration in air and in water*

Focal Distance. The theoretical laws of air/water passage are verified in experiments:

– The distance between the nodal point image and the CCD matrix undergoes a multiplicative factor equivalent to the index in which the camera is plunged (see [LAV 00b]).

– Table 1.5 shows the relationship between the focal distances in air and in water. If we integrate the determination uncertainties of focal distances estimated by the process of calibration, this ratio is very close to 1.333.

	f-air	f-water	ratio (f-water/f-air)
fx(pix)	375.65	499.12	1.329
fy(pix)	375.81	501.97	1.336

Table 1.5. *Comparison of the focal distance in air and in water*

Coordinates of the principal point (U_0, V_0). The position of the principal point (intersection of the optical axis and CCD matrix) is relatively stable between the two experiments (Table 1.6).

	Air	Water
u0(pixels)	390.87	391.81
v0(pixels)	291.76	292.30

Table 1.6. *Comparison of the coordinates of the principal point in air and in water*

Distortion. Figure 1.18a shows a joint representation of the distortion curves obtained from the series of measurements in air and water. We can observe the importance of the radial effect during the experimentation in air.

If it is assumed that this distortion is mainly radial, the curves must then verify:

$$1.333\big(u + d_{\text{air}}(u)\big) = u' + d_{\text{water}}(u')$$

Figure 1.18b shows the prediction of distortion curves in water obtained from data calculated in air. As we can note, the superposition is extremely good. Since the view field in air is larger than in water, prediction of distortion of the edge of image in water corresponds to very precise measurements in air since this part of the image is perfectly visible. This is not the case for images in water because of the experimental difficulty in obtaining correct views of the edges of images there.

It seems obvious that the use of an underwater camera can be freed from *in situ* calibration. In fact, underwater shots are not always easy to obtain especially if we want to take shots of convergent images, uniformly distributed on a CCD sensor. Then, a pre-calibration in air is a remarkable solution and leads to very interesting results.

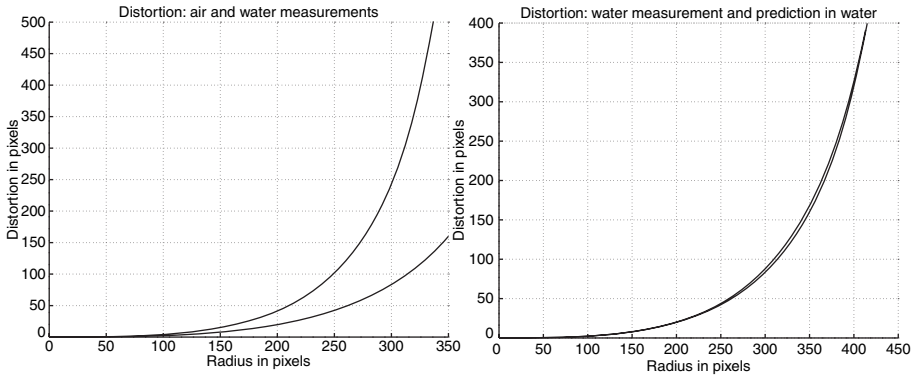


Figure 1.18. *Distortions: (a) curves drawn from air measurements and water measurements, (b) superposition distortion in water and prediction of distortion in water from distortion in air*

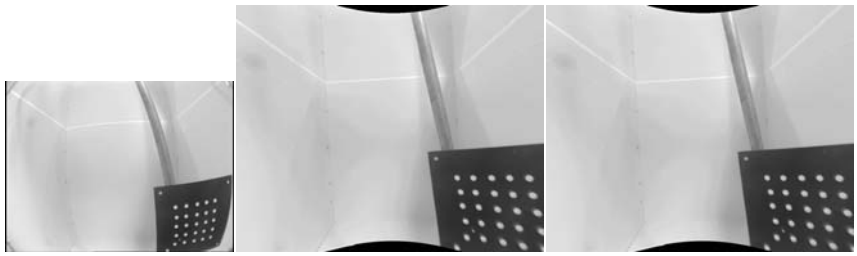


Figure 1.19. *Distortion: (a) original view (768 × 576 pixels), (b) non-distorted view (968 × 776 pixels) from data obtained in water, (c) non-distorted view from data predicted from air*

1.5.4. *Calibration of zooms*

1.5.4.1. *Recalling optical properties*

Changing the focal distance. Change of the focal distance induced in a zoom is obtained by the displacement of a moving block of convergent lenses. This modification in diopter configuration causes an apparent enlargement of the image. However, the transformation undergone by the image is not a simple affinity and we show [LAV 93] that the perspective properties of the device are largely modified.

Two major modifications must be taken into account:

– Modification of the focal distance (fx, fy).

– Modification (often significant) of the distance between the optical center equivalent to the device and the object (T_z). This phenomenon taking place as a result of the displacement of the principal reference marker of the thick optical model can induce axial displacements of a significant amplitude (see [LAV 93] for more details).

To conclude, the zoom action induces modifications on the intrinsic and extrinsic parameters of the vector of calibration.

From a purely theoretical point of view, it will be advisable to consider as many independent principal reference markers as (zoom/focusing) configurations of the lens.

In practice, however, it should be noted that the estimate precision of the principal point decreases with the increase in the focal distance. Indeed, the higher the focal distance, the more the perspective model tends towards the orthographical model in scale, which no longer allows the determination of (u_0, v_0) (we have compensation with T_x and T_y parameters).

For a zoom with great enlargement, the solution consists of joining the principal point with the expanding focus to determine the average coordinates of (u_0, v_0) adapted to the zoom considered [LI 96]. This approximation is particularly valid since the phenomenon of axial translation of the zoom is significant.

1.5.4.2. *Estimate of the principal point*

This estimate is done by analyzing a sequence of images by zooming. Each point of the same object is followed throughout the sequence and we estimate the line in the context of least squares passing by each coordinate (see Figure 1.21). The calculation of the intersection of beam lines provides the expansion focus of the zoom.

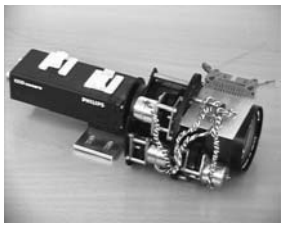


Figure 1.20. *Camera equipped with a zoom*

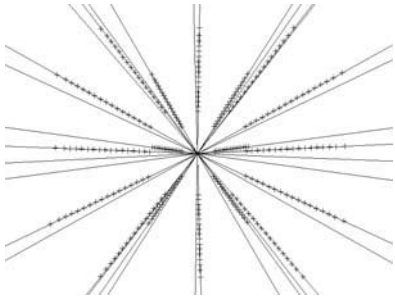


Figure 1.21. *Determination of expanding focus*

Figure 1.22. *Active head of the Sweden's Royal Institute of Technology (KTH-Stockholm)*

1.5.4.3. *Experiments*

Calibration of a zoom must be obtained for all configurations of focal distance and the useful focusing by the device. Figure 1.23 shows the sampling realized during experiments. The amplitudes of displacements are normalized.

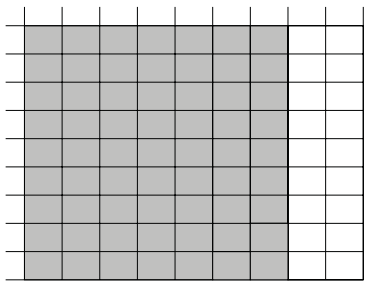


Figure 1.23. *Tested configurations (zones in gray correspond to zoom combinations/focusing realized during the experiments)*

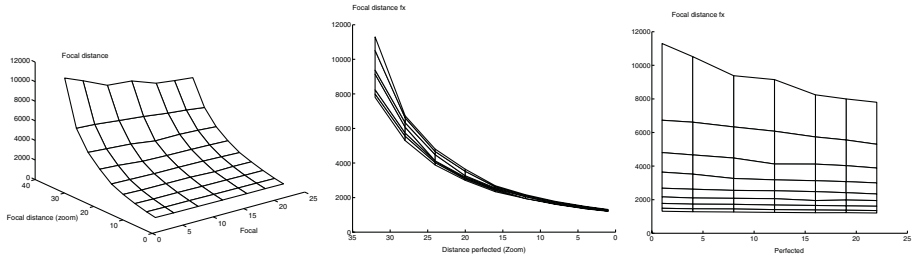


Figure 1.24. *Surface representing the variations of the focal distance according to the focusing and zoom adjustment*

Figures 1.24 shows the calibration results of zooms, which equip the active head of KTH in Stockholm (Professor Eklundh); see Figure 1.22. The curves highlight the effect coupled on the focal distance of a zoom change and also of a focusing change. An exhaustive analysis of the results can be found in [LI 95].

1.6. Bibliography

[AME 84] American Society for Photogrammetry, *Manual of Photogrammetry*, 4th ed., 1984.

[DEM 95] DEMENTHON D. and DAVIS L., "Model-based object pose in 25 lines of code", *International Journal of Computer Vision*, vol. 15, p. 123, 141, June 1995.

[FAU 87] FAUGERAS O. and TOSCANI G., "Camera calibration for 3D Computer Vision", *Proc. of CVPR*, Tokyo, Japan, 1987.

[LAV 93] LAVEST J., RIVES G. and DHOME M., "3D Reconstruction by Zooming", *IEEE Trans. on Robotics and Automation*, vol. 9, no. 2, p. 196–207, April 1993.

[LAV 98] LAVEST J., VIALA M. and DHOME M., "Do we really need an accurate calibration pattern to achieve a reliable camera calibration", *Fifth European Conference on Computer Vision, ECCV98, Freiburg, Germany*, vol. 1, p. 158–174, June 1998.

[LAV 00a] LAVEST J. and DHOME M., "Comment calibrer des objectifs à très courte focale ?", *Conf. Reconnaissnace des Formes et Intelligence Artificielle*, Paris, France, p. 81–90, 2000.

[LAV 00b] LAVEST J., RIVES G. and LAPRESTE J., "Underwater-camera calibration", *Proc. European Conference on Computer Vision*, vol. 2, Dublin, Ireland, p. 654–672, 2000.

[LI 94] LI M., "Camera Calibration a head-eye system for Active Vision", in EKLUNDH J. (Ed.), *Proc. of 3rd European Conf. on Computer Vision*, ECCV94, Stockholm, vol. 1, p. 543–554, May 1994.

[LI 95] LI M. and LAVEST J., "Some aspects of zoom lens calibration", *Technical Report KTH-NADA-CVAP*, no. 172, 1995.

[LI 96] LI M. and LAVEST J., "Some aspects of zoom lens calibration", *IEEE PAMI*, vol. 18, no. 11, 1996.

[PRE 92] PRESS W., TEUKOLSKY S., VETTERLING W. and FLANNERY B., *Numerical Recipies in C*, Cambridge University Press, 2nd ed., 1992.

[TOR 81] TORLEGÅRD K., "Accuracy Improvement in Close Range Photogrammetry", *Schriftenreihe, Wissenschaftlicher Studiengang Vermessungswesen, Hochschule der Bunderwehr München*, vol. 5, September 1981.

This page intentionally left blank

Chapter 2

Self-Calibration of Video Sensors

2.1. Introduction

This chapter deals with the fundamental problem of self-calibration of a camera with the help of a set of points paired between various images. A method is developed based on Kruppa equations, which are well-known in the context of this application. We make use of the decomposition of fundamental matrix in singular values to derive remarkably simplified Kruppa equations in a purely algebraic method. This enables us, in particular, to solve the problem of choosing two Kruppa equations among the larger group of equations derived using the traditional method. In this method, the equations are derived very simply and we by no means make use of the geometric interpretation with an absolute conic as a base, or that related to infinite plane, and we do not explicitly use epipoles, whose estimate is known to be unstable. Finally and especially, this method is implemented, compared and tested successfully to find the intrinsic parameters of various cameras from noisy synthetic data and several real images. It is also shown that the quality of the results obtained makes it possible to validate the approach until reliable 3D measurements are obtained with the help of images.

A camera is a projective machine [FAU 92, FAU 93, HAR 92b], which makes it possible to obtain the image **m**, in the retinal plane, from a point **M** in space. This is obtained by intersecting the retinal plane with the optical radius,

---

Chapter written by Rachid DERICHE.

which passes by the center of projection or the optical center and by the point **M**. This operation is linear in projective coordinates and uses a rank 3 3×4 matrix, known as the perspective projection matrix, which depends on five intrinsic parameters and six other parameters, known as extrinsic parameters, which determine the position and orientation of the camera (3 of rotation and 3 of translation).

Calibration of a camera consists of estimating its perspective projection matrix, which enables us to trace its set of intrinsic and extrinsic parameters. This is generally carried out by means of a certain number of landmarks whose 3D coordinates and their 2D projections are easily identifiable and known with high degree of accuracy. This technique, very much used by the photogrammetry community [ZEL 52, FAI 75, SLA 80, WOL 83] and that of computer vision [BRO 71, TSA 87, TSA 86, FAU 86, TOS 87] in order to trace the retina metrics, is however very difficult and constraining owing to the fact that it requires the use of a set of very precise 3D and 2D landmarks, which are often not available in the case of the application considered.

The self-calibration of a camera thus consists of tracing its intrinsic parameters, and thus the retina metrics, without using 3D landmarks. Only the information present in the images is accessible i.e., we have only 2D landmarks extracted from images taken by the camera to be self-calibrated. In the past decade, this problem of capital importance in several applications has attracted the attention of the computer vision community, which developed a fine analysis of the constraints that the intrinsic parameters of a moving camera taking various views must respect. In this chapter, we will only consider the case where the intrinsic parameters are assumed to be constant. For recent works related to self-calibration in the case of variable intrinsic parameters, or complementary works on calibration by specific movements or stereoscopic systems see [HEY 97, POL 98, POL 99, AGA 98, AGA 99, LOU 00a, DEV 96, VIE 94, VIE 99, VIE 96b, VIE 96a, ENC 94].

In the case of 2 images perceived by a camera whose intrinsic parameters are supposed to be constant, we can show that the knowledge of at least 8 points paired between the two views makes it possible to reach, in a unique way, a 3×3 matrix of row 2 known as the *fundamental matrix* from which two polynomial constraints on the intrinsic parameters can be exhibited. Various approaches were developed in order to show these constraints. In this chapter we will focus on a detailed presentation of the approach known as Kruppa *equations*, which enable us to efficiently solve the problem of

self-calibration. To achieve this, we will start by presenting the approach known as *Huang-Faugeras* constraints [HUA 89] and the approach developed by *Trivedi* [TRI 88], as they constitute a good introduction to Kruppa equations. The approach known as *Huang-Faugeras* constraints [HUA 89] indeed makes it possible to deduce, by way of a fundamental matrix structure, a polynomial equation of order 8 in the intrinsic parameters, whereas the *Trivedi* approach [TRI 88] makes it possible to show that there exists not only 1 but 2 polynomial constraints of order 8. Then, Luong [LUO 92, LUO 97] showed that the polynomial *Huang-Faugeras* constraint can be divided into two independent polynomial relations, equivalent to two of *Trivedi's* equations. A purely geometric approach, and its only algebraic equivalent are then presented in order to simplify these constraints and to reduce their order from 8 to 4. These new polynomial constraints, known as *Kruppa* equations in honor of the Austrian mathematician Erwin Kruppa, are then remarkably simplified by way of decomposition into singular values of the fundamental matrix.

Several approaches of self-calibration were developed within the computer vision community. As a complement to what is presented in this chapter, see the works by Luong [LUO 92], Zeller [ZEL 96a, ZEL 96b], Heyden and Åström [HEY 96], Pollefeys and Van Gool [POL 97b, POL 97a] and Luong and Faugeras [LUO 97]. Some authors were interested in particular types of movements to self-calibrate. Thus, Hartley [HAR 97a] showed that in the case of a pure rotation, self-calibration can be solved with the assistance of a linear algorithm. Other cases of interest, of which some include stereoscopic systems, were studied by Armstrong *et al*. [ARM 96, ZHA 93, HOR 94, BRO 96, HOR 98]. Lastly, an inventory of critical movements of the cameras for which Kruppa equations degenerate and do not provide information or provide only a partial information on the intrinsic parameters (translatory movements, planes, rotation around a point, etc.) was started by Zeller in [ZEL 96a] and was continued by Sturn [STU 97] and Zisserman *et al*. [ZIS 98].

This chapter is organized as follows. This first section provides an introduction to the problem of self-calibration; section 2.2 contains some fundamental notations and reminders; section 2.3 is devoted to recalling *Huang-Faugeras* constraints [HUA 89] and those of *Trivedi* [TRI 88], which constitute a good introduction to Kruppa equations. The latter are then derived in section 2.4, initially with the help of a traditional geometric approach and then with the help of an algebraic approach, which will make it

possible to simplify them remarkably. The last section is finally devoted to the presentation of some experimental results on synthetic and real data.

2.2. Reminder and notation

A projective camera model is assumed. This model appeals to projective geometry, of which some basic recalls and concepts can be found in [MOH 93, FAU 93, MUN 92].

This model projects a 3D point $\mathbf{M} = [x, y, z]^t$ in space in a 2D point $\mathbf{m} = [u, v]^t$ in the image, through projection matrix 4×3, which is noted by $\mathbf{P}$

$$s\hat{m} = \mathbf{P}\hat{\mathbf{M}} \tag{2.1}$$

where s is an arbitrary scale factor, but non-zero, and the notation $\hat{\mathbf{p}}$ is such that if $\mathbf{p} = [x, y, \ldots]^t$, then $\hat{\mathbf{p}} = [x, y, \ldots, 1]^t$.

We consider a system of stereo cameras and provide two 2D points $\mathbf{m_1}$ and $\mathbf{m_2}$ resulting from the projection of the same physical point $\mathbf{M}$ in space. Then, we have the 2 following relations:

$$s_1\hat{\mathbf{m}}_\mathbf{1} = \mathbf{P_1}\hat{\mathbf{M}}, \quad s_2\hat{\mathbf{m}}_\mathbf{2} = \mathbf{P_2}\hat{\mathbf{M}} \tag{2.2}$$

If we place ourselves in the reference mark associated with the 1st camera, the projection matrices are then given by:

$$\mathbf{P_1} = [\mathbf{A} \mid \mathbf{0}], \quad \mathbf{P_2} = [\mathbf{A'R} \mid \mathbf{A't}],$$

where $\mathbf{R}$ and $\mathbf{t}$ respectively represent the rotation matrix and the translation vector associated with the rigid movement between the 2 cameras. Note that a similar relation exists in the case of the two views taken by a single moving camera.

Matrix $\mathbf{A}$ and matrix $\mathbf{A}'$ are the matrices known as intrinsic parameters. They are given by the following well-known form [FAU 95]:

$$\mathbf{A} = \begin{bmatrix} \alpha_u & -\alpha_u \cot\theta & u_0 \\ 0 & \alpha_v / \sin\theta & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{A'} = \begin{bmatrix} \alpha'_u & -\alpha'_u \cot\theta' & u'_0 \\ 0 & \alpha'_v / \sin\theta' & v'_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

where the parameters α_u (resp. α'_u) and α_v (resp. α'_v) correspond to the focal distances in pixels along the image axes, and θ (resp. θ') represents the angle between the 2 axes of the image and (u_0, v_0) (resp. (u'_0, v'_0)) the coordinates of the principal point in the reference mark image (i.e., the point defined by the intersection of the optical axis with the image plane).

The ratio $\frac{\alpha_v}{\alpha_u}$ (resp. $\frac{\alpha'_v}{\alpha'_u}$) is known under the name of *aspect ratio* or scale factor. When the 2 images are consequently acquired by the same camera, $\theta = \theta'$ and $\frac{\alpha_v}{\alpha_u} = \frac{\alpha'_v}{\alpha'_u}$, even if the focal distances are not identical. Practically, θ and θ' are very close to $\frac{\pi}{2}$ for real cameras. [FAU 93]. Moreover, if the pixels are square, which is the case for many recent cameras, the *aspect ratio* or scale factor is equal to one.

By eliminating the scalars s_1 and s_2 and $\mathbf{M}$ of 2 projection equations, we obtain the following relation, which connects two 2D points resulting from the same physical point $\mathbf{M}$:

$$\hat{\mathbf{m}}_\mathbf{2}{}^t \mathbf{F} \hat{\mathbf{m}}_\mathbf{1} = 0 \tag{2.4}$$

In this equation, the matrix F, known as the *fundamental matrix*, is given by:

$$\mathbf{F} = \mathbf{A}'^{*}[\mathbf{t}]_\times \mathbf{R} \mathbf{A}^{-1} \tag{2.5}$$

where $\mathbf{A}'^{*} = (\mathbf{A}'^{-1})^t$ is the adjoint *matrix* of $\mathbf{A}'$ and the notation $[\mathbf{x}]_\times$ represents the anti-symmetric matrix $[\mathbf{x}]_\times$ such that $\mathbf{x} \times \mathbf{y} = [\mathbf{x}]_\times \mathbf{y}$ irrespective of $\mathbf{y}$:

$$[\mathbf{x}]_\times = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}$$

The fundamental matrix $\mathbf{F}$ describes the epipolar geometry between the 2 views considered.

We finally conclude by recalling what is the essential matrix given by $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$. This matrix $\mathbf{E}$ is related to the fundamental matrix $\mathbf{F}$ through relation (2.5), which becomes (in the case of two identical cameras):

$$\mathbf{F} = \mathbf{A}^{*} \mathbf{E} \mathbf{A}^{-1} \tag{2.6}$$

This makes it possible to write $\mathbf{E}$ as a function of $\mathbf{F}$, as follows:

$$\mathbf{E} = \mathbf{A}^t\mathbf{F}\mathbf{A} \tag{2.7}$$

As noted by Trivedi [TRI 88], the symmetric matrix $\mathbf{E}\mathbf{E}^t$ is independent of rotation $\mathbf{R}$ since:

$$\mathbf{E}\mathbf{E}^t = [\mathbf{t}]_\times \mathbf{R}\mathbf{R}^t \left([\mathbf{t}]_\times\right)^t = [\mathbf{t}]_\times \left([\mathbf{t}]_\times\right)^t \tag{2.8}$$

By substituting equation (2.7) in these equations and since:

$$\mathbf{F}\mathbf{A} = \mathbf{A}^*\mathbf{E} \tag{2.9}$$

we obtain:

$$\mathbf{F}\mathbf{K}\mathbf{F}^t = \mathbf{A}^*[\mathbf{t}]_\times \left([\mathbf{t}]_\times\right)^t \mathbf{A}^{-1} \tag{2.10}$$

where $\mathbf{K}$ is the symmetric matrix $\mathbf{A}\mathbf{A}^t$, which will intervene later on during algebraic derivations of Kruppa equations.

2.3. Huang-Faugeras constraints and Trivedi's equations

Before presenting the Kruppa equations in a simplified way in the following section, we recall *Huang-Faugeras* constraints here [HUA 89] and also Trivedi's equations [TRI 88] and we discuss their equivalence.

2.3.1. *Huang-Faugeras constraints*

As the essential matrix $\mathbf{E}$ is the product $[\mathbf{t}]_\times\mathbf{R}$ of an antisymmetric matrix by a matrix of rotation, its row is always 2. According to Huang and Faugeras [HUA 89], for the matrix to be essential, it is necessary and it is enough that it satisfies the constraint of row 2, plus the constraint that the 2 non-zero singular values of $\mathbf{E}$ must be equal. This last constraint is equivalent to:

$$\mathrm{trace}^2\left(\mathbf{E}\mathbf{E}^\mathbf{t}\right) - 2\,\mathrm{trace}\left(\left(\mathbf{E}\mathbf{E}^\mathbf{t}\right)^2\right) = 0 \tag{2.11}$$

Using relation (2.7), we then obtain the following constraint, which uses only the fundamental matrix $\mathbf{F}$ and matrix $\mathbf{A}$, which is unknown in the case of the self-calibration problem,

$$\mathrm{trace}^2\left(\mathbf{A}^\mathbf{t}\mathbf{F}\mathbf{A}\mathbf{A}\mathbf{F}^\mathbf{t}\mathbf{A}\right) - 2\,\mathrm{trace}\left(\left(\mathbf{A}^\mathbf{t}\mathbf{F}\mathbf{A}\mathbf{A}\mathbf{F}^\mathbf{t}\mathbf{A}\right)^2\right) = 0 \tag{2.12}$$

Owing to the fact that the fundamental matrix is like $\det(\mathbf{F}) = 0$, the constraint of row on $\mathbf{E}$ is always verified. On the other hand, equation (2.12) is a polynomial constraint of order 8 in the coefficients of $\mathbf{A}$, i.e., of order 8 in intrinsic parameters. This constraint can then be used within the context of the self-calibration application: given a moving camera observing a certain number of characteristic points, each movement will create a fundamental matrix, which will make it possible to write a polynomial constraint in the coefficients of $\mathbf{A}$. In principle, this paves the way for recovering the intrinsic parameters of the camera with a sufficient number of movements.

2.3.2. *Trivedi's constraints*

Another manner to approach the constraints on $\mathbf{F}$ and $\mathbf{A}$ is to proceed as Trivedi did in [TRI 88], making good use of the particular shape of the symmetric matrix $\mathbf{E}\mathbf{E}^{\mathbf{t}}$, which depends only on the parameters of the translation vector $\mathbf{t}$ to derive a set of 2 independent polynomial constraints. By developing equation (2.8), we find the matrix $\mathbf{S} = \mathbf{E}\mathbf{E}^{\mathbf{t}}$ in the following form:

$$\mathbf{S} = \begin{bmatrix} t_2^2 + t_3^2 & -t_1 t_2 & -t_1 t_3 \\ -t_2 t_1 & t_3^2 + t_1^2 & -t_2 t_3 \\ -t_3 t_1 & -t_3 t_2 & t_1^2 + t_2^2 \end{bmatrix}$$

Owing to the fact that the symmetric matrix $\mathbf{E}\mathbf{E}^{\mathbf{t}}$ depends only on the 3 parameters of the translation vector $\mathbf{t}$, the 3 diagonal elements and the three non-diagonal elements of $\mathbf{S}$ are connected by the following 3 relations:

$$4S_{ij}^2 - \big(\operatorname{trace}(S) - 2S_{ii}\big) * \big(\operatorname{trace}(S) - 2S_{jj}\big) = 0 \qquad (2.13)$$

$$1 \leq i, j \leq 3 \qquad (2.14)$$

These three polynomial constraints, which correspond only to 2 independent equations and an identity by considering the nullity of the determinant of $\mathbf{E}$, are of degree four in the coefficients of $\mathbf{E}$, and thus of degree eight in the coefficients of $\mathbf{A}$.

These 2 constraints can then be used within the context of an application of self-calibration by following the same protocol as that described in the preceding section about *Huang-Faugeras* constraint. However, here we will have 2 equations for each movement, instead of only one as in the preceding case.

2.3.3. *Discussion*

Luong [LUO 92, LUO 97] showed that the polynomial constraint given by (2.12) and known as *Huang-Faugeras* can be divided into two independent polynomial relations which are equivalent to the 2 *Trivedi's* equations given by (2.14). Like Luong, we can notice that the equations obtained from *Trivedi* and *Huang-Faugeras* constraints present the disadvantage of corresponding to polynomials of order 8 in the matrix coefficients of intrinsic parameters $\mathbf{A}$, whereas, as it will be seen in the following section, Kruppa equations correspond to 2 polynomials of degree two only in matrix coefficients $\mathbf{K} = \mathbf{A}\mathbf{A}^{\mathbf{t}}$, and thus of degree four in matrix coefficients of intrinsic parameters $\mathbf{A}$. This explains the complete idea of the approach with the help of Kruppa equations, the idea being mainly based on the handling of polynomials of smaller degree. The simplified Kruppa equations that we will present later on in this chapter have the uniqueness of avoiding explicit calculation of epipoles, known to be unstable, [LUO 98] and making use of the fundamental matrices between different views used in order to achieve a reliable estimate of the intrinsic parameters.

2.4. Kruppa equations

In this section, we start by finding the well-known Kruppa equations in the field of self-calibration by a purely geometric interpretation using the absolute conic [FAU 90, MAI 92]. We then develop a purely algebraic second approach, i.e., without using the absolute conic or infinite plane [ZEL 96a, POL 97a], which will then be used in order to simplify these equations to a great extent.

2.4.1. *Geometric derivation of Kruppa equations*

Geometric derivation of Kruppa equations is based on a significant projective invariant, i.e. the absolute conic, which enables us to geometrically express the rigidity constraint of a movement. Let us take a point $\tilde{\mathbf{M}}^{\mathbf{t}} = [x_1, x_2, x_3, x_4]$; the absolute conic Ω is defined as the intersection of the quadratic equation $x_1^2 + x_2^2 + x_3^2 + x_4^2 = 0$ with the infinite plane Π_∞. The absolute conic Ω is thus defined by $x_1^2 + x_2^2 + x_3^2 = 0$ and $x_4 = 0$ equations. [FAU 95]. Therefore, it is easy to show that each point $\mathbf{m}$ belonging to the projection ω of Ω in the second image belongs to the cone of equation $\mathbf{m}^t\mathbf{A}^{-t}\mathbf{A}^{-1}\mathbf{m} = 0$. Thus, the matrix of intrinsic parameters $\mathbf{A}$ defines the equation of the absolute conic image. If the intrinsic parameters do not vary, the absolute conic image is invariant when the camera is moved.

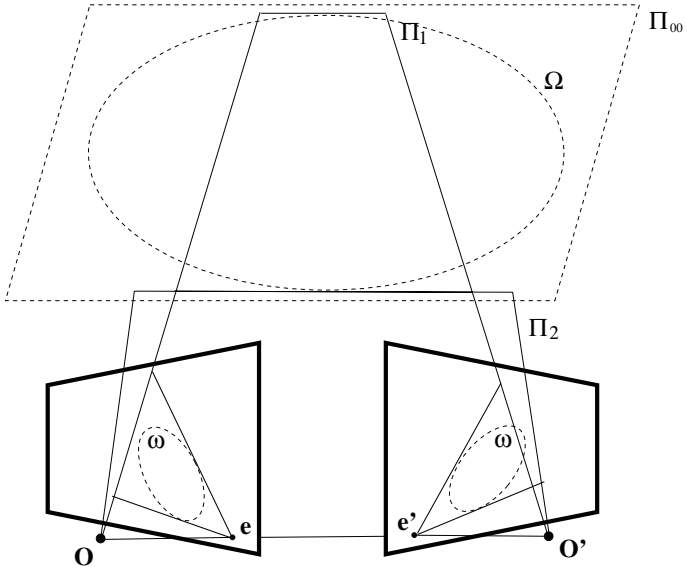


Figure 2.1. *Geometric interpretation of Kruppa equations*

As illustrated in Figure 2.1, there are 2 planes Π_1 and Π_2 made by a straight line OO' and a tangent to Ω. These 2 planes intersect the retinal planes with two pairs of straight lines tangential to ω, which are connected by epipolar homography.

The set of tangents in a cone also form a cone, known as *dual* cone. The matrix of dual cone ω has an equation $\mathbf{l}^t\mathbf{A}\mathbf{A}^t\mathbf{l} = 0$, where $\mathbf{l}$ is a line tangent to ω. This implies that for any point $\mathbf{m}$ on tangents to ω in the second image, we have:

$$(\mathbf{e}' \times \mathbf{m})^t\mathbf{A}\mathbf{A}^t(\mathbf{e}' \times \mathbf{m}) = 0$$

The epipolar line $\mathbf{F}^t\mathbf{m}$ corresponding to $\mathbf{m}$ in the first image is also a tangent at ω. Consequently, the invariance of ω in rigid transformation gives:

$$\left(\mathbf{F}^t\mathbf{m}\right)^t\mathbf{A}\mathbf{A}^t\left(\mathbf{F}^t\mathbf{m}\right) = 0$$

These equations in a simple manner express the constraint that the tangent epipolar lines to ω in the second image correspond to the tangent epipolar lines to ω in the first image, and induce two quadratic equations in the coefficients

of $\mathbf{K} = \mathbf{A}\mathbf{A}^{\mathbf{t}}$ corresponding to the two well-known Kruppa equations:

$$\mathbf{F}\mathbf{A}\mathbf{A}^t\mathbf{F}^t = \beta\big([\mathbf{e}']_\times\big)^t\mathbf{A}\mathbf{A}^t\big([\mathbf{e}']_\times\big), \tag{2.15}$$

where β is a non-zero, arbitrary scalar. Since the intrinsic parameters are five in number, theoretically, three views are sufficient to calculate them.

2.4.2. *An algebraic derivation of Kruppa equations*

In this section, an algebraic derivation of Kruppa equations is developed. We start by deriving a solution for the epipole $\mathbf{e}'$ in the second image, given by $\mathbf{F}^t\mathbf{e}' = \mathbf{0}$, using equation (2.5), which gives the form of the fundamental matrix. This gives:

$$\mathbf{A}^{-t}\mathbf{R}^t\big([\mathbf{t}]_\times\big)^t\mathbf{A}^{-1}\mathbf{e}' = \mathbf{0} \tag{2.16}$$

Owing to the fact that $([\mathbf{t}]_\times)^t\mathbf{t} = \mathbf{0}$, we obtain the following solution for $\mathbf{e}'$:

$$\mathbf{e}' = \lambda\mathbf{A}\mathbf{t}, \tag{2.17}$$

where λ is a non-zero scalar. This also gives the following relation for $\mathbf{t}$:

$$\mathbf{t} = \lambda'\mathbf{A}^{-1}\mathbf{e}', \tag{2.18}$$

where $\lambda' = 1/\lambda$. Equation (2.18) allows us to obtain the following relation for the $[\mathbf{t}]_\times$[1] matrix:

$$[\mathbf{t}]_\times = \lambda'\det\big(\mathbf{A}^{-1}\big)\mathbf{A}^t[\mathbf{e}']_\times\mathbf{A} \tag{2.19}$$

By substituting the latter relation in equation (2.10), we directly get Kruppa equations:

$$\mathbf{F}\mathbf{K}\mathbf{F}^t = \gamma[\mathbf{e}']_\times\mathbf{K}\big([\mathbf{e}']_\times\big)^t, \tag{2.20}$$

where γ is a non-zero scalar. We observe that since $([\mathbf{e}']_\times)^t = -[\mathbf{e}']_\times$, equation (2.20) is identical to equation (2.15), which was geometrically derived. Since $\mathbf{F}\mathbf{K}\mathbf{F}^t$ is a symmetric matrix, equation (2.20) corresponds to

1. Using the relation $[\mathbf{M}\mathbf{u}]_\times = \det(\mathbf{M})\mathbf{M}^*[\mathbf{u}]_\times\mathbf{M}^{-1}$, where $\mathbf{M}$ is a non-singular matrix.

the following scalar equations obtained by eliminating γ:

$$\begin{aligned}\frac{\mathbf{FKF}^t_{11}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{11}} &= \frac{\mathbf{FKF}^t_{12}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{12}} \\ &= \frac{\mathbf{FKF}^t_{22}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{22}} = \frac{\mathbf{FKF}^t_{13}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{13}} \\ &= \frac{\mathbf{FKF}^t_{23}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{23}} = \frac{\mathbf{FKF}^t_{33}}{\left([\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)_{33}}\end{aligned} \tag{2.21}$$

where $\mathbf{A}_{ij}$ refers to matrix element $\mathbf{A}$, indexed by i, j. However, these equations are linearly dependent, as we have:

$$\left(\mathbf{FKF}^t - \gamma[\mathbf{e}']_\times \mathbf{K}\left([\mathbf{e}']_\times\right)^t\right)\mathbf{e}' = \mathbf{0} \tag{2.22}$$

Now, only 2 independent equations remain among the pair of equations given by (2.21). These equations are polynomials of order 2 in the coefficients of $\mathbf{K}$ and thus of 4 in the coefficients of $\mathbf{A}$. This makes them, *a priori*, more suitable for the problem of self-calibration than the Huang-Faugeras equations [HUA 89] or those of Trivedi [TRI 88] presented in the preceding section. Therefore, a possible protocol is to start by estimating $\mathbf{K}$, then $\mathbf{A}$ from $\mathbf{K}$, by Cholesky's decomposition[2].

However, at this stage, some questions arise on the choice of using 2 equations among all those given by (2.22). This problem can be solved, either by the choice of a parametrization of epipolar geometry as in [FAU 90, MAY 92], or by arbitrarily taking an equation in order to estimate the scale factor and to reflect it on the rest by arbitrarily choosing 2 among the remaining 5 [BOU 98]. The idea of taking all equations into account can also be considered. The approach described in the following section will make it possible to easily answer the question relating to the most convenient choice of equations, by directly deriving 3 equations. These equations have a close link with those deduced by Luong [LUO 92] with the help of a different step, based on changing rather astute reference marks, which amounts to generalizing Kruppa equations by considering that the absolute conic can

2. Cholesky's decomposition of a positive definite matrix $\mathbf{B}$ is a matrix $\mathbf{C}$ such that $\mathbf{B} = \mathbf{C}^t\mathbf{C}$ [GOL 89].

have 2 different images in each retina. It is this same idea of Luong that is found in the recent article by Hartley [HAR 97b], where he derives Kruppa equations directly from **F**, by operating changes of reference mark identical to those of Luong. These equations are identical to those presented in the following section, but this chapter presents a different and a remarkably simple approach. We can also note the fact that nowadays, no experimental result has been produced on these simplified equations of Kruppa. Luong used his simplified relations to show the similarity in the *Trivedi* and *Huang-Faugeras* constraints. Hartley followed this method to show that Kruppa relations can be directly expressed from the fundamental matrix but no experimental result was reported. In the following section, the method presented by us is implemented, compared and successfully tested to find the intrinsic parameters of various cameras from noisy synthetic data and several real images. It is also shown that the quality of the results obtained makes it possible to validate the approach to a great extent, until reliable 3D measurements are obtained from images.

2.4.3. *Simplified Kruppa equations*

In this section, in a purely algebraic method, we use the decomposition in singular values [GOL 89] of the fundamental matrix to derive remarkably simplified Kruppa equations. This makes it possible, in particular, to solve the problem of selecting two Kruppa equations to use among the larger group of equations derived using the traditional method and not explicitly requiring an estimate of epipoles:

$$\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^t \tag{2.23}$$

Owing to the fact that **F** is of row 2, the diagonal matrix **D** is of form:

$$D = \begin{bmatrix} r & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

r and s are the square roots of eigenvalues of the FF^t matrix, whereas $\mathbf{U}$ and $\mathbf{V}$ are 2 orthogonal matrices.

Using equation (2.23), the epipole in the second image $\mathbf{e}'$ can therefore be deduced in a very simple manner. Indeed:

$$\mathbf{F}^t\mathbf{e}' = \mathbf{V}\mathbf{D}^t\mathbf{U}^t\mathbf{e}' = \mathbf{0} \tag{2.24}$$

Since **D** is a diagonal matrix whose last element is zero, this gives us the following direct solution for $\mathbf{e}'$:

$$\mathbf{e}' = \delta \mathbf{U}\mathbf{m}, \quad \mathbf{m} = [0, 0, 1]^t, \quad \delta \neq 0.$$

As a result, the matrix $[\mathbf{e}']_x$ is equal to:

$$[\mathbf{e}']_\times = \mu \mathbf{U}\mathbf{M}\mathbf{U}^t, \tag{2.25}$$

where μ is a non-zero scalar and $\mathbf{M} = [\mathbf{m}]_\times$ is given by:

$$\mathbf{M} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

By incorporating equation (2.25) into (2.19) and then into (2.10), we obtain a new expression of Kruppa equations:

$$\mathbf{F}\mathbf{K}\mathbf{F}^t = \nu \mathbf{U}\mathbf{M}\mathbf{U}^t\mathbf{K}\mathbf{U}\mathbf{M}^t\mathbf{U}^t, \quad \nu \neq 0 \tag{2.26}$$

By multiplying the two members of (2.26) by $\mathbf{U}^t$ and $\mathbf{U}$, respectively, and since **U** is an orthogonal matrix, we finally obtain the following well simplified equations:

$$\mathbf{D}\mathbf{V}^t\mathbf{K}\mathbf{V}\mathbf{D}^t = \nu \mathbf{M}\mathbf{U}^t\mathbf{K}\mathbf{U}\mathbf{M}^t \tag{2.27}$$

Owing to the simple forms of **D** and **M**, equations (2.27) correspond only to 3 linearly dependent equations. In fact, by denoting $\mathbf{u_1}$, $\mathbf{u_2}$ and $\mathbf{u_3}$ the vector columns of matrix **U**, and $\mathbf{v_1}$, $\mathbf{v_2}$ and $\mathbf{v_3}$ the vector columns of matrix **V**, the expressions are simplified as follows:

$$\mathbf{D}\mathbf{V}^t\mathbf{K}\mathbf{V}\mathbf{D}^t = \begin{bmatrix} r^2\mathbf{v_1}^t\mathbf{K}\mathbf{v_1} & rs\mathbf{v_1}^t\mathbf{K}\mathbf{v_2} & 0 \\ sr\mathbf{v_2}^t\mathbf{K}\mathbf{v_1} & s^2\mathbf{v_2}^t\mathbf{K}\mathbf{v_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M}\mathbf{U}^t\mathbf{K}\mathbf{U}\mathbf{M}^t = \begin{bmatrix} \mathbf{u_2}^t\mathbf{K}\mathbf{u_2} & -\mathbf{u_2}^t\mathbf{K}\mathbf{u_1} & 0 \\ -\mathbf{u_1}^t\mathbf{K}\mathbf{u_2} & \mathbf{u_1}^t\mathbf{K}\mathbf{u_1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

which finally gives the following 3 linearly dependent equations:

$$\frac{r^2\mathbf{v_1}^t\mathbf{K}\mathbf{v_1}}{\mathbf{u_2}^t\mathbf{K}\mathbf{u_2}} = \frac{rs\mathbf{v_1}^t\mathbf{K}\mathbf{v_2}}{-\mathbf{u_2}^t\mathbf{K}\mathbf{u_1}} = \frac{s^2\mathbf{v_2}^t\mathbf{K}\mathbf{v_2}}{\mathbf{u_1}^t\mathbf{K}\mathbf{u_1}} \tag{2.28}$$

In these 3 equations, only 2 are independent, which are Kruppa simplified equations deduced in a simple way by algebraic method.

In addition to this new and very simplified way that leads to Kruppa equations, it is to be noted that the use of SVD (singular value decomposition) mainly allowed us to automatically deduce 3 equations among the 6 present in the original formula. The problem of choosing the 3 equations was solved either by the choice of a parametrization of epipolar geometry as in [FAU 90, MAY 92] or by arbitrarily taking an equation in order to estimate the scale factor and to reflect it on the rest by arbitrarily choosing 2 among the remaining 5 [BOU 98].

We can notice that in the case of a calibrated camera, the matrix of intrinsic parameters can be assumed to be equal to identity matrix 3×3. Consequently, considering the essential matrix $\mathbf{E}$ instead of the fundamental matrix $\mathbf{F}$, it is easy to show that equation (2.28) is reduced to $r^2 = s^2$, which implies that $r = s$, as shown in [HUA 89] and mentioned previously in section 2.3.

In the following section, we implement and test the method described until now to determine the intrinsic parameters of different cameras from various real images.

2.5. Implementation

In this section, we implement Kruppa equations by minimizing the non-linear criterion, which is associated to them. According to the approach described in [ZEL 96a, ZEL 96b], the equations derived in section 2.4.3 are treated within a non-linear optimization context and are solved in an iterative way. The choice of criterion, that of the initial conditions and the optimization method are presented and discussed.

2.5.1. *The choice of initial conditions*

Let $\mathbf{S_F} = [r, s, \mathbf{u_1}^t, \mathbf{u_2}^t, \mathbf{v_1}^t, \mathbf{v_2}^t]^t$ be the 14×1 vector formed by the SVD parameters of $\mathbf{F}$. Let $\frac{\rho_i(\mathbf{S_F},\mathbf{K})}{\phi_i(\mathbf{S_F},\mathbf{K})}$, $i = 1 \ldots 3$ be the 3 ratios defined by equation (2.28). Each pair of image defines a fundamental matrix, which makes it possible to write two polynomial equations in the parameters of $\mathbf{K}$:

$$\begin{aligned} \rho_1(\mathbf{S_F},\mathbf{K})\phi_2(\mathbf{S_F},\mathbf{K}) - \phi_1(\mathbf{S_F},\mathbf{K})\rho_2(\mathbf{S_F},\mathbf{K}) = 0 \\ \rho_1(\mathbf{S_F},\mathbf{K})\phi_3(\mathbf{S_F},\mathbf{K}) - \phi_1(\mathbf{S_F},\mathbf{K})\rho_3(\mathbf{S_F},\mathbf{K}) = 0 \end{aligned} \tag{2.29}$$

This system of equations is of degree two in 5 unknown (elements of the matrix $\mathbf{A}$). An initial approximation of the principal point in the image is given by the center of the image. Moreover, if angle θ is supposed to be equal to $\frac{\pi}{2}$, the number of unknown factors in (2.29) is reduced to two, i.e. the elements K_{11} and K_{22} of the matrix $\mathbf{K}$ connected to parameters α_u and α_v. Thus, the system of (2.29) becomes unknown pair by pair and can be solved analytically.

The system can have a maximum of $2^2 = 4$ solutions of which some can be eliminated by taking into account the constraints. Thus, K_{11} and K_{22} must be such that the matrix $\mathbf{K}$ is real and is defined positive and the *aspect ratio* or scale factor must be close to the unit. By having M images acquired with constant intrinsic parameters, $N \leq \frac{M(M-1)}{2}$ fundamental matrices can be estimated and produce N systems of the form (2.29), which have a maximum of $4N$ solutions for α_u and α_v. Several strategies can be used in order to choose good initial conditions for α_u and α_v among the set of $4N$ solutions. Three different strategies were tested (choice of an arbitrary solution, an average solution and a median solution), which enabled us to note that the various solutions were so close that they produced practically no variation on the results of non-linear optimization of the criterion, which we will present in the following section.

2.5.2. *Optimization*

Let $\pi_{ij}(\mathbf{S_F}, \mathbf{K})$ be the difference of ratios $\frac{\rho_i(\mathbf{S_F},\mathbf{K})}{\phi_i(\mathbf{S_F},\mathbf{K})} - \frac{\rho_j(\mathbf{S_F},\mathbf{K})}{\phi_j(\mathbf{S_F},\mathbf{K})}$ and let $\sigma^2_{\pi_{ij}}(\mathbf{S_F}, \mathbf{K})$ be its variance, approximated by[3]:

$$\sigma^2_{\pi_{ij}}\big(\mathbf{S_F}, \mathbf{K}\big) = \frac{\partial \pi_{ij}\big(\mathbf{S_F}, \mathbf{K}\big)}{\partial \mathbf{S_F}} \mathbf{\Lambda}_{\mathbf{S_F}} \frac{\partial \pi_{ij}\big(\mathbf{S_F}, \mathbf{K}\big)}{\partial \mathbf{S_F}}^t , \tag{2.30}$$

where $\mathbf{\Lambda}_{\mathbf{S_F}}$ is the 14×14 covariance matrix associated with $\mathbf{S_F}$ and $\frac{\partial \pi_{ij}(\mathbf{S_F},\mathbf{K})}{\partial \mathbf{S_F}}$ is the Jacobian of $\pi_{ij}(\mathbf{S_F}, \mathbf{K})$ in $\mathbf{S_F}$.

Since $\mathbf{S_F}$ is a function of $\mathbf{F}$, its covariance matrix $\mathbf{\Lambda}_{\mathbf{S_F}}$ is estimated by:

$$\mathbf{\Lambda}_{\mathbf{S_F}} = \frac{\partial \mathbf{S_F}}{\partial \mathbf{F}} \mathbf{\Lambda}_{\mathbf{F}} \frac{\partial \mathbf{S_F}}{\partial \mathbf{F}}^t , \tag{2.31}$$

3. If $\mathbf{x}$ is a random vector of average $\mathbf{x_0}$ and of covariance matrix $\mathbf{\Lambda_x}$, the vector matrix $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is equal in the first order to $\mathbf{\Lambda_y} = \frac{\partial \mathbf{f}(\mathbf{x_0})}{\partial \mathbf{x_0}} \mathbf{\Lambda_x} \frac{\partial \mathbf{f}(\mathbf{x_0})}{\partial \mathbf{x_0}}^t$; see [FAU 93] for more details.

where $\mathbf{\Lambda_F}$ is the 9×9 fundamental covariance matrix[4] and $\frac{\partial \mathbf{S_F}}{\partial \mathbf{F}}$ is the Jacobian value of $\mathbf{S_F}$ in $\mathbf{F}$. This last stage, i.e., the estimation of SVD derivatives, is explained in [PAP 00a, PAP 00b]. Variances $\sigma^2_{\pi_{ij}}(\mathbf{S_F}, \mathbf{K})$ are used for weighing residues $\pi_{ij}(\mathbf{S_F}, \mathbf{K})$ based on the confidence granted to equations.

Matrix $\mathbf{K}$ is the solution of the non-linear problem:

$$\mathbf{K} = \arg\min_{\tilde{\mathbf{K}}} \sum_{i=1}^{N} \left(\frac{\pi^2_{12}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})}{\sigma^2_{\pi_{12}}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})} + \frac{\pi^2_{13}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})}{\sigma^2_{\pi_{13}}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})} + \frac{\pi^2_{23}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})}{\sigma^2_{\pi_{23}}(\mathbf{S_{F}}_i, \tilde{\mathbf{K}})} \right) \quad (2.32)$$

As each fundamental matrix gives 2 independent equations and $\mathbf{K}$ has 5 unknown factors, a minimum of 3 pairs of views are necessary. Minimization of the sum of differences of square ratios $\pi_{ij}(\mathbf{S_F}, \mathbf{K})$ in (2.32) experimentally provided better results than the minimization of the sum of square polynomials of (2.29) [LUO 97]. Although the third equation (i.e., $\pi_{23}(\mathbf{S_F}, \mathbf{K})$) in (2.32) is dependent on the other two [LUO 97], it can also be added in the criterion.

Minimization of (2.32) is carried out using the Levenberg-Marquardt algorithm [AND 94], by using the initial solutions as estimated in the preceding part. In addition to the estimate of $\mathbf{K}$, the minimization procedure of (2.32) also makes it possible to estimate its covariance matrix [AND 94]. Matrix $\mathbf{A}$ is extracted from $\mathbf{K}$ in three steps: $\mathbf{A}^{-t}$ is estimated by Cholesky's decomposition of $\mathbf{K}^{-1}$, transposed and is finally inversed to obtain $\mathbf{A}$.

2.6. Experimental results

The approach presented was validated on synthetic and real results, and the estimate of matrix $\mathbf{A}$ was used to trace the measurements of 3D angle and the ratios of 3D segment lengths directly from a set of images. This, in addition to $\mathbf{A}$, requires an estimate of the homography of infinite plane $\mathbf{H}_\infty$ carried out by initially considering the essential matrix, from the associated fundamental matrix using (2.7), and by breaking it up into a rotation matrix $\mathbf{R}$ and translation $\mathbf{t}$, such that $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ (see [HAR 92a] for more details). Finally, $\mathbf{H}_\infty$ is estimated from $\mathbf{R}$ and $\mathbf{A}$.

4. This matrix is assumed to be given, for example, by the procedure described in [CSU 97].

2.6.1. *Estimation of angles and length ratios from images*

Knowledge of $\mathbf{A}$ enables us to obtain 3D Euclidean measurements from images. Thus, the angle between two 3D segments L_1 and L_2 can be estimated as follows [ZEL 96a, ZEL 96b].

Let $\mathbf{l_1}$, $\mathbf{l'_1}$ and $\mathbf{l_2}$, $\mathbf{l'_2}$ be the projections of L_1 and L_2 in 2 images. Using Laguerre's formula, the cosine of the angle between L_1 and L_2 is given by:

$$\cos\big(L_1, L_2\big) = \frac{\big|S\big(\mathbf{v_1}, \mathbf{v_2}\big)\big|}{\sqrt{S\big(\mathbf{v_1}, \mathbf{v_1}\big)S\big(\mathbf{v_2}, \mathbf{v_2}\big)}}, \tag{2.33}$$

where $\mathbf{v_1}$ and $\mathbf{v_2}$ are the projections in the first image of the intersections of L_1 and L_2 with the *infinite plane*, and:

$$S\big(\mathbf{m}, \mathbf{n}\big) = \mathbf{m}^t \mathbf{A}^{-t} \mathbf{A}^{-1} \mathbf{n} \tag{2.34}$$

Points $\mathbf{v_1}$ and $\mathbf{v_2}$ are determined by:

$$\mathbf{v_1} = \mathbf{l_1} \times {\mathbf{H}_\infty}^t \mathbf{l'_1}, \quad \mathbf{v_2} = \mathbf{l_2} \times {\mathbf{H}_\infty}^t \mathbf{l'_2}, \tag{2.35}$$

where $\mathbf{H}_\infty$ is the *homography of infinite plane*, which is given by $\mathbf{H}_\infty = \mathbf{A}\mathbf{R}\mathbf{A}^{-1}$.

In addition to angle measurements, length ratios can also be estimated. Let A, B, C and D be four 3D points, whose corresponding 2D projections are: (a, a'), (b, b'), (c, c') and (d, d'). Therefore, we can show [ZEL 96a] that the length ratio of 3D segments AB and CD is given by:

$$\begin{aligned}\frac{AB}{CD} = &\sqrt{\frac{S\big(V_{ab}, V_{ab}\big)}{S\big(V_{cd}, V_{cd}\big)} \cdot \frac{S\big(V_{ac}, V_{ac}\big)S\big(V_{bc}, V_{bc}\big) - S^2\big(V_{ac}, V_{bc}\big)}{S\big(V_{ab}, V_{ab}\big)S\big(V_{ac}, V_{ac}\big) - S^2\big(V_{ab}, V_{ac}\big)}} \\ &\cdot \sqrt{\frac{S\big(V_{bd}, V_{bd}\big)S\big(V_{cd}, V_{cd}\big) - S^2\big(V_{bd}, V_{cd}\big)}{S\big(V_{bc}, V_{bc}\big)S\big(V_{bd}, V_{bd}\big) - S^2\big(V_{bc}, V_{bd}\big)}},\end{aligned} \tag{2.36}$$

where:

$$\mathbf{v}_{ij} = \mathbf{l}_{ij} \times \mathbf{H}_\infty^t \mathbf{l}'_{ij}, \quad ij \in \{ab, ac, bc, bd, cd\} \tag{2.37}$$

is the projection in the 1st image of intersection points of the straight line defined by 3D points I and J and infinite plane, and $\mathbf{l}_{ij}$ is the straight line defined by image points i and j, given by vectorial product $i \times j$. Operator $S(\cdot, \cdot)$ is defined by (2.34).

2.6.2. *Experiments with synthetic data*

In this section, we give some experimental results drawn from noisy synthetic data. These results particularly illustrate the sensitivity to noise of the measured angles and the 3D length ratios calculated from estimated images and intrinsic parameters.

Three rigid displacements of cameras were simulated and 300 3D points randomly distributed according to Gaussian law were projected on the four positions of simulated cameras. Various values of Gaussian noise were considered. The six fundamental matrices were estimated and their SVD were calculated. Minimization of (2.32) was carried out with 4 unknown factors, i.e., the angle between the axes was assumed to be known and equal to $\pi/2$. The intrinsic parameters of the simulated camera are given in Table 2.1. The estimated parameters are given in Table 2.2. The left column indicates the standard variation of the noise added to points in the image, whereas the second column indicates the intrinsic parameters estimated by the method presented. In order to validate the approach within the required context as that of 3D measurement, we give an analysis of the errors made by estimating 3D angles and length ratios carried out with the help of a few segments connecting synthetic 3D points. 100 angles and 100 length ratios were thus calculated from a set of randomly selected synthetic 3D points. We then used noisy projections of the 3D points used, as well as the intrinsic parameters estimated from images to find the angles and the length ratios of the points selected from (2.33). The average and standard variations of the relative error between the actual values and the estimated values are summarized in Table 2.2. The third column of this table contains angle measurements whereas the fourth column relates to length ratios. The values in between

840	0	310
0	770	270
0	0	1

Table 2.1. *Real intrinsic parameters of* **A**

brackets relate to errors when we use intrinsic parameters and non-estimated real fundamental matrices. The only source of error considered is then that associated with noisy 2D points. To some extent, these values represent a limit beyond which it is impossible to proceed since in this case $\mathbf{A}$ and $\mathbf{F}$ are perfectly known. Table 2.2 makes it possible to verify that the introduction of estimated parameters in the measuring process increases the error slightly, which validates the estimate of parameters using the proposed method.

Std. Var. Noise	Estimated Matrix $\mathbf{A}$			Angles Rel. Err. Ave.	Angles Rel. Err. Std. Var.	Length ration Rel. Err. Ave.	Length ration Rel. Err. Std. Var.
0.0	840.732 0 0	−1.44521e-14 770.606 0	310.058 270.606 1	0.0038 (4.95e-06)	0.0064 (7.95e-06)	0.0192 (1.94e-05)	0.0705 (6.32e-05)
0.1	840.216 0 0	0 770.008 0	310.905 267.742 1	0.0076 (0.0059)	0.0130 (0.0099)	0.0190 (0.0200)	0.0438 (0.0505)
0.5	828.684 0 0	1.39949e-14 769.914 0	320.451 262.45 1	0.0466 (0.0449)	0.0726 (0.0776)	0.0972 (0.0924)	0.1748 (0.1921)
1.0	836.937 0 0	0 752.633 0	329.008 289.929 1	0.0971 (0.0999)	0.2534 (0.3480)	0.1290 (0.1336)	0.2738 (0.2864)
1.5	830.949 0 0	2.81749e-14 770.784 0	327.062 261.244 1	0.1220 (0.1034)	0.2895 (0.2397)	0.1657 (0.1716)	0.2789 (0.3125)

Table 2.2. *Results of synthetic data*

2.6.3. *Experiments with real data*

The following results were obtained from a dataset of fundamental matrices estimated on several sequences of real images taken by a traditional camera in Santorini island (house sequence and Orthodox Church sequence) and in Crete (sequence of Knossos Palace) in Greece.

In all these sequences, the fundamental matrices were estimated from points paired by the method described in [ZHA 95] and the angle θ was assumed to be known and equal to $\pi/2$; the matrix $\mathbf{K}$ in (2.32) is thus parametrized with four unknown factors. More experimental results can be found in [LOU 99, LOU 00b].

(a) (b)

Figure 2.2. *Images 4 and 5 of the house sequence*

The first experiment uses eight images of size 709×429 of a house at Santorini in Greece. Images 4 and 5 of this sequence are shown in Figures 2.2a and b. Self-calibration was carried out using 28 fundamental matrices defined by all pairs of images. The matrix of the intrinsic parameters estimated is shown in Table 2.3. Figures 2.3a and b illustrate the line segments used to calculate the angles indicated in Table 2.4. More precisely, the left column indicates the pair of line segments used to define the angle, the second column indicates the real value of the angle, the third column gives the cosine of the angle estimated from images by (2.33) and finally, the fourth column gives the angles corresponding to the estimated cosines. The estimated intrinsic parameters also help to calculate the length ratios using the segments shown in Figures 2.3c and d. The estimated length ratios are shown in Table 2.5. The left column indicates the pair of line segments used, the middle column gives the true ratio values and the right column gives the sizes estimated from images using (2.36). Finally, Figure 2.4 illustrates various views of a 3D model of the house. This model was reconstructed from the estimate of intrinsic parameters of the camera.

709.32	−1.96171e-14	353.646
0	736.494	183.207
0	0	1

Table 2.3. *House sequence: matrix of estimated intrinsic parameters*

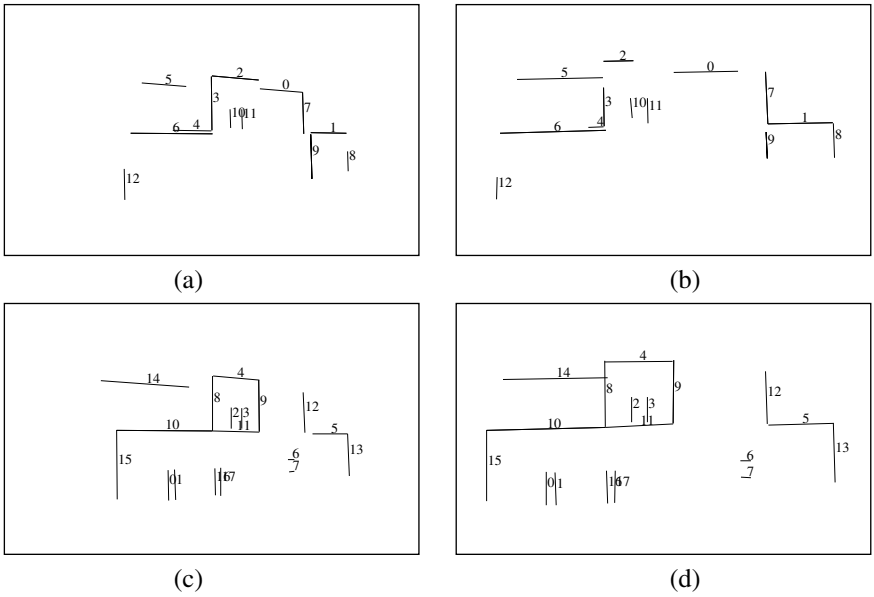


Figure 2.3. *Line segments of the house sequence: (a)–(b) segments used for measuring angles, (c)–(d) segments used for measuring length ratios*

Line segments of Figures 2.3a–b

Angle segments	Real angle (deg)	Estimated cosine	Estimated angle (deg)
0 – 1	0	0.991603	7.430145
1 – 2	0	0.995566	5.397296
2 – 3	90	0.0187033	88.928315
3 – 4	90	0.00298511	89.828965
5 – 6	0	0.999325	2.106069
7 – 8	0	0.999797	1.153783
7 – 9	0	0.999875	0.906795
10 – 11	0	0.99365	6.460356
9 – 12	0	0.999659	1.495708
6 – 9	90	0.0194928	88.883073

Table 2.4. *House sequence: in situ observation and angle estimation*

Line segments of Figures 2.3c–d

Line Segment	Real length ratio	Estimated length ratio
0 – 1	1.0	0.9352
2 – 3	1.0	1.11188
4 – 5	1.0	0.866016
6 – 7	1.0	0.994141
8 – 9	1.0	1.40652
10 – 11	2.0	2.06752
12 – 13	1.0	0.950633
10 – 14	1.0	1.04678
8 – 15	1.0	0.875412
16 – 17	1.0	1.02571

Table 2.5. *House sequence: in situ observation and estimated length ratios*

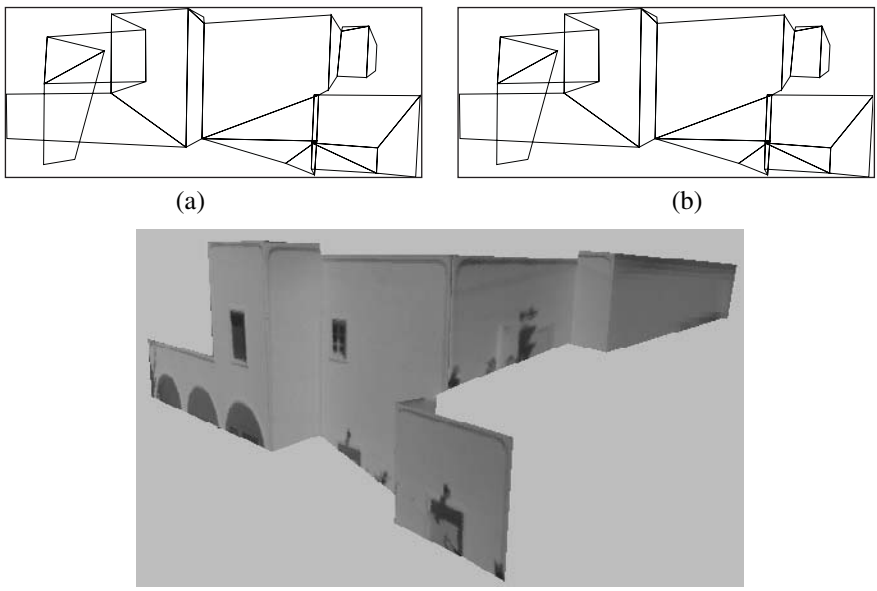


Figure 2.4. *Top and side views of the 3D model reconstructed from the house sequence and the estimated intrinsic parameters. Note the orthogonality and the parallelism of the reconstructed walls. Images (a) and (b) show the 3D positions of the camera used and (c) includes the texture in the reconstructed surfaces*

(a) (b)

Figure 2.5. *Images 1 and 2 of the Orthodox Church sequence*

In the second experiment, we treat the six images of a sequence representing an Orthodox Church at Santorini in Greece. Images 1 and 2 of the sequence are illustrated in Figures 2.5a and b. These images are of dimensions of 709×429 pixels. Self-calibration was carried out using 16 fundamental matrices defined by a set of images. The matrix of the estimated parameters is given in Table 2.6. Tables 2.7 and 2.8 give angles

604.926	0	377.86
0	712.944	313.882
0	0	1

Table 2.6. *Church sequence: matrix of estimated intrinsic parameters*

Line segments of Figures 2.6a–b

Angle segments	True estimate (deg)	Angle Estimate	Cosine angle (deg)
0 – 1	0	0.999697	1.411016
1 – 2	90	0.051252	87.062189
3 – 4	0	0.999969	0.448352
5 – 6	90	0.0126704	89.274019
1 – 6	0	0.999958	0.524615
7 – 8	90	0.00205859	89.882051
9 – 10	0	0.99978	1.202867
2 – 11	90	0.0227189	88.698188
2 – 12	90	0.048568	87.216162
1 – 13	90	0.0302124	88.268694
3 – 14	90	0.0859321	85.070376
15 – 16	90	0.0640689	86.326605

Table 2.7. *Church sequence: in situ observation and estimated angles*

Line segments of Figures 2.6c–d

Line segments	Real length ratio	Estimated length ratio
0 – 1	1.0	1.07089
2 – 3	1.0	1.11651
4 – 5	1.0	0.92611
6 – 7	1.0	1.04813
8 – 9	1.0	0.861372
10 – 11	1.0	1.02708
12 – 13	1.0	0.936046
14 – 15	1.0	0.960333
16 – 17	1.0	1.05697
18 – 19	1.0	0.930897

Table 2.8. *Church sequence: in situ observation and estimated length ratios*

and length ratios calculated from images estimated, intrinsic parameters and line segments in Figures 2.6a–b and c–d, respectively. Two views of the reconstructed 3D model are shown in Figure 2.7.

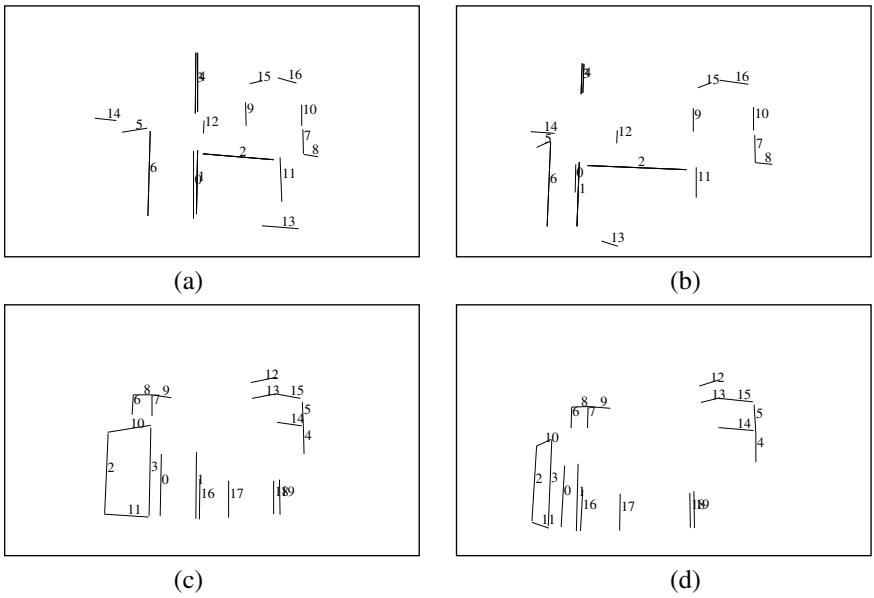
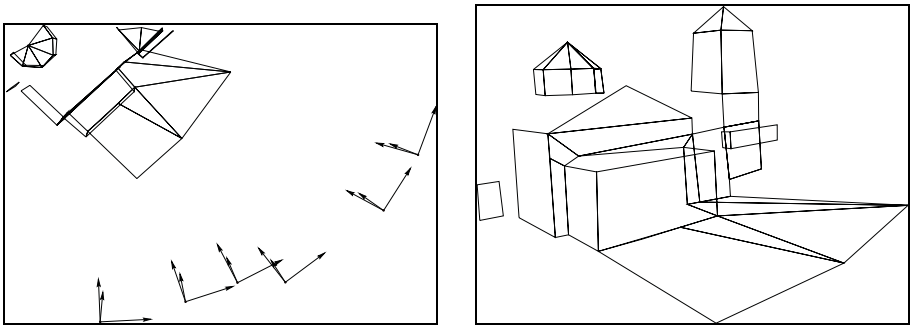


Figure 2.6. *Line segments of the Orthodox Church sequence: (a)-(b) line segments used for the measurement of 3D angles, (c)-(d) line segments used for the measurement of 3D length ratios*



(a) (b)

Figure 2.7. *The top and front views of the 3D model of the church reconstructed from the estimated intrinsic parameters: (a) shows the camera positions. Note the orthogonality and parallelism of the walls and the isosceles triangle which forms the dome*

Figure 2.8. *Images 1 and 2 of Knossos sequence*

The third experiment relates to a triplet of images of the Palace of Knossos in Crete. Two images of size 880×579 are illustrated in Figures 2.8. Self-calibration was carried out using three associated fundamental matrices. Figure 2.9 shows several views of the 3D model reconstructed with and without texture.

2.7. Conclusion

In this chapter, we were interested in the fundamental problem of self-calibration of a camera from a set of points paired between various images. We mainly focused on the well-known Kruppa equation method. After an introduction to these equations and a reminder regarding how to obtain them by geometric approach and its algebraic equivalent, we presented

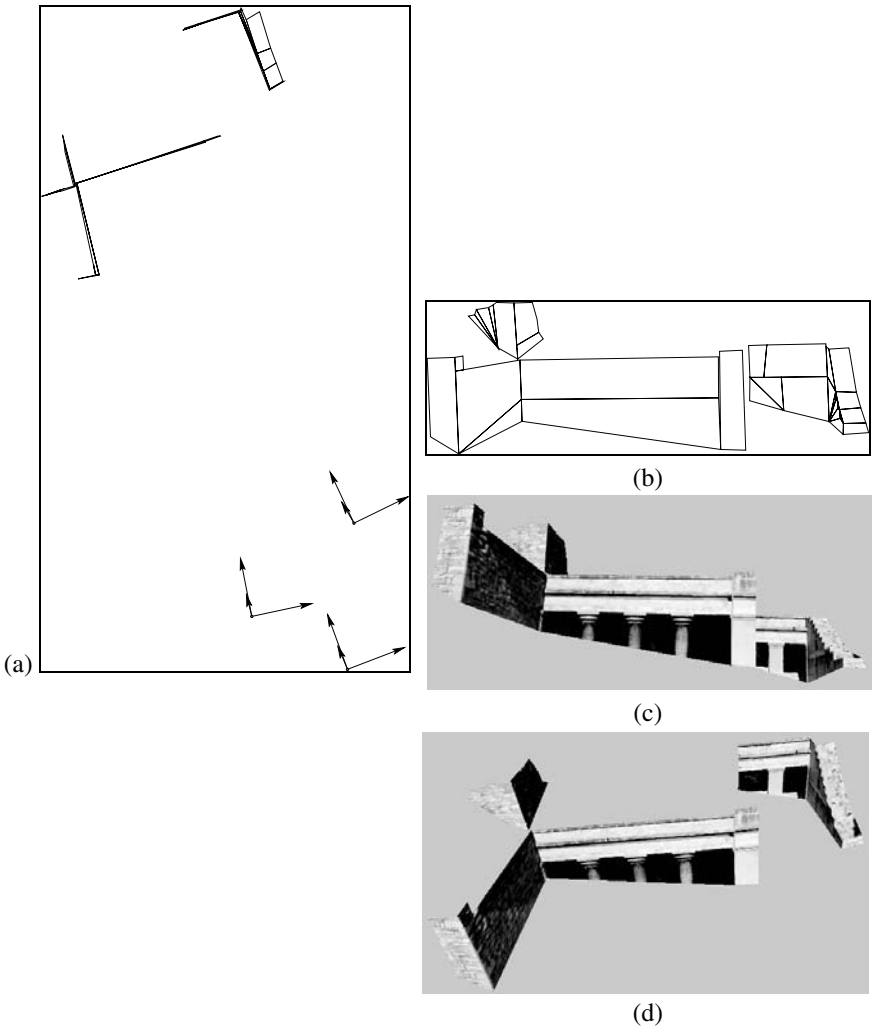


Figure 2.9. *Top and side views of the 3D Knossos model reconstructed using estimated intrinsic parameters: (a) and (b) show the top and side views of the model, with (a) illustrating the 3D positions of the camera. (c) and (d) show the two side views with texture*

new simplified Kruppa equations with the help of the decomposition of the fundamental matrix into singular values. Finally and mainly, this method was implemented, compared and successfully tested to find the intrinsic parameters of various cameras from noisy synthetic data and several real images. The quality of results obtained enabled us to validate the approach until reliable 3D measurements were obtained from images.

2.8. Acknowledgement

The author wishes to thank Manolis Lourakis for his invaluable assistance in obtaining the results which illustrate this chapter.

2.9. Bibliography

[AGA 98] DE AGAPITO L., HAYMAN E. and REID I.L., "Self-calibration of a rotating camera with varying intrinsic parameters", *British Machine Vision Conference*, Southampton, UK, BMVA Press, September 1998.

[AGA 99] DE AGAPITO L., HARTLEY R. and HAYMAN E., "Linear Calibration of a Rotating and Zooming Camera", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, vol. 1, Fort Collins, Colorado, IEEE Computer Society, p. 15–21, June 1999.

[AND 94] ANDERSON E., BAI Z., BISHOF C., DEMMEL J., DONGARRA J., CROZ J.D., GREENBAUM A., HAMMARLING S., MCKENNEY A., OSTROUCHOV S. and SORENSEN D., *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, PA 19104-2688, 2nd ed., 1994.

[ARM 96] ARMSTRONG M., ZISSERMAN A. and HARTLEY R., "Self-calibration from Image Triplets", *Fourth European Conference on Computer Vision*, p. 3–16, April 1996.

[BOU 98] BOUGNOUX S., "From Projective to Euclidean Space under any Practical Situation, a Criticism of Self-Calibration", *IEEE International Conference on Computer Vision*, p. 790–796, 1998.

[BRO 71] BROWN D.C., "Close-Range Camera Calibration", *Photogrammetric Engineering*, vol. 37, no. 8, p. 855–866, 1971.

[BRO 96] BROOKS M.J., DE AGAPITO L., HUYNG D.Q. and BAUMELA L., "Direct Methods for Self-Calibration of a Moving Stereo Head", *Fourth European Conference on Computer Vision*, vol. II, p. 415–426, April 1996.

[CSU 97] CSURKA G., ZELLER C., ZHANG Z. and FAUGERAS O., "Characterizing the Uncertainty of the Fundamental Matrix", *CVGIP: Image Understanding*, vol. 68, no. 1, p. 18–36, October 1997.

[DEV 96] DEVERNAY F. and FAUGERAS O., "From Projective to Euclidean Reconstruction", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, San Francisco, USA IEEE, p. 264–269, June 1996.

[ENC 94] ENCISO R., VIÉVILLE T. and FAUGERAS O., "Approximation du Changement de Focale et de Mise au Point par une Transformation Affine à Trois Paramètres", *Traitement du Signal*, vol. 11, no. 5, p. 361–372, 1994.

[FAI 75] FAIG W., "Calibration of Close-Range Photogrammetry Systems: Mathematical Formulation", *Photogrammetric Engineering and Remote Sensing*, vol. 41, no. 12, p. 1479–1486, 1975.

[FAU 86] FAUGERAS O. and TOSCANI G., "The Calibration Problem for Stereo", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Miami Beach, p. 15–20, June 1986.

[FAU 90] FAUGERAS O. and MAIBANK S., "Motion from point matches: multiplicity of solutions", *The International Journal of Computer Vision*, vol. 4, no. 3, p. 225–246, 1990.

[FAU 92] FAUGERAS O., "What can be seen in three dimensions with an uncalibrated stereo rig?", *Proceedings of the 2nd ECCV*, p. 563–578, May 1992.

[FAU 93] FAUGERAS O., *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, 1993.

[FAU 95] FAUGERAS O., "Stratification of 3-D vision: projective, affine, and metric representations", *Journal of the Optical Society of America A*, vol. 12, no. 3, p. 465–484, March 1995.

[GOL 89] GOLUB G. and LOAN C.V., *Matrix Computations*, The John Hopkins University Press, Baltimore, 2nd ed., 1989.

[HAR 92a] HARTLEY R.I., "Estimation of Relative Camera Positions for Uncalibrated Cameras", in SANDINI G. (Ed.), *Proceedings of the 2nd European Conference on Computer Vision*, Santa Margherita, Italy, Springer-Verlag, p. 579–587, May 1992.

[HAR 92b] HARTLEY R., GUPTA R. and CHANG T., "Stereo from Uncalibrated Cameras", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Urbana Champaign, IL, IEEE, p. 761–764, June 1992.

[HAR 97a] HARTLEY R., "Self-Calibration of Stationary Cameras", *The International Journal of Computer Vision*, vol. 22, no. 1, p. 5–24, February 1997.

[HAR 97b] HARTLEY R.I., "Lines and points in three views and the trifocal tensor", *The International Journal of Computer Vision*, vol. 22, no. 2, p. 125–140, March 1997.

[HEY 96] HEYDEN A. and ASTRÖM K., "Algebraic varieties in multiple view geometry", *Fourth European Conference on Computer Vision*, vol. II, p. 671–682, 1996.

[HEY 97] HEYDEN A. and ÅSTRÖM K., "Euclidean reconstruction from image sequences with varying and unknown focal length and principal point", *Comp. Vision and Pattern Rec.*, IEEE Computer Society Press, p. 438–443, 1997.

[HOR 94] HORAUD R., DORNAIKA F., BOUFAMA B. and MOHR R., "Self-Calibration of a Stereo Head Mounted onto a Robot Arm", in EKLUNDH J. (Ed.), *Proceedings of the 3rd European Conference on Computer Vision*, Stockholm, Sweden, Springer-Verlag, p. 455–462, 1994.

[HOR 98] HORAUD R. and CSURKA G., "Self-Calibration and Euclidean Reconstruction Using Motions of a Stereo Rig", *Proceedings of the 6th International Conference on Computer Vision*, Bombay, India, IEEE Computer Society, IEEE Computer Society Press, January 1998.

[HUA 89] HUANG T.S. and FAUGERAS O.D., "Some Properties of the E Matrix in Two-View Motion Estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, p. 1310–1312, 1989.

[LOU 99] LOURAKIS M.I. and DERICHE R., Camera Self-Calibration Using the Singular Value Decomposition of the Fundamental Matrix: From Point Correspondences to 3D Measurements, Research Report no. 3748, INRIA Sophia-Antipolis, August 1999.

[LOU 00a] LOURAKIS M.I. and DERICHE R., Camera Self-Calibration Using the Kruppa Equations and the SVD of the Fundamental Matrix: The Case of Varying Intrinsic Parameters, Research Report no. 3911, INRIA Sophia-Antipolis, 2000.

[LOU 00b] LOURAKIS M.I. and DERICHE R., "Camera Self-Calibration Using the Singular Value Decomposition of the Fundamental Matrix", *Proc. of the 4th Asian Conference on Computer Vision*, vol. I, p. 403–408, January 2000.

[LUO 92] LUONG Q.-T., Matrice Fondamentale et Calibration Visuelle sur l'Environnement-Vers une plus grande autonomie des systèmes robotiques, PhD Thesis, University of Paris-Sud, Center d'Orsay, December 1992.

[LUO 97] LUONG Q.-T. and FAUGERAS O., "Self-Calibration of a Moving Camera from Point Correspondences and Fundamental Matrices", *The International Journal of Computer Vision*, vol. 22, no. 3, p. 261–289, 1997.

[LUO 98] LUONG Q.-T. and FAUGERAS O., "On the determination of epipoles using cross-ratios", *CVGIP: Image Understanding*, vol. 71, no. 1, p. 1–18, July 1998.

[MAY 92] MAIBANK S.J. and FAUGERAS O.D., "A Theory of Self-Calibration of a Moving Camera", *The International Journal of Computer Vision*, vol. 8, no. 2, p. 123–152, 1992.

[MOH 93] MOHR R., "Projective Geometry and Computer Vision", in CHEN C.H., PAU L.F. and WANG P.S.P. (Eds.), *Handbook of Pattern Recognition and Computer Vision*, Chapter 2.4, p. 369–393, World Scientific Publishing Company, 1993.

[MUN 92] MUNDY J. L. and ZISSERMAN A., *Geometric Invariance in Computer Vision*, MIT Press, 1992.

[PAP 00a] PAPADOPOULO T. and LOURAKIS M., Estimating the Jacobian of the Singular Value Decomposition: Theory and Applications, Research report, INRIA Sophia-Antipolis, 2000.

[PAP 00b] PAPADOPOULO T. and LOURAKIS M., "Estimating the Jacobian of the Singular Value Decomposition: Theory and Applications", *Proc. of the 6th European Conference on Computer Vision*, June 2000.

[POL 97a] POLLEFEYS M. and GOOL L.V., "Self-calibration from the Absolute Conic on the Plane at Infinity", *Proc. CAIP'97*, LNCS vol.1296, Kiel, Germany, Springer-Verlag, p. 175–182, 1997.

[POL 97b] POLLEFEYS M. and VAN GOOL L., "A Stratified Approach to Metric Self-Calibration", *IEEE International Conference on Computer Vision and Pattern Recogition*, p. 407–412, 1997.

[POL 98] POLLEFEYS M., KOCH R. and GOOL L.V., "Self-Calibration and Metric Reconstruction in Spite of Varying and Unknown Internal Camera Parameters", *IEEE International Conference on Computer Vision*, p. 90–95, 1998.

[POL 99] POLLEFEYS M., KOCH R. and GOOL L.V., "Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters", *The International Journal of Computer Vision*, vol. 32, no. 1, p. 7–25, August 1999.

[SLA 80] SLAMA C.C. (Ed.), *Manual of Photogrammetry*, American Society of Photogrammetry, 4th ed., 1980.

[STU 97] STURM P., "Critical motion sequences for monocular self-calibration and uncalibrated Euclidean reconstruction", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, p. 1100–1105, 1997.

[TOS 87] TOSCANI G., Système de Calibration optique et perception du mouvement en vision artificielle, PhD Thesis, Paris-Orsay, 1987.

[TRI 88] TRIVEDI H.P., "Can Multiple Views Make up for Lack of Camera Registration", *Image and Vision Computing*, vol. 6, no. 1, p. 29–32, February 1988.

[TSA 86] TSAI R., "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Miami Beach, p. 364–374, June 1986.

[TSA 87] TSAI R.Y., "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses", *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, p. 323–344, 1987.

[VIÉ 94] VIÉVILLE T., "Autocalibration of Visual Sensor Parameters on a Robotic Head", *Image and Vision Computing*, vol. 12, 1994.

[VIÉ 96a] VIÉVILLE T. and FAUGERAS O., "The First Order Expansion of Motion Equations in the Uncalibrated Case", *CVGIP: Image Understanding*, vol. 64, no. 1, p. 128–146, July 1996.

[VIÉ 96b] VIÉVILLE T., FAUGERAS O.D. and LUONG Q.-T., "Motion of Points and Lines in the Uncalibrated Case", *The International Journal of Computer Vision*, vol. 17, no. 1, p. 7–42, January 1996.

[VIÉ 99] VIÉVILLE T. and LINGRAND D., "Using Specific Displacements to analyze Motion without Calibration", *IJCV*, vol. 31, no. 1, p. 5–29, 1999.

[WOL 83] WOLF P., *Elements of Photogrammetry*, McGraw-Hill, New York, 1983.

[ZEL 52] ZELLER M., *Textbook of Photogrammetry*, H.K. Lewis & Company, London, 1952.

[ZEL 96a] ZELLER C., Calibration Projective Affine et Euclidienne en Vision par Ordinateur, PhD Thesis, École Polytechnique, February 1996.

[ZEL 96b] ZELLER C. and FAUGERAS O., Camera Self-Calibration from Video Sequences: the Kruppa Equations Revisited, Research Report no. 2793, INRIA, February 1996.

[ZHA 93] ZHANG Z., LUONG Q.-T. and FAUGERAS O., Motion of an Uncalibrated Stereo Rig: Self-Calibration and Metric Reconstruction, Research Report no. 2079, INRIA Sophia-Antipolis, 1993.

[ZHA 95] ZHANG Z., DERICHE R., FAUGERAS O. and LUONG Q.-T., "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry", *Artificial Intelligence Journal*, vol. 78, p. 87–119, October 1995.

[ZIS 98] ZISSERMAN A., LIEBOWITZ D. and ARMSTRONG M., "Resolving Ambiguities in Auto-Calibration", *Philosophical Transactions of the Royal Society of London, SERIES A*, vol. 356, no. 1740, p. 1193–1211, 1998.

Chapter 3

Specific Displacements for Self-calibration

3.1. Introduction: interest to resort to specific movements

Position of the problem

Let us consider the observation of three-dimensional scenes in which one or several fixed objects are moving. Our point of view is that of a single camera, which can move in this scene. Generally this camera will not be calibrated and will be able to change its focal length, orientation, focus and aperture, thus all visual sensor parameters.

Thus, we are interested in the study of the movement of such cameras or that of observed objects. This also involves the analysis of the structure of three-dimensional scenes and the modeling of these cameras in terms of equations. More precisely, we are interested in recovering the scene structure, object/camera movement and camera calibration using singular cases, i.e. taking specific cases into account.

This study can be applicable to several cameras: it is sufficient to consider that the change in camera corresponds to a spatial displacement of the sensor. However, a more detailed study would be necessary if several cameras are simultaneously used to film the same scene.

Chapter written by Diane LINGRAND, François GASPARD and Thierry VIÉVILLE.

Use of specific cases

There are specific cases for which general methods for the estimation of movement do not apply. For example, in the case of a purely rotating object, the principal matrix discussed in the previous chapters does not exist and hence methods based on its estimation cannot be used. In this case, we are interested in another quantity, homography, which connects the points from one image to another.

In practice, a number of video sequences present specific cases: a pedestrian or a car move while remaining in contact with the ground (thus yielding a surface movement or planar movement), a monitoring camera is generally fixed, a robot moves only according to certain fixed degrees of freedom, etc. (thus without movement).

On the other hand, such particular cases can provide simplified equations making it possible to obtain information that cannot be observed in general: hence, we can, on the one hand, determine the parameters that generally cannot be determined and, on the other hand, obtain simpler equations which enable a more exact numerical estimation. In practice, we are almost always in the presence of, or close to, a particular case of a system whose movements are controlled (vehicle, robot, camera placed on a support).

Similarly, even in a non-critical particular case, the result is better with a simplified model, consisting of lesser parameters than that with a more general model [VIÉ 99].

Study of singularities

The study of specific cases is thus motivated by two factors: the first concerns the simplification of equations and reduction of parameters to be estimated, and the second is the fact that in a specific case, erroneous data can completely defeat the estimation. Thus, as explained in [TOR 97], considering the line interpolation from a set of almost colinear points, some outliers can completely bias the estimation. On the other hand, certain authors, for example [TRI 98a], have noticed that many cameras and geometry of scenes correspond to degenerated cases and it is necessary to study and exploit the properties of these specific cases in order to ensure the general method stability.

Definitions of such “degenerated cases” or of “degenerated data” have been given by different authors. Torr [TOR 95] considers data that does not allow

us to determine the solution to the movement problem in a unique way as degenerated. As for Sturm [STU 97b], he considers that a set of images is critical when there are several solutions to the Euclidean calibration problem (in this case, it shows that the hyperplane at infinity cannot be identified in a single way and that the affine reconstruction cannot be carried out).

Sturm [STU 97a] carried out a complete study of critical cases and an exhaustive study concerning all the specific cases was recently carried out by Lingrand [LIN 99].

3.2. Modeling: parametrization of specific models

3.2.1. *Specific projection models*

The most commonly used model is the perspective projection model. However, different approximations have been used, especially by Aloimonos [ALO 90], Dementhon [DEM 89], Horaud and Christy [HOR 94, HOR 97], and Soatto and Perona [SOA 95].

Poelman and Kanade [POE 94] used the para-perspective model on sequence of images and found better results than for the orthographic model when the translation movement along the optical axis or orthogonal to the optical axis is significant. However, it is necessary to know the optical center as well as the focal distance for the para-perspective model. If these parameters are not known, then the orthographic model is equivalent to it.

Boufama and Weinshall [BOU 94] carried out a comparison of orthographic and perspective projection models with algorithms using invariants on three sequence of images with different view field sizes. The orthographic model proves to be more effective for a small view field and the perspective model is more effective for a bigger view field; thus, authors use an algorithm combining these two approaches. This type of approach by stratification has been used by other authors for movement. Lingrand and Viéville [LIN 95] used the orthographic model in cases where movement of the camera preserves its retinal plane. Quan [QUA 96] used the same model, but for sequence of images.

From the works of Lingrand [LIN 99], we can summarize the principle of perspective projection and its diverse approximations presented in Figure 3.1. Note that the latter split into two stages comprising the projection of 3D points

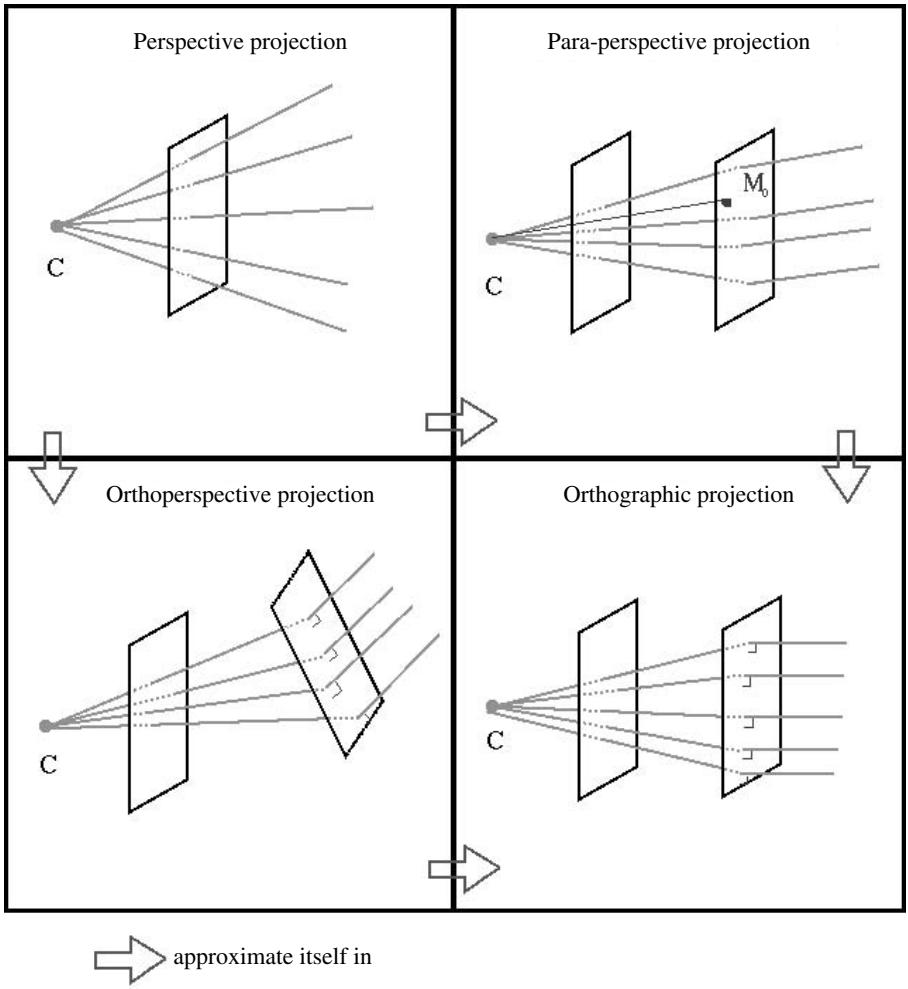


Figure 3.1. *Perspective projection and its various approximations. The arrows indicate the hierarchy existing between these different approximations*

for the observed object in an auxiliary plane passing through the "mean" point, followed by a central projection:

– *Perspective projection*: projection modeled using a pinhole assumption; all light rays pass through the optical center of the sensor.

– *Para-perspective projection*: projection of 3D points in the direction of observation from the average point on the auxiliary plane parallel to the image plane.

– *Orthoperspective projection*: orthogonal projection on the auxiliary plane for which the normal is given by the direction of observation of the average point.

– *Orthographic projection to scale*: orthogonal projection on the auxiliary plane parallel to the image plane.

The perspective model is the most general one. It can be approximated under similar experimental conditions in a para-perspective or orthoperspective projection. These two approximations can be approximated by orthographic projection.

Taking into account the poor algebraic simplification of the perspective projection, this is used, see for example [ALO 90], only in the case of its orthographic approximation. On the other hand, some authors, having developed algorithms for specific cases in orthographic projection [TOM 91, POE 93, KOE 84], have succeeded in generalizing their methods to para-perspective cases, which was not possible for orthoperspective cases. Finally, since orthoperspective projection has the same validity field as para-perspective projection and is more complicated than the latter, we consider only para-perspective or orthographic approximations in the following.

The para-perspective model has the advantage of making the projection equations linear but has the disadvantage of introducing two additional parameters. The orthographic model combines the advantages of making the equations linear without introducing additional parameters.

These two models are related to the scene through a mean depth point but preserve both scale factors (close objects appear bigger than far away objects).

However, the para-perspective model remains more precise than the orthographic model, as it takes into account the position of objects in the periphery (seen through a different angle).

Lingrand [LIN 99] proposes a general projection model, which models integrate all

$$\kappa\,\mathbf{m} = \begin{pmatrix} \hat{\alpha_u} & \gamma & \lambda\hat{\beta}_u + \mu u_0 & (1-\mu)\hat{u_0} \\ 0 & \hat{\alpha_v} & \lambda\hat{\beta}_v + \mu v_0 & (1-\mu)\hat{v_0} \\ 0 & 0 & \mu & (1-\mu) \end{pmatrix} \mathbf{M} = \mathbf{A}\mathbf{M} \qquad (3.1)$$

In this model, we distinguish two types of parameters:

– Modal parameters: λ and μ.

– Intrinsic parameters: α_u, α_v, β_u, β_v, u_0, v_0 and γ.

The modal parameters $\lambda \in \{0, 1\}$ and $\mu \in \{0, 1\}$ determine the projection model.

	λ	μ
Pure perspective projection	1	1
Orthographic projection	0	0
Para-perspective projection	1	0

3.2.2. *Specifications of internal parameters of the camera*

The projection models are parametrized by intrinsic parameters, which are assumed constant by most of the self-calibration methods. However, in active vision applications, this is not correct anymore [VIÉ 94b]. However, different remarks can be made.

Enciso [ENC 95] has experimentally shown that the ratio $\frac{\alpha_u}{\alpha_v}$ of the scale factors is constant for ordinary cameras (cameras, camescopes, etc.) and that the orthogonal parameter γ can be taken as negligible. Willson [WIL 94] has confirmed this property for visual sensors of high quality. Luong and Viéville [LUO 96], Tomasi and Kanade [TOM 92] as well as Zisserman and Liebowitz [ZIS 98] have removed the orthogonal parameter from their works. Heyden and Åström [HEY 97] have used these properties and have thus shown that by knowing the focus expansion and the orthogonal parameter, we can carry out a reconstruction with close similarities. Pollefeys and Van Koch [POL 97] have generalized these results by proving that self-calibration is possible in cases where the intrinsic parameters vary as soon as we know that $\gamma = 0$. Moreover, for Pollefeys and Van Koch, to consider that all intrinsic parameters are constant is not a realistic hypothesis whereas to consider that they are all variable is too general. Hence, in [POL 97], they always use the hypothesis of the knowledge of certain intrinsic parameters or the linear relations between them, from Kruppa equations [FAU 93], obtaining a relation between the singular values of constraints: these have to be non-zero in order to avoid any singularity.

Studies on the variation of focal length and zoom ([BOB 94], [ENC 95] and [GAS 97]) have highlighted that the coordinates of a main point can be considered as constants during weak focal length variations but that it was difficult to model the displacement of the principal point during a large focal length variation. Lavest [LAV 92] as well as Enciso [ENC 95] and [POL 95] used the zoom and also other parameters such as the focus change for reconstruction, or vergence, exactly like Bobet [BOB 94], within the framework of a stereoscopic head. Brooks and De Agapito [BRO 96] also studied the self-calibration of a stereo head as well as the associated degeneration, whereas Gaspard studied zoom in the case of monocular sequences [GAS 96]. Heyden [HEY 97] pursued his studies with the objective of a Euclidean reconstruction by variation of the focal length on a sequence of images.

3.2.3. *Taking into account specific displacements*

Taking specific displacements into account was studied first and is still widely used. This is due to several reasons. The first one is that many movements of objects or cameras are specific: a car on a highway, movements made by robotic systems whose degree of freedom is restricted or constrained, etc. The second reason is more mathematical: the fundamental matrix corresponding to the perspective projection vanishes for pure rotation, while the fact certain components of the movement are equal to zero can also simplify the expression of the fundamental matrix by reducing its number of parameters.

The study of pure rotation was conducted by a number of authors. Some of them assume that the intrinsic parameters are constant, while others consider that they can vary; some consider that the rotation is unknown, while others consider that its angle or its axis is known. Hartley [HAR 94b] considers a sequence of images taken by a camera in the axis of pure rotation passing through its optical center. Moreover, considering that the intrinsic parameters of the camera are constant, Hartley shows with algebraic reasoning that it is easy to find the intrinsic parameters of the camera. Later on, De Agapito and Hayman [AGA 98] studied the case of pure rotation with variable intrinsic parameters. Viéville [VIÉ 94a], in the case where the angle of rotation is known, calculated the intrinsic parameters, the optical center, the axis of rotation and the coordinates of 3D points to a close scale factor by carrying out at least four different rotations on the same axis, for two axes, and by following at least two points.

More recently, different authors have studied a certain number of specific cases. Sturm [STU 97A] studied the conditions of general self-calibration method degeneration and made explicit all critical movements for which self-calibration methods cannot be applied. Horaud and Christy [HOR 97] carried out Euclidean and affine reconstructions using different controlled movements by a robot (known movement with known parameters).

The case of pure translation has been less widely studied. However, Armstrong and Zisserman [ARM 94] as well as Pollefeys and Van Gool [POL 96] studied translation, and translation and zoom combinations.

Sturm [STU 97a], Wiles and Brady [WIL 96], and Armstrong and Zisserman [ARM 96] were interested in the reconstruction of scenes by planar movement. Sturm showed that we cannot remove the ambiguity on solutions without additional information while Wiles, Brady, Armstrong and Zisserman studied the different camera models for planar movement as well as the simplification of fundamental and homography matrices that follow them. Sturm also considered linear movements.

More generally, a certain number of authors studied specific cases and explained the related simplifications in the expressions of the fundamental matrix or homography in order to be able to avoid, on the one hand, critical equations and, on the other hand to benefit from a smaller number of parameters. Viéville and Lingrand [VIÉ 96b, VIÉ 99] examined these specific cases of translations and rotations and put in place a criterion for automatic selection of cases corresponding in the best way to the data. Torr and Zisserman [TOR 95] followed the same method in order to improve their estimation method of basic parameters of the robust algorithm of RANSAC: PLUNDER (*Pick Least UNDEgenerate Randomly*). Torr was closely interested in methods of model selection. Clarke [CLA 97] in his thesis studied another sample of specific cases of movement but with constant intrinsic parameters and without explaining the process of selecting the case.

From the algebraic developments of [VIÉ 99], it was shown that if we only consider algebraic constraints of a degree higher than 4, all specific models are models of:

– axis rotations or pure rotations with fixed intrinsic parameters;

– pure translations with or without variations of intrinsic parameters;

– frontal movement with fixed intrinsic parameters;

– movements leaving the retina invariant;

Displacement class	Parametrization, constraint	N
[p*t1r*]	$\mathbf{F} = 0$	0
[p1t5r0]	$\mathbf{F} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	0
[p1t[34]r0]	$\mathbf{F} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ -a & -b & 0 \end{pmatrix}, \|\mathbf{F}\| = 1$	1
[p1t2r0]	$\mathbf{F} = \begin{pmatrix} 0 & c & a \\ -c & 0 & b \\ -a & -b & 0 \end{pmatrix}, \|\mathbf{F}\| = 1$	2
[p2t[34]r1]	$\mathbf{F} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{pmatrix}, \|\mathbf{F}\| = 1$	4
[p2t[234]r0]	$\mathbf{F} = \begin{pmatrix} 0 & f & a \\ -f & 0 & b \\ c & d & e \end{pmatrix}, cb - ad = 0, \|\mathbf{F}\| = 1$	4

Table 3.1. *Models of movements leading to a specific form of the principal linear matrix in relation to parameters; N is the number of the degrees of freedom of the model*

which are the linear models given in Tables 3.1 and 3.2 from the following conventions:

[**p1**]	Constant intrinsic parameters
[**p2**]	Variable intrinsic parameters
[**t1**]	No translation
[**t2**]	Non-zero translation along the optical axis
[**t34**]	Zero translation along the optical axis
[**t5**]	Translation only along the optical axis
[**r0**]	No rotation
[**r1**]	Rotation along the optical axis
[**r2**]	General rotation

This gives, in a certain sense, an exhaustive view of all the specific cases of visual displacements, which can be perceived by non-calibrated monocular vision. Detailed development is given by Lingrand [LIN 99].

3.2.4. *Relation with specific properties in the scene*

The case of planar scenes has been widely studied as it corresponds in the case of perspective model of the camera, to the study of homography (i.e. collineation), just as in the case of pure rotation. Gaspard and Viéville [GAS 96] used homography to extract different planar structures in non-calibrated monocular scenes. Triggs [TRI 98b] studied self-calibration in the case of planar scenes but with constant intrinsic parameters.

Torr and Fitzgibbon [TOR 98b] studied different hypotheses concerning the relation between the correspondence of points for three consecutive images by considering them two at a time: homography then homography (H-H), homograph then, fundamental matrix (H-F), fundamental matrix then homography (F-H), or fundamental matrix, fundamental matrix (F-F). These authors proposed an automatic identification mechanism for one of the cases and then in the presence of homography, with the help of the GRIC criterion [TOR 98a], determined whether it refers to a plane structure or a pure rotation. This algorithm makes it possible to overcome the problem of point tracking on a sequence of images. It avoids the images posing problems and allows us to find the structure in a robust way.

Similarly, in Table 3.2, as analyzed in [VIÉ 96c], all specific cases where the projected movement is described by a specific linear homography as compared to the parameters as listed; this including "pure rotation" cases where the homography H_∞ is that of an infinite plane, i.e., cases without translation. The specific homography cases connected to a planar movement and dependent on the orientation of this plane as compared to the rotation, translation, or to the optical axis are explained in [LIN 99].

3.3. Self-calibration of a camera

In the field of self-calibration of a camera, a number of authors [MAY 92, VIÉ 96a, HAR 94a, SHA 94b] defined the symmetric matrix K:

$$\begin{aligned}\mathbf{K} &= \mathbf{A}\mathbf{A}^T \\ &\equiv \begin{pmatrix} \alpha_u^2 + u_0^2 + \alpha_v^2 \cot(\theta)^2 & u_0 v_0 - \alpha_v^2 \cot(\theta) & u_0 \\ u_0 v_0 - \alpha_v^2 \cot(\theta) & a_v^2 \dfrac{1}{\sin(\theta)^2} + v_0^2 & v_0 \\ u_0 & v_0 & 1 \end{pmatrix}\end{aligned} \tag{3.2}$$

Displacement classification	Parametrization	N
[p*t1r0]	$\mathbf{H} = \mathbf{I}$	0
[p1t[34]r0]	$\mathbf{H} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{pmatrix}$	2
[p2t1r0]	$\mathbf{H} = \begin{pmatrix} c & 0 & a \\ 0 & c & b \\ 0 & 0 & 1 \end{pmatrix}$	4
[p1t[34]r1]	$\mathbf{H} = \begin{pmatrix} c & d & a \\ -d & c & b \\ 0 & 0 & 1 \end{pmatrix}$	4
[p2t[34]r1]/	$\mathbf{H} = \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix}$	6
[p2t[234]r2]	$\mathbf{H} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{pmatrix}$	8

Table 3.2. *Specific homography models whose constraints are linear as compared to the components; N is the number of parameters of the model. We can always assume $H^{22} = 1$ to remove the non-determination of the scale factor in this case*

This is in correspondence with the matrix of intrinsic parameters A as shown in the following equations:

$$
\begin{aligned}
u_0 &= \frac{K^{02}}{K^{22}} \\
v_0 &= \frac{K^{12}}{K^{22}} \\
a &= \frac{K^{11}}{K^{22}} - v_0^2 \\
b &= \frac{K^{01}}{K^{22}} - u_0 v_0 \\
\alpha_v &= \sqrt{a - b} \\
\theta &= \arccos\left(\sqrt{b/a}\right) \\
c &= \frac{K^{00}}{K^{22}} - u_0^2 - b \\
\alpha_u &= \sqrt{c}
\end{aligned}
$$

which are defined as soon as $K^{22} \neq 0$, $a > b \geq 0$, and $c \geq 0$ in an unambiguous way so that by definition $\alpha_u \geq 0$ and $\alpha_v \geq 0$. This matrix is in fact the inverse of the matrix which parametrizes the "absolute conic", which geometrically characterizes the calibration of a camera (see [MAY 92]).

If we make explicit the fact that $\mathbf{R}$ is an orthogonal matrix in the definition $\mathbf{H}_\infty = \mathbf{A}'\mathbf{R}\mathbf{A}^{-1}$ we obtain:

$$\mathbf{K}' \equiv \mathbf{H}_\infty \, \mathbf{K} \, \mathbf{H}_\infty^T \tag{3.3}$$

which corresponds to five independent linear equations and we derive:

$$\tilde{\mathbf{s}} \, \mathbf{K}' \, \tilde{\mathbf{s}} \equiv \mathbf{F} \, \mathbf{K} \, \mathbf{F}^T \tag{3.4}$$

which corresponds to the two equations of Kruppa that define the self-calibration of a camera [MAY 92]. The present derivation method is algebraically the simplest one [ZEL 94].

Unfortunately, in general, we have for $N + 1$ views, $5(N + 1)$ parameters to calculate while $2N$ equations are available. Hence, it is necessary to introduce additional constraints, for example, that these parameters are constant [LUO 95] or that some of them are fixed [HEY 97].

In these two cases, Kruppa equations are numerically very unstable [LUO 95] except if the disparity between them is significant [ZEL 94].

In a simpler way, we can use a three parameter projection model to represent the projection of a point $\mathbf{M}$, whose coordinates are expressed in the frame of reference attached to the camera. On a point $\mathbf{m}$ of the retina:

$$Z\mathbf{m} = \mathbf{A}\mathbf{M} \quad \text{with } \mathbf{A} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.5}$$

Here, we suppose that the two axes of the retinal plane are orthogonal and that the ratio between the horizontal and the vertical focal lengths is known and is equal to 1.

Different authors have experimentally proved that this model is sufficient for regularly used cameras [VIÉ 95a] and also for visual sensors of high quality [WIL 94] or, more recently [BOU 98, HEY 97].

3.3.1. *Usage of pure rotations or points at the horizon*

The affine calibration is known, if we have an estimation of the homography transformation corresponding to the infinite plane (i.e., to the horizon) $\mathbf{H}_\infty$. This is the case for a pure rotation or when we can identify at least four remote points. In this case, from (3.3) we can directly calculate the calibration parameters of a view from those of the previous views. We can thus propagate the calibration information within the sequence of images, knowing this affine calibration in only one of these views [VIÉ 96a]. More generally, it is necessary to introduce 5 other constraints, for example, by assuming constant calibration parameters [HAR 94b, VIÉ 96].

In the last case, it was proved that at least two rotations are necessary, i.e., three views [VIÉ 96], which gives a relatively heavy experimental paradigm. More precisely, two rotations around different axes are required to select the intrinsic parameters [HAR 94b, VIÉ 96c]. Indeed, one rotation does not allow us to pick the horizontal and vertical scale factor but a linear combination of these two parameters [VIÉ 94a].

Identifying the $\mathbf{H}_\infty$ matrix, which corresponds to the "rotational" part of the displacement, can be done using different methods as discussed by [VIÉ 96c].

[VIÉ] proposes the simplest calibration method from pure rotations and especially the most numerically stable and most efficient to implement since it requires only one single rotation and provides an explicit solution in the least squared sense. Now, let us explain this method.

A plausible assumption practice [ENC 95, WIL 94] consists of considering only the following three parameters: the focal length f and the position of the principal point (u_0, v_0). This is sufficient to define calibration.

In this case we have:

$$u_0 = \frac{K^{02}}{K^{22}}$$

$$v_0 = \frac{K^{12}}{K^{22}}$$

$$f^2 = \frac{K^{00}K^{22} - \left(K^{02}\right)^2}{\left(K^{22}\right)^2}$$

$$= \frac{K^{11}K^{22} - \left(K^{12}\right)^2}{\left(K^{22}\right)^2} \tag{3.6}$$

and matrix K verifies two homogenous quadratic constraints:

$$\begin{aligned} \theta = \frac{\Pi}{2} &\Longrightarrow K^{01}K^{22} = K^{02}K^{12} \\ \alpha_u = \alpha_v &\Longrightarrow K^{00}K^{22} - \left(K^{02}\right)^2 = K^{11}K^{22} - \left(K^{12}\right)^2 \end{aligned} \tag{3.7}$$

The first is connected to the fact that we neglect the parameter of linear distortion and the second is connected to the fact that the ratio of the horizontal/vertical scale is known and leads to unity.

Hence, from (3.3), using (3.7), we obtain two quadratic equations on each set of parameters with respect to (u_0, v_0, f^2). These equations can be used as measurement equations in a recursive estimation process as tried by Heyden [HEY 97]; then, we are able to calibrate automatically after three rotations, even in the case of variable parameters.

However, note that if these equations are effective, they are not used in practice as they are numerically unstable.

Moreover, [VIÉ 99] show that the usage of the three parameter model does not give a particular form to Kruppa equations. Indeed, equation (3.4) is equivalent to:

$$\mathbf{K}' \equiv \left(\mathbf{F} + \mathbf{s}\mathbf{x}^T\right) \mathbf{K} \left(\mathbf{F} + \mathbf{s}\mathbf{x}^T\right)^T$$

for an undetermined vector $\mathbf{x}$ which can always be chosen to verify the constraints of (3.7), which certainly simplifies the equations but does not bring about any changes.

This explains why it is necessary to use specific displacements.

3.3.2. *Pure rotation and fixed parameters*

Let us assume that we perform a pure rotation without modifying the intrinsic parameters. In practice, if we perform a zoom, we not only modify the intrinsic parameters but generate a translation of the optical center; hence, we induce parallax (displacement of points on the retina depends on the depth of the point), which is very difficult to analyze.

Thus, a safe paradigm is to propose such an "aerobic" paradigm to self-calibrate a camera: rotation zoom.

In this case, we have $\mathbf{t} = 0$ and $\mathbf{A} = \mathbf{A}'$, and we can estimate $\mathbf{H}_\bullet \equiv \mathbf{H}_\infty$ since the retinal movement of the points corresponds to this homography. Moreover, this verifies the property $\det(\mathbf{H}_\infty) = 1$ making it possible to fix the scale ratio. We then obtain:

$$\begin{aligned} \mathbf{H}_\infty = \frac{\mathbf{H}_\bullet}{\det\left(\mathbf{H}_\bullet\right)^{\frac{1}{3}}} \quad \text{and} \quad \rho = \frac{\mathbf{A}\mathbf{u}}{\det(\mathbf{A})} \\ \Longrightarrow \mathbf{O} = \frac{\mathbf{H}_\infty - \mathbf{H}_\infty^{-1}}{2} = \sin(\theta)\,\mathbf{K}\,\tilde{\rho} \end{aligned} \tag{3.8}$$

This form is equivalent to the Luong decomposition of the matrix $\mathbf{H}_\infty$ [VIÉ 96] but the algebraic form proposed here is simpler and numerically more effective. Note that the matrix $\mathbf{O}$ verifies two algebraic constraints: $\det(\mathbf{O}) = \mathrm{trace}(\mathbf{O}) = 0$.

Although, in general, at least two rotations are necessary, in the case of a model with three parameters, by eliminating ρ, we obtain the following linear equations (in u_0, v_0, f^2):

$$\begin{cases} O^{00} - u_0 O^{20} = 0 \\ O^{11} - v_0 O^{21} = 0 \\ \left(O^{01} + O^{10}\right) - u_0 O^{21} - v_0 O^{20} = 0 \\ O^{20} f^2 + O^{00} u_0 + O^{01} v_0 + O^{02} = 0 \\ O^{21} f^2 + O^{10} u_0 + O^{11} v_0 + O^{12} = 0 \end{cases} \tag{3.9}$$

that are redundant and can be resolved to the least squares, which makes it possible to obtain an explicit solution given here:

$$\begin{cases} u0 = \dfrac{O^{20\,3} O^{00} + \left(O^{00} - O^{11}\right) O^{21\,2} O^{20} + \left(O^{01} + O^{10}\right) O^{21\,3}}{O^{20\,2} O^{21\,2} + O^{20\,4} + O^{21\,4}} \\ v0 = -\dfrac{\left(-O^{01} - O^{10}\right) O^{20\,3} + \left(O^{00} - O^{11}\right) O^{21} O^{20\,2} - O^{21\,3} O^{11}}{O^{20\,2} O^{21\,2} + O^{20\,4} + O^{21\,4}} \\ f^2 = -\dfrac{\left(O^{21} O^{11} + O^{20} O^{01}\right) v0}{O^{20\,2} + O^{21\,2}} - \dfrac{\left(O^{20} O^{00} + O^{21} O^{10}\right) u0}{O^{20\,2} + O^{21\,2}} \\ \qquad - \dfrac{O^{21} O^{12} + O^{20} O^{02}}{O^{20\,2} + O^{21\,2}} \end{cases}$$

and defined except if:

$$O^{20} = -\frac{\sin(\theta)}{f} u_y = 0 \quad \text{and} \quad O^{21} = \frac{\sin(\theta)}{f} u_x = 0 \tag{3.10}$$

which corresponds to a rotation around the optical axis. This is a configuration that is easy to avoid.

This then leads to a recommendation *to calibrate a camera by making a pure rotation or by making a rotational movement by seeing at least 4 points at the horizon.* It is perhaps the simplest and the most efficient strategy.

Moreover, we can also easily calculate, if necessary, the vector ρ since: $\mathbf{O}\rho = 0$ as well as the angle of rotation from equation (3.8).

3.3.3. *Rotation around a fixed axis*

Let us now assume that it is impossible to make a pure rotation, for example, when the camera makes a zoom: in this case the optical center may be far behind the mechanical center of the camera and always varies with the focal length [ENC 95, WIL 94]. Hence, it is not possible to assume pure rotation for a zoom except for some very specific hardware such as [PAH 92].

To easily generalize the previous approach to such cases, we assume that we carry out a rotation around a fixed axis in, let us say, direction $\mathbf{u}$, this vector being unary 1 (i.e. $\|\mathbf{u}\| = 1$) and passing through a point $\mathbf{C}$, which is the "center of rotation". In fact, we can choose any point on the axis of rotation Δ and to simplify the calculations, we take the point $\mathbf{C}$ such that $\vec{\mathbf{OC}} \perp \mathbf{u}$. We note by θ, the angle of rotation.

It is known [BEA 95, VIÉ 94a, VIÉ 95b] that such a rigid displacement is characterized by the fact that the resulting translation of off-centering the axis of rotation is orthogonal to this axis.

More precisely [VIÉ 94a, VIÉ 99], we can write the precise form taken by the fundamental matrix

$$\mathbf{F} \equiv \|\mathbf{C}\| \begin{bmatrix} \sin(\theta)\, \tilde{\mathbf{f}}_0 \\ +\big(1 - \cos(\theta)\big) \left[\mathbf{f}_1\, \mathbf{f}_2^T + \mathbf{f}_2\, \mathbf{f}_1^T\right] \end{bmatrix} \tag{3.11}$$

with:

$$\mathbf{f}_0 = \lambda^2 \frac{2}{\det(A)} \mathbf{A} \frac{\mathbf{C} \wedge \mathbf{u}}{\|\mathbf{C}\|}$$

$$\mathbf{f}_1 = \lambda \mathbf{A}^{-1T} \mathbf{u}$$

$$\mathbf{f}_2 = \lambda \mathbf{A}^{-1T} \frac{\mathbf{C} \wedge \mathbf{u}}{\|\mathbf{C}\|}$$

while:

$$\mathbf{f}_0^T \mathbf{f}_1 = 0 \tag{3.12}$$

which is the single constraint between these vectors [VIÉ 99].

If we assume that we know the fundamental matrix (up to a close scale factor) and the angle of rotation θ, then, we can explicitly calculate the intrinsic parameters:

$$\begin{cases} u_0 = f_0^0 - \gamma\, f_2^0 / f_0^2 \\ v_0 = f_0^1 - \gamma\, f_2^1 / f_0^2 \\ f = \dfrac{\sqrt{\gamma\left(f_0^0 f_2^0 + f_0^1 f_2^1 + f_0^2 f_2^2\right) - \gamma^2\left(\left(f_2^0\right)^2 + \left(f_2^1\right)^2\right)}}{f_0^2} \end{cases} \tag{3.13}$$

up to a variable indetermination γ.

Similarly, we can make explicit two linear equations:

$$\begin{cases} f_2^1\left(f_0^2 u_0 - f_0^0\right) - f_2^0\left(f_0^2 v_0 - f_0^1\right) = 0 \\ f^2 + u_0^2 + v_0^2 - \left(f_0^0 u_0 + f_0^1 v_0 + \mu f_2^2\right) / f_0^2 = 0 \end{cases} \tag{3.14}$$

with:

$$\mu = \frac{f_2^0\left(f_0^2 u_0 - f_0^0\right) + f_2^1\left(f_0^2 u_0 - f_0^1\right)}{\left(f_2^0\right)^2 + \left(f_2^1\right)^2}$$

The first equation gives us a linear equation with respect to (u_0, v_0) to be solved with at least two displacements; the second directly gives the focal length from the estimation of (u_0, v_0).

Hence, *two rotations around fixed axes* are necessary to self-calibrate a visual sensor. Numerical considerations suggest using a horizontal and a vertical rotation [VIÉ 94a].

We can also use these two equations as measurement equations on a Kalman filter.

Furthermore, we also obtain the extrinsic parameters:

$$\mathbf{u} \equiv \mathbf{A}^T \mathbf{f}_1 \quad \text{and} \quad \mathbf{C} \equiv \mathbf{u} \wedge \mathbf{A}^{-1} \mathbf{f}_0 \tag{3.15}$$

Here, as soon as the intrinsic parameters are estimated, we recover the direction of the axis of rotation and its position up to a scale factor, for this monocular system.

It is necessary to note that these equations are degenerated only if:

$$f_0^2 = \frac{2}{f^2}\left[C^0 u^1 - C^1 u^0\right] = 0 \tag{3.16}$$

i.e. if, in the image plane, the principal point and the projections of the axis of rotation and of point $\mathbf{C}$ are colinear. In practice, this situation arises only in the case where the rotation is around the optical axis, as in the preceding case.

From a theoretical point of view, we can say that this case is a simple extension of a case of pure rotation where the component related to the translation can be eliminated. The result is, by itself, less interesting since we do not have an explicit solution for the three parameters, but only two equations. However, they are very simple, linear, and can be easily used in the estimation process.

Hence, we are now faced with a second process of active self-calibration, whose performances have been studied by [VIÉ 94a] and which obtain a precision of 3–4 pixels for the principal point and 1 to 2% for the focal length.

3.4. Perception of depth

3.4.1. *Usage of pure translations*

To calibrate a camera, the best movement is a rotation, since movements in the image do not depend on the depth. On the contrary, to perceive the distance of an object from the camera, pure translation is the most appropriate

movement, since movements in the image are inversely proportional to the depth:

$$\mathbf{m}' = \frac{Z}{Z'}\mathbf{m} + \frac{1}{Z'}\mathbf{s} \Longrightarrow \pi' = \frac{1}{Z'} = \frac{\|\mathbf{m}' \wedge \mathbf{m}\|}{\|\mathbf{m}' \wedge \mathbf{s}\|} \tag{3.17}$$

where $\mathbf{m}$ and $\mathbf{m}'$ are two corresponding points related to the perspective projection of a 3D point $\mathbf{M}$ for two different image acquisitions. The depth of this point with respect to each image Z and Z', respectively. s corresponds to the expansion focus due to the pure translation.

In this case, the vector $\mathbf{s}$ is easily estimated up to a scale factor from at least two points, since

$$\left[\underbrace{\mathbf{m} \wedge \mathbf{m}'}_{\mathbf{d}}\right]^T \mathbf{s} = 0. \tag{3.18}$$

For two couples of points and by considering the vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ defined in (3.18) we have $\mathbf{s} \equiv \mathbf{d}_1 \wedge \mathbf{d}_2$.

Here we consider a translation *without* modification of intrinsic parameters.

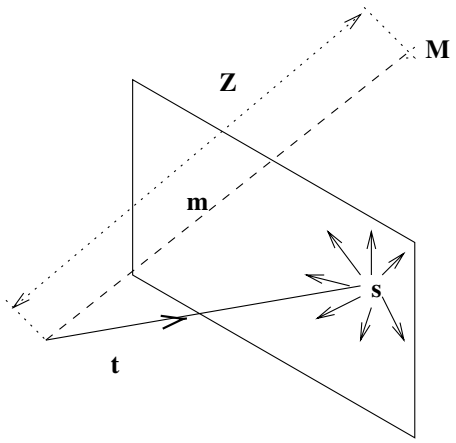


Figure 3.2. *Geometric elements for a pure translation*

The depth Z is the Euclidean distance along the optical axis from the optical center point. We also consider the *proximity* $\pi = \frac{1}{Z}$, this parameter being numerically better conditioned in the equations.

To reconstruct the point, it is necessary to know the intrinsic parameters since $\mathbf{M}' = Z'\mathbf{A}^{-1}\mathbf{m}'$. However, without this, we can still estimate a quantity, which makes it possible to decide the *relative distance of the two points as compared to the camera* or if a point is *at the horizon* (i.e. $\pi \simeq 0$). In fact, we are in a case of an "affine calibration" since the collineation of the plane at infinity is simply defined by $\mathbf{H}_\infty = \mathbf{I}$.

The calculation of proximity is better conditioned for points $\mathbf{m}'$ non-colinear with $\mathbf{s}$, i.e., far away from this point in the image. The point $\mathbf{s} = \mathbf{At}$ is the projection of the translation on the retina, called the expansion focus or the epipole according to the context, since these definitions are equivalent. This means that it will be better to use lateral translations (horizontal or vertical) rather than longitudinal ones (aligned with the optical axis).

Regarding parametrization of the movement in this very common case (for example, see [SHA 94a, VAN 94]), we have:

$$\mathbf{F} = \tilde{\mathbf{s}} \quad \text{and} \quad \mathbf{H}_\infty \equiv \mathbf{I} \tag{3.19}$$

which are thus very easy to use.

In particular, the anti-symmetry of the fundamental matrix is characteristic due to a pure translation, which makes it possible to detect such a condition and also to control external parameters in order to obtain such a type of movement.

In the case of a planar object in pure translation, [VIÉ 99] show that the corresponding homography is of the form:

$$\mathbf{H} \equiv \mathbf{I} + \frac{\mathbf{s}}{\|\mathbf{s}\|}\nu^T \tag{3.20}$$

This homography is *characterized by the fact that two of its eigenvalues are equal* and that we can easily calculate the underlying parameters (of course, to a close scale factor):

$$\mathbf{s} \equiv \mathbf{u}_{1a} \wedge \mathbf{u}_{1b}; \quad \nu = \frac{\lambda_0/\lambda_1 - 1}{\frac{\mathbf{s}}{\|\mathbf{s}\|}^T \mathbf{u}_0}\mathbf{u}_0 \tag{3.21}$$

where $\mathbf{u}_{1a}$ and $\mathbf{u}_{1b}$ are the two eigenvectors corresponding to the same eigenvalue λ_1, while $\mathbf{u}_0$ is the eigenvector corresponding to the other eigenvalue λ_0.

Thus, it is easy to detect if a homography corresponds to a planar movement of pure translation and to calculate its characteristics, as in the calibrated case (see [FAU 93] for a review).

3.4.2. *Retinal movements*

Pure translation is not the only type of specific movement that facilitates the visual perception of the structure of the observed scene. Another class of displacements, which is also useful are movements leaving the image plane (the retina) invariant.

It is shown [LIN 96] that it corresponds to a rigid movement such that:

$$t^2 = u^0 = u^1 = 0 \tag{3.22}$$

i.e., it is composed of a rotation around the optical axis and horizontal and vertical translations.

In this case, the calibration parameters can be fixed or variable.

Hence, we have a fundamental matrix of the form:

$$\mathbf{F} = \begin{pmatrix} 0 & 0 & \lambda t^1 \\ 0 & 0 & -\lambda t^0 \\ F^{20} & F^{21} & F^{22} \end{pmatrix} \tag{3.23}$$

with:

$$F^{20} = \frac{f'}{f}\big(- F^{02}\cos(\theta) - F^{12}\sin(\theta)\big)$$

$$F^{21} = \frac{f'}{f}\big(F^{02}\sin(\theta) - F^{12}\cos(\theta)\big)$$

$$F^{22} = -\big(F^{02}u'_0 + F^{20}u_0 + F^{12}v'_0 + F^{21}v_0\big)$$

This matrix is subject to four constraints:

$$F^{00} = F^{01} = F^{10} = F^{11} = 0 \tag{3.24}$$

and [VIÉ 99] proves that reciprocally these constraints involve either $t^0 = t^1 = t^2 = 0$, while the matrix is either undefined or $u^0 = u^1 = t^2 = 0$, which shows that these conditions are characteristic of a retinal displacement.

From this matrix, we can easily obtain the optimal:

– angle of rotation, in the least square sense:

$$\theta = \arctan\left(\frac{F^{21}}{F^{20}}\right) - \arctan\left(\frac{F^{12}}{F^{02}}\right) \tag{3.25}$$

– the direction of translation: $\mathbf{t} \equiv \mathbf{s} \equiv (F^{02}, F^{12}, 0)^T$;

– the variation in the focal length:

$$\frac{f'}{f} = \sqrt{\frac{\left(F^{20}\right)^2 + \left(F^{21}\right)^2}{\left(F^{02}\right)^2 + \left(F^{12}\right)^2}} \tag{3.26}$$

– and a linear equation on the positions of the principal point across the expression F^{22}, while these four conditions completely define the matrix $\mathbf{F}$, which by itself has 4 degrees of freedom.

The last two equations correspond to the Kruppa equations for self-calibration, whereas the first two give us related the extrinsic parameters.

We can also characterize the matrix $H_\infty = \begin{pmatrix} a & -b & c \\ b & a & d \\ 0 & 0 & 1 \end{pmatrix}$, allowing us to obtain:

– the angle of rotation: $\theta = \arctan(\frac{b}{a})$,

– the new calibration parameters from the preceding ones:

$$\begin{cases} u'_0 = c + \dfrac{f'}{f}\big(\cos(\theta)u_0 - \sin(\theta)v_0\big) \\ v'_0 = d + \dfrac{f'}{f}\big(\cos(\theta)v_0 + \sin(\theta)u_0\big) \\ f' = f\sqrt{a^2 + b^2} \end{cases} \tag{3.27}$$

whereas it is not possible to self-calibrate in this case. If the parameters are constant, then we calculate the coordinates of the principal point but not the focal length. This situation corresponds to a pure rotation around the optical axis and we find the same singularities already pointed out in this case.

If we consider the observation of a planar structure during a retinal movement, we obtain a homography matrix of the form:

$$\mathbf{H} = \begin{pmatrix} a & b & d_u \\ c & d & d_v \\ 0 & 0 & 1 \end{pmatrix} \tag{3.28}$$

which is thus an affine transformation of the image. Reciprocally, such a transformation corresponds either to (i) a retinal movement $t^2 = s^2 = u^0 = u^1 = 0$ or (ii) to the fronto-parallel movement of a plane [VIÉ 99, VIÉ 96b].

In the general cases, where parameters vary, it is not possible to infer these six coefficients since there are 11 unknown quantities and generally, there are no usable constraints at this stage. On the other hand, in the case where calibration parameters are constant [VIÉ 99, VIÉ 96b], we obtain:

$$\begin{cases} \theta = 2\arctan(Z) \\ t^0 = \lambda\,\cos(\alpha) \\ t^1 = \lambda\,\sin(\alpha) \\ n^0 = \dfrac{t^0\big(H^{00} - H^{11}\big) + t^1\big(H^{01} - H^{10}\big)}{\big(t^0\big)^2 + \big(t^1\big)^2} \\ n^1 = \dfrac{t^1\big(H^{11} - H^{00}\big) + t^0\big(H^{01} - H^{10}\big)}{\big(t^0\big)^2 + \big(t^1\big)^2} \end{cases} \tag{3.29}$$

with:

$$\begin{aligned} &\big(1 + H^{00} + H^{11} + H^{11}H^{00} - H^{01}H^{10}\big)Z^2 + 2\big(H^{01} - H^{10}\big)Z \\ &\quad + \big(1 - H^{00} - H^{11} + H^{11}H^{00} - H^{01}H^{10}\big) = 0 \end{aligned}$$

and: $\alpha = \frac{1}{2}\arctan(2\,\frac{\alpha_1}{\alpha_2})$ defined by:

$$\begin{aligned} \alpha_1 = &\big(H^{11}H^{01} + H^{00}H^{10}\big) - \cos(\theta)\big(H^{01} + H^{10}\big) \\ &+ \sin(\theta)\big(H^{11} - H^{00}\big) \end{aligned}$$

$$\alpha_2 = \Big(\big(H^{00}\big)^2 + \big(H^{01}\big)^2 - \big(H^{10}\big)^2 - \big(H^{11}\big)^2 \Big) - 2\Big(\cos(\theta)\big(H^{11} - H^{00}\big) + \sin(\theta)\big(H^{10} + H^{01}\big) \Big)$$

Hence, we have two solutions for the displacement as in the calibrated case [VIÉ 96b].

Moreover, we are left with a linear equation for the intrinsic parameters (u_0, v_0):

$$\begin{aligned} &\sin(\alpha)\big(H^{02} - u_0\big(1 - H^{00}\big) + v_0 H^{01}\big) \\ &\quad = \cos(\alpha)\big(H^{12} + u_0 H^{10} - v_0\big(1 - H^{11}\big)\big) \end{aligned} \tag{3.30}$$

while there is no constraint on the focal length.

Note that the specific case of affine movement, i.e., a simple translation of the image, corresponds to a pure translation of a fronto-parallel plane without variation in intrinsic parameters. In this case $\mathbf{s} \equiv (d_u, d_v, 0)$. It is interesting to have such a geometric interpretation for a given explicit model.

3.4.3. *Variation of the focal length*

Another specific movement that helps us to give an indication on the depth is the "zoom" corresponding to the change in the focal length, as studied by several authors [POL 95, ENC 96, LAV 93, GAS 97].

The key point is that the zoom does not correspond only to a simple variation of intrinsic parameters but also to a translation of the optical center during the increase or decrease of the focal length. More precisely, such a translation cannot be put into direct relation with the variation of the principal point as the mechanical process (screw, slide, etc.) can be rather complex and is not always deterministic [BOB 94, ENC 94]. Similarly, we cannot truly fix the variation of the principal point [WIL 93, VIÉ 95a] as compared to other parameters and it is necessary to assume a non-deterministic variation of these parameters.

Hence, we suppose that there is a translation primarily along the optical axis, but not necessarily.

On the other hand, a zoom is carried out without any rotation of the image plane [WIL 94].

Under these assumptions, the fundamental matrix has the following form:

$$\mathbf{F} = \begin{pmatrix} 0 & -\lambda t^2 & \lambda f t^1 - F^{01} v_0 \\ \lambda t^2 & 0 & -\lambda f t^0 + F^{01} u_0 \\ -\lambda f' t^1 + F^{01} v'_0 & \lambda f' t^0 - F^{01} u'_0 & F^{22} \end{pmatrix} \tag{3.31}$$

with:

$$F^{22} = \left(u_0 v'_0 - v_0 u'_0\right) F^{01} - u_0 F^{20} - u'_0 F^{02} - v_0 F^{21} - v'_0 F^{12}$$

which is constrained by:

$$F^{00} = 0, \quad F^{11} = 0, \quad F^{01} = -F^{10} \tag{3.32}$$

this condition being necessary and sufficient.

In this case, we have 4 independent parameters since the 7 degrees of freedom of a fundamental matrix obey 3 constraints. More precisely, by knowing (u_0, v_0, f), we can recover, if $F^{01} = -\lambda t^2 \neq 0$ (which is expected for a zoom, while if $t^2 = 0$, it will be the previous case):

$$\mathbf{t} \equiv \begin{pmatrix} F^{12} - F^{01} u_0 \\ -F^{02} - F^{01} v_0 \\ f F^{01} \end{pmatrix} \text{ and } \begin{cases} u'_0 = \dfrac{f'}{f} u_0 - \dfrac{\frac{f'}{f} F^{12} + F^{21}}{F^{01}} \\ \\ v'_0 = \dfrac{f'}{f} v_0 + \dfrac{\frac{f'}{f} F^{02} + F^{20}}{F^{01}} \end{cases} \tag{3.33}$$

Hence, we recover $\mathbf{t}$ up to a scale factor and also recover the variation of the position of the principal point.

On the other hand, we cannot recover f'/f as this expansion factor is an unknown factor of the parametrization of the movement [VIÉ 96c]. Equations (3.31) with $\det(\mathbf{F}) = 0$ and equations (3.33) being equivalent, we cannot recover another parameter. Thus, the "zoom", which is precisely a variation of the focal distance, does not enable the estimation of this variation at this stage. Even if we were assuming a "perfect" zoom with $t^0 = t^1 = 0$, we obtain an estimation of (u_0, v_0) and (u'_0, v'_0) but not an estimation relative to the focal length.

If we are able to estimate the affine calibration, i.e., the homography of the points at the horizon, we then obtain a matrix of the form:

$$\mathbf{H}_\infty \equiv \begin{pmatrix} \frac{f'}{f} & 0 & u'_0 - u_0\frac{f'}{f} \\ 0 & \frac{f'}{f} & v'_0 - v_0\frac{f'}{f} \\ 0 & 0 & 1 \end{pmatrix} \tag{3.34}$$

which makes it possible to directly follow the evolution of the calibration parameters in coherence with equation (3.3).

Similarly, if we consider the observation of a planar structure [VIÉ 99, GAS 97], the matrix of the associated homography is then an algebraic object with 7 degrees of freedom:

$$\mathbf{H} \equiv \begin{pmatrix} \frac{f'}{f} + s^0\nu^0 & s^0\nu^1 & u'_0 - u_0\frac{f'}{f} + s^0\nu^2 \\ s^1\nu^0 & \frac{f'}{f} + s^1\nu^1 & v'_0 - v_0\frac{f'}{f} + s^1\nu^2 \\ s^2\nu^0 & s^2\nu^1 & 1 + s^2\nu^2 \end{pmatrix} \tag{3.35}$$

and verifies only one constraint:

$$H^{01}H^{20}H^{20} - H^{00}H^{21}H^{20} - H^{10}H^{21}H^{21} + H^{11}H^{21}H^{20} = 0$$

If the plane is not fronto-parallel, we obtain:

– 2 degrees of freedom for the projection of the translation up to a scale factor:

$$\mathbf{s} \equiv \left(H^{01}H^{20}, H^{10}H^{21}, H^{20}H^{21}\right)$$

– 2 evolution equations of the principal point:

$$\begin{cases} u'_0 = \frac{f'}{f}u_0 + \frac{H^{01}}{H^{21}} + \frac{f'}{f}\frac{H^{01}H^{22} - H^{02}H^{21}}{H^{01}H^{20} - H^{00}H^{21}} \\ v'_0 = \frac{f'}{f}v_0 + \frac{H^{10}}{H^{20}} + \frac{f'}{f}\frac{H^{10}H^{22} - H^{12}H^{20}}{H^{10}H^{21} - H^{11}H^{20}} \end{cases}$$

However, again, we do not get any equation for $\frac{f'}{f}$, since the 3 remaining degrees of freedom are related to the normal $\mathbf{n}$ of the plane, parametrized by $\nu = \mathbf{A}^{-T} A^{-T} \frac{\mathbf{n}}{d}$,

$$\nu = \begin{pmatrix} \mu H^{20} \\ \mu H^{21} \\ \mu H^{22} - \dfrac{1}{s^2} \end{pmatrix} \text{ with } \mu = \frac{f'}{f} \frac{2}{\left(H^{00} + H^{11}\right) s^2 - H^{20} s^0 - H^{21} s^1}$$

are also functions of $\frac{f'}{f}$, while s^2 here is supposed to be different from 0.

This indetermination phenomenon is explained by the fact that a zoom corresponds to an expansion of space, which is not measured in the monocular case (see [VIÉ 96a] for an algebraic development). We can also interpret this mechanism by noticing that in the first order [VIÉ 00], we only estimate the "collision time", which combines the relative variation of the focal length with the variation of the distance of the object along the optical axis.

This calibration problem, unless a pre-calibration on a known object [LI 96] is used or unless a more specific model [POL 95] is used, requires, *a priori*, to work with a projective or affine model as, for example, in [HAY 96].

However, the caveat was overcome by Gaspard [GAS 97] who noticed that the form of the given matrix in (3.34) is characteristic of a fronto-parallel plane as seen in (3.15) with:

$$\mathbf{H} \equiv \begin{pmatrix} \dfrac{f'}{f} & 0 & u'_0 - u_0 \dfrac{f'}{f} + s^0 n^2 \\ 0 & \dfrac{f'}{f} & v'_0 - v_0 \dfrac{f'}{f} + s^1 n^2 \\ 0 & 0 & 1 + s^2 n^2 \end{pmatrix} \tag{3.36}$$

and then developed a paradigm that makes it possible to distinguish it from other fronto-parallel planes, if the plane at the horizon is present in the scene, thus resolving the problem of estimating $\frac{f'}{f}$.

The knowledge of the value of $\mathbf{H}_\infty$, corresponding to an affine calibration of the camera makes it possible to construct the scene observed. Such reconstructions are of interest since affine transformations preserve the medium, parallelism, order of depth, etc. A number of perceptual tasks such as the positioning of a point as compared to a plane [ROB 93], its application

to avoid obstacles [ZEL 94] or the calculation of the convex casing of an object of the scene [ROB 93], without explicit knowledge of the intrinsic parameters of the camera can be realized in this way.

Reconstruction of the depth from a fronto-parallel plane

The similarity of the homography formed between a fronto-parallel plane and the collineation of the plane at infinity leads us to consider a reconstruction, not from $\mathbf{H}_\infty$ but from a homography $\mathbf{H}$ corresponding to a fronto-parallel plane.

By taking the vectorial product with $\mathbf{m}'$, in the equation $\mathbf{m}' = \mathbf{H}_\infty \mathbf{m} + \frac{1}{z}\mathbf{s}$ [GAS 96] which parametrizes the retinal movement, we obtain:

$$\mathbf{m}' \wedge \mathbf{H}_\infty \, \mathbf{m} + \frac{\mathbf{m}' \wedge \mathbf{s}}{Z} \equiv 0$$

where $\mathbf{H}_\infty = \mathbf{H} - \mathbf{s}\,\nu^T$, with $\nu = (0, 0, \frac{1}{d})^T$.

A step further:

$$\mathbf{m}' \wedge \mathbf{H}\mathbf{m} + \left(\frac{1}{Z} - \frac{1}{d}\right)\left(\mathbf{m}' \wedge \mathbf{s}\right) \equiv 0$$

Then, we reconstruct $\frac{1}{\tilde{Z}}$ with:

$$\frac{1}{\tilde{Z}} = \frac{1}{Z} - \frac{1}{d} \tag{3.37}$$

The reconstructed depth is then $\tilde{Z}$, related to the depth, as represented in Figure 3.3, by:

$$\tilde{Z} = \frac{Z}{1 - \frac{Z}{d}}$$

By following the direction of the translation, which is *a priori* unknown, we thus reconstruct either the depth z or its inverse.

A similar case occurs when retrieving $\mathbf{H}_\infty$ among a set of homographs. As explained [ROB 95b], we can estimate the relative position of points as compared to a plane. This makes it possible to find the triangles related to the extremes of the scene (in depth). If we now assume that points are at the horizon, we can sort the scene depths. However, we could also mismatch the horizon plane with the closest plane in this case.

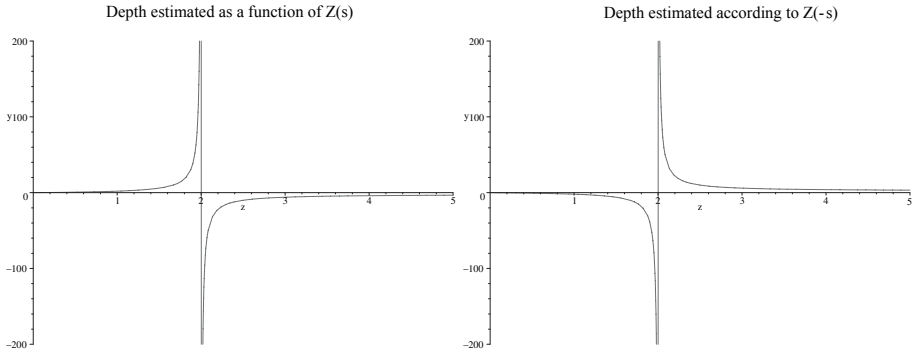


Figure 3.3. *Reconstruction of the depth from a fronto-parallel plane following the direction of translation, a priori unknown, we reconstruct either the depth Z or its inverse*

If we assume $\mathbf{s}$ is known, i.e., the direction of the translation, we can then differentiate the closest from the farthest from the enclosed equation. In the case of a zoom, we can assume this known sign since the fact of carrying out a zoom (or a back zoom) will control the direction of the translation in Z. Specific cases relating not only to the nature of movement but also to the geometry of the observed scene make it possible for us to obtain here a reconstruction of the scene, with depth ordered since the function represented in Figure 3.3 is strictly increasing.

3.5. Estimating a specific model on real data

Constitution of a set of measurements

In practice, an early vision module detects points of interest in an image, corresponding to the points of the strongest curvature of the image intensity signal (corners, junctions between two lines, spots, etc.) in such a way that they are correctly localized. Very often the Harris corner detector [HAR 88] is used as described in detail by [ZHA 95] in this context, but we can also suggest interactive methods [BOU 97], which make it possible to have a better control over the choice of the data and allow us to refine the localization using correlation methods.

In any case, it is necessary to accept the idea that there are erroneous data, for example, because of ambiguity in the image, erroneous matching, or because the point belongs to a non-rigid object or because its movement does not correspond to the movement to be evaluated.

If we consider a sequence of images, we have to face the following problem: (a) on the one hand it is necessary to have *maximum* disparity between the images, which will help to estimate the movement so that the numerical conditioning of these equations is correct, while (b) we will also need a *minimum* disparity to easily put into correspondence the visual landmarks from one image to the other.

We easily overcome this problem by following each point in the sequence of images using consecutive images and by applying the equations to the two extreme images of the sequence. By doing so, we also remove unstable measurement points, which will not be followed along the entire sequence; we avoid ambiguities relating to one aspect of the viewed scene in a particular angle, etc. To obtain correct accuracy for the measurement, we can (for example, see [VIÉ 94a]) interpolate the path of each point using a polynomial model or using splines to obtain an under-pixel precision.

Definition of an estimation criterion

In practice, it is interesting to estimate not only the parameters of the different models analyzed here, but also to be able to determine the best model given a dataset. Such procedures have been analyzed by [LIN 99] and then tackled again at the level of software integration [VIE 01].

In order to eliminate a fundamental matrix [TOR 95, ZHA 95], a useful criterion is based on the distance of the point to the epipolar line defined by matrix $\mathbf{F}$:

$$\mathbf{F}_{\bullet} = \arg\min_{\mathbf{F}} \left[\biguplus_{\{\mathbf{m}\}} \frac{\left(\mathbf{m}'^{T}\mathbf{F}\mathbf{m}\right)^{2}}{\sqrt{\left[\left(\bar{\mathbf{F}}^{T}\mathbf{m}'\right)^{0}\right]^{2} + \left[\left(\bar{\mathbf{F}}^{T}\mathbf{m}'\right)^{1}\right]^{2}}} \right]. \tag{3.38}$$

Assuming that each data point has an identical precision, this quantity is expressed in pixel^2 and corresponds to the mean quadratic distance of the point to its epipolar in each image.

If we freeze the denominator, we obtain a simple quadratic criterion with respect to the components of $\mathbf{F}$, hence leading to linear equations, which suggests minimizing this criterion in an iterative manner [HAR 95]. There are also variants of this criterion which are adapted to the calculation of the structure of the scene and not only to its movement [STU 96], whereas if we

consider that each measurement point is defined with a given precision, the criterion can be weighted accordingly.

More generally, it is a non-linear quadratic criterion and the notation $\biguplus$ depends on the selected statistical criterion: for the least squares method, it is a simple sum, whereas in the case of robust methods such as least squares medians, it corresponds to a median [HUB 81]. Other methods such as RANSAC [BOL 81] are based on the separation of errors in increasing order to remove the highest ones, i.e. those which run a higher risk of mismatching (for example, see [SHA 93, ZHA 95]).

Perhaps we can recommend, among all these methods, a very simple mechanism of random picking, which is defined as: randomly pick a minimum number of measurements and estimate the corresponding parameter. Then, calculate the error for all the measurements. Select a maximum error threshold and count the number of measurements that are below this threshold. This method is statistically justified and is very easy to implement.

In any case, once a parameter is estimated, it is necessary to refine using some least squares criteria coherent data to eliminate the influence of additive noise [ROU 87]. At this stage, such a criterion can be statistically biased and the method in [KAN 92], implemented in [VIE 01] avoids this problem.

A step further, new methods that take into account the physical nature of the parameters have been introduced and are, in the case of local minimization, faster and more effective [GAS 00].

In the case of the estimation of a homography matrix, we can write[1] a similar criterion [VIÉ 96c]:

$$\mathbf{H}_{\bullet} = \arg\min_{\mathbf{H}} \left[\biguplus_{\{\mathbf{m}\}} \frac{\left\| \left(\left(\mathbf{h}^2\right)^T \mathbf{m} \right) \mathbf{m}' - \mathbf{H}\,\mathbf{m} \right\|^2}{\left(\left(\mathbf{h}^2\right)^T \mathbf{m} \right)^2} \right] \tag{3.39}$$

1. The relation $\mathbf{m}' = \frac{\mathbf{H}\,\mathbf{m}}{((\mathbf{h}^2)^T\,\mathbf{m})}$ is a vectorial form of: $\begin{cases} u' = \dfrac{H^{00}u + H^{01}v + H^{02}}{H^{20}u + H^{21}v + H^{22}} \\ v' = \dfrac{H^{10}u + H^{11}v + H^{12}}{H^{20}u + H^{21}v + H^{22}} \\ 1 = 1 \end{cases}$

by noting $\mathbf{H} = \begin{bmatrix} [\mathbf{h}^0]^T \\ [\mathbf{h}^1]^T \\ [\mathbf{h}^2]^T \end{bmatrix}$.

which corresponds to the quadratic distance in pixel^2 of the prediction error of the displacement model and has the same properties as the preceding criterion.

Beyond this formulation, other methods propose to estimate linear criteria by introducing new intermediate parameters [MEE 91, TRI 98a, TAU 91, LEE 99] whose redundancy has to be tackled later. For example, [LEE 99] proposed a reprojection method on a linearization of the initial problem, which is biased [KAN 96], even if this bias is negligible in practice [LEE 99]. Other mechanisms are based on the "renormalization" of the criterion, as developed by [CHO 99].

3.5.1. *Application of the estimation mechanism to model inference*

It appears that the methods discussed previously are very robust with respect to erroneous data and make it possible to distinguish different movements in the scene. Thus, we can collect the displacement measurement of the main part of the scene (for example, the movement of the background compared to the camera or that of an object largely covering the view field). Then, by reapplying the methods on the measurements, which have been rejected during this first estimation, we can expect to calculate the movement of another object in movement and iteratively, to segment the objects on the basis of their movement.

In our case, it is also necessary to compare different models applied to the same dataset. To this purpose, we consider that *the minimized quadratic criterion corresponds, under Gaussian additive noise hypothesis, to a distribution* χ^2, i.e., if we note by $\epsilon^2_{\mathbf{m}}$, the quadratic error of each measurement given in criteria (3.38) or (3.39):

$$\chi^2_d = \frac{\sum_{\{\mathbf{m}\}} \epsilon^2_{\mathbf{m}}/\sigma^2}{\mathrm{card}(\{\mathbf{m}\}) - N} \tag{3.40}$$

where σ^2 is the variance in pixel^2 of the measurement, about one pixel. Here N is the number of parameters of the model, and $\mathrm{card}(\{\mathbf{m}\})$ is the number of measurements.

Thus, we can compare two distributions using the Fisher test [VIE 01] or from the family of tests of Akaike [LIN 99]. [VIE 01] shows that we compare these values of χ^2 under assumptions that are never verified (linearity, additive Gaussian noise, etc.) and so propose, for example, a comparison of the type:

$$\chi^2_{d'} \leqslant \chi^2_d e^{-\kappa \frac{d'-d}{d'}}$$

for $d' > d$ with $\kappa \in [1 \ldots 10]$ to be adjusted according to the application. It has also been experimentally noted [LIN 99] that for good data, these different tests give equivalent results, whereas for data which are more disturbed, no test actually makes a distinction possible. The choice of this comparison criterion is thus not critical.

Hence, with such a paradigm we can select a model with a minimum number of parameters (even if, in the absolute, its residual error is higher), i.e. best adapted to the data.

3.5.2. *Some experimental results*

Let us finally propose some experimental results, which illustrate these developments.

These experiments have been conducted in [VIÉ 99] and then completed by [LIN 99] as well as [GAS 96] and reported in [VIÉ 01].

Figures 3.4 and 3.5 show us two experimental results, which make it possible to test the inference mechanism proposed here, for both interior or exterior scenes.

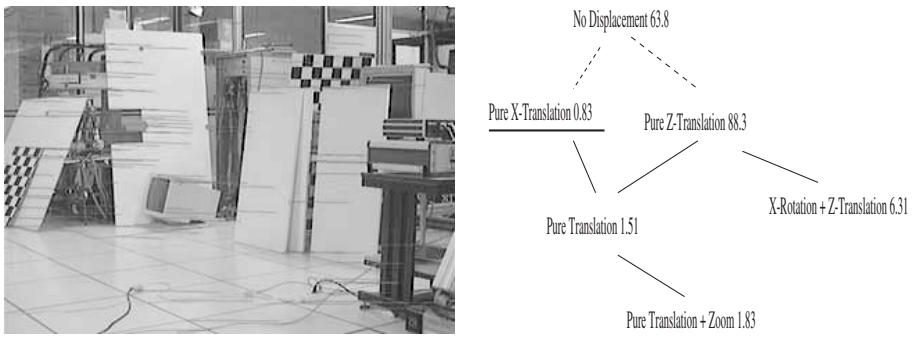


Figure 3.4. *Partial view of the hierarchy of models for a specific displacement, i.e. a translation along the X axis. For details, see the text*

The "cost" of the model is given in pixels; it is the square root of the mean quadratic error between the position of the measurement point in the image and the position predicted by the model, which is a standard in the field (for example, see [LUO 93, TOR 97, ZHA 94]).

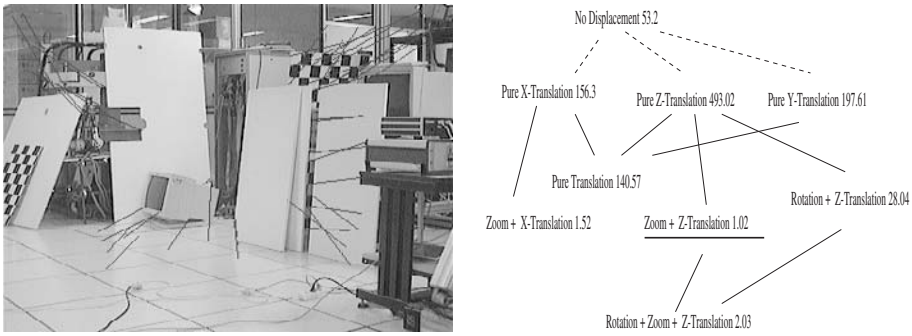


Figure 3.5. *Partial view of the hierarchy of models for a specific displacement, i.e. a zoom of the camera. For details, see the text*

In both cases, the model corresponds to the type of expected movement. In [LIN 99] other experimental results have been obtained with manual movements of the camera, which qualitatively corresponded to a specific movement; in each of the cases, they have been estimated by a coherent model with respect to the displacement being carried out.

As a complement to this series of experiments, from a small hierarchy of models [GAS 96] has shown that with this method we can separate different objects of the scene which present planar structures, as illustrated in Figure 3.6. On the left of Figure 3.6, we see the segmentation of planar structures, including the bottom of the plane. This allows us to estimate the plane at the horizon and thus the rotational part of the movement. On the right, the same system helps to separate two objects in movement, each vehicle being seen as a planar object. This type of approach has also been used to drive a vehicle or for other visual tasks [VIÉ 96c].

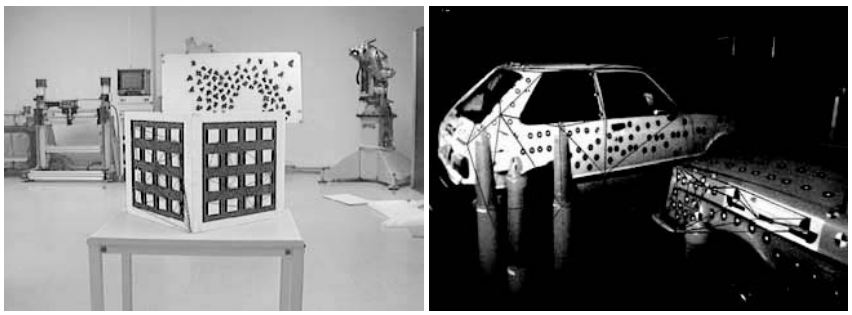


Figure 3.6. *Detection of planar structures: estimation by parametric modeling in a complex environment of different movements (on the right) or view planes (on the left)*

These results are characteristic of the estimation methods, which are all based on the kind of random selection methods refined by "non-linear least squares" optimization [CHO 99, MEE 91, TAU 91, KAN 96, ZHA 97].

3.5.3. *Application at the localization of a plane*

Calibrations from planes have been recently studied in [STU 99]. Assuming that the position of the points in the plane is *a priori* known, we try, from its observation in the retina, to estimate the position of this plane in space and possibly to estimate the camera intrinsic parameters. The representation of the projection of a point in this plane in the retina can be obtained from a homography [FAU 93]. Indeed, a 3D point $\mathbf{M}$ expressed in the absolute index $\mathcal{R}$ and belonging to the plane $\mathcal{P}$ of normal $\mathbf{n}$ ($\|\mathbf{n}\| = 1$), situated at a distance d from the origin of the index verifies:

$$\mathbf{n}^T \mathbf{M} = d \Longleftrightarrow \frac{\mathbf{n}^T \mathbf{M}}{d} = 1 \tag{3.41}$$

The projection of this point is:

$$Z' \mathbf{m}' = \mathbf{A}\, \mathbf{M}' \tag{3.42}$$

where $\mathbf{M}'$ is the point $\mathbf{M}$ expressed in the index $\mathcal{R}'$ connected to the camera. By explicitly writing the change of index between $\mathcal{R}$ and $\mathcal{R}'$, and from rotation $\mathbf{R} = \begin{pmatrix} \mathbf{r}_0^T \\ \mathbf{r}_1^T \\ \mathbf{r}_2^T \end{pmatrix}$ and translation $\mathbf{t}$, we obtain:

$$\mathbf{m}' \equiv \mathbf{A} \left[\mathbf{R}\, \mathbf{M} + \mathbf{t}\right]$$

$$\mathbf{m}' \equiv \mathbf{A} \left[\mathbf{R} + \frac{\mathbf{t}\, \mathbf{n}^T}{d}\right] \mathbf{M}$$

$$\mathbf{m}' \equiv \mathbf{H}\, \mathbf{M}$$

More precisely and without loss of generality, we can choose $\mathcal{R}$ in such a way that $\mathbf{n} = (0,\, 0,\, 1)^T$ and $d = 1$ (study of the equation plane $Z = 1$). Thus, we obtain:

$$\lambda\, \mathbf{H} = \mathbf{A} \left[\mathbf{R} + \mathbf{t}\, \mathbf{z}^T\right] \tag{3.43}$$

From a view, we can then estimate the homography $\mathbf{H}$, i.e. 9 equations parametrized by 10 unknown quantities ($\lambda + \underbrace{\mathbf{A}}_{3} + \underbrace{\mathbf{R}}_{3} + \underbrace{\mathbf{t}}_{3}$).

In the case where the intrinsic parameters are known, we can determine $\mathbf{R}$ and $\mathbf{t}$ (position of the plane) from (3.43) by removing λ. Indeed, from the quadratic constraints on the vectors of the rotation matrix, we can write:

$$\begin{cases} \mathbf{t} = \lambda\, \mathbf{A}^{-1}\, \mathbf{H}\, \mathbf{z} - \mathbf{R}\, \mathbf{z} \\ \mathbf{r}_0 = \lambda\, \mathbf{A}^{-1}\, \mathbf{H}\, \mathbf{x} \quad \text{with } \mathbf{x} = (1,0,0)^T \\ \mathbf{r}_1 = \lambda\, \mathbf{A}^{-1}\, \mathbf{H}\, \mathbf{y} \quad \text{with } \mathbf{y} = (0,1,0)^T \\ \mathbf{r}_2 = \mathbf{r}_0 \wedge \mathbf{r}_1 \\ \lambda = \sqrt{\dfrac{k_x + k_y}{k_x^2 + k_y^2}} \quad \text{with } k_x = \mathbf{x}^T\, \mathbf{H}^T\, \mathbf{A}^{-T}\, \mathbf{A}^{-1} \mathbf{H}\, \mathbf{x} \\ \qquad\qquad\qquad\quad \text{and } k_y = \mathbf{y}^T\, \mathbf{H}^T\, \mathbf{A}^{-T}\, \mathbf{A}^{-1} \mathbf{H}\, \mathbf{y} \end{cases} \tag{3.44}$$

Having estimated 7 parameters, there are two equations, which are a function of the intrinsic parameters.

Factorized Kruppa equations

We describe these equations from the quadratic constraints on $\mathbf{r}_0$ and $\mathbf{r}_1$ and we obtain:

$$\mathbf{r}_0^T\, \mathbf{r}_0 = \mathbf{r}_1^T\, \mathbf{r}_1 \quad \text{and} \quad \mathbf{r}_0^T\, \mathbf{r}_1 = 0 \tag{3.45}$$

i.e.:

$$\mathbf{x}^T\, \mathbf{H}^T \mathbf{K}^{-1}\, \mathbf{H}\, \mathbf{x} = \mathbf{y}^T\, \mathbf{H}^T \mathbf{K}^{-1}\, \mathbf{H}\, \mathbf{y} \quad \text{and} \quad \mathbf{x}^T\, \mathbf{H}^T \mathbf{K}^{-1}\, \mathbf{H}\, \mathbf{y} = 0$$

with:

$$\mathbf{K}^{-1} = \left(\mathbf{A}\, \mathbf{A}^T\right)^{-1} \equiv \begin{pmatrix} 1 & 0 & -u_0 \\ 0 & 1 & -v_0 \\ -u_0 & -v_0 & u_0^2 + v_0^2 + f^2 \end{pmatrix}$$

These equations are linear in f^2 and quadratic in (u_0, v_0).

Hence, we can either:

– use a system where the focal distance (in pixels) is known and, as a consequence, estimate the coordinates of the principal point (u_0 and v_0); or

– calculate f from u_0 and from v_0, since the position of the principal point is determined by another method [LAV 93].

Using synthetic data, drawn randomly (position of the plane and intrinsic parameters), we note that the resolution of these equations is unstable, as presented in Figure 3.7 in the case where we try to estimate the focal length according to the principal point.

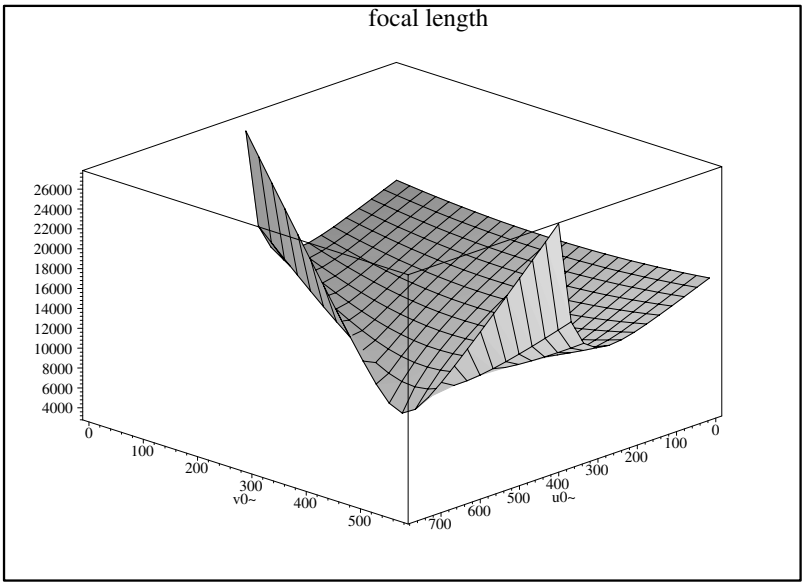


Figure 3.7. *Illustration of the instability of the resolution*

Case of several images

Now let us consider the case of a zoom where we observe a plane with different focal lengths. As earlier, we model the zoom only by assuming that rotation $\mathbf{R}$ does not vary during the zoom. In this case we have, if N is the number of images, $9N$ equations and $7N + 3$ unknowns: $\underbrace{\lambda_i}_{N} + \underbrace{\mathbf{A}_i}_{3N} + \underbrace{\mathbf{R}}_{3} + \underbrace{\mathbf{t}_i}_{3N}$. However, these equations are not independent since if $\mathbf{H}_i$ is the homography for the image i, we have:

– $\forall i, j\ \mathbf{H}_i[3,1] = \mathbf{H}_j[3,1]$

– $\forall i, j\ \mathbf{H}_i[3,2] = \mathbf{H}_j[3,2]$

– a third non-linear constraint.[2]

Then, there remain only $6N + 3$ independent equations for $6N + 4$ unknowns since we can, from the constraints, fix a common scale factor for all homographies. We are not able to calibrate the camera by observing a fixed plane with a zoom.

Specific cases

Given that the equations connecting the intrinsic parameters of the system depend directly on the constraints of the rotation matrix, it is suitable to consider certain specific cases when, for example, the plane is vertical, horizontal, or fronto-parallel. In the latter situation, where the rotation is around the optical axis[3], the homography is of the following form:

$$\mathbf{H}_z \equiv \begin{pmatrix} f\cos(\alpha) & -f\sin(\alpha) & \big(ft_0 + u_0\big(1+t_2\big)\big) \\ f\sin(\alpha) & f\cos(\alpha) & \big(ft_1 + v_0\big(1+t_2\big)\big) \\ 0 & 0 & \big(1+t_2\big) \end{pmatrix} \tag{3.46}$$

Thus, there are $7 + 1$ unknowns and $5 + 1$ equations. Moreover, we have only one quadratic equation according to f and the scale factor k. As a consequence, such degenerated situations have to be taken into account in a calibration application. If we now consider a rotation around the axis y, the

2.

$$\begin{aligned} \forall i,j \quad & -\mathbf{H}_j[3,1]^2\mathbf{H}_j[1,2]\mathbf{H}_i[2,2] + \mathbf{H}_j[3,1]\mathbf{H}_j[1,2]\mathbf{H}_i[2,1]\mathbf{H}_j[3,2] \\ & + \mathbf{H}_j[1,1]\mathbf{H}_j[3,2]\mathbf{H}_j[3,1]\mathbf{H}_i[2,2] - \mathbf{H}_j[1,1]\mathbf{H}_j[3,2]^2\mathbf{H}_i[2,1] \\ & + \mathbf{H}_i[1,2]\mathbf{H}_j[3,1]^2\mathbf{H}_j[2,2] - \mathbf{H}_i[1,2]\mathbf{H}_j[3,1]\mathbf{H}_j[2,1]\mathbf{H}_j[3,2] \\ & - \mathbf{H}_i[1,1]\mathbf{H}_j[3,2]\mathbf{H}_j[3,1]\mathbf{H}_j[2,2] + \mathbf{H}_i[1,1]\mathbf{H}_j[3,2]^2\mathbf{H}_j[2,1] \\ & = 0 \end{aligned}$$

between $\mathbf{H}_i$ and $\mathbf{H}_j$.

3. As a convention, the positions of the points, known *a priori*, in the (2D) plane are then expressed in an index, such that, the corresponding 3D points belong to the equation plane $Z = 1$.

homography is written in the form:

$$\mathbf{H}_y \equiv \begin{pmatrix} f\cos(\alpha) - u_0\sin(\alpha) & 0 & f\big(\sin(\alpha)+t_0\big) + u_0\big(\cos(\alpha)+t_2\big) \\ -v_0\sin(\alpha) & f & ft_1 + v_0\big(\cos(\alpha)+t_2\big) \\ -\sin(\alpha) & 0 & \cos(\alpha)+t_2 \end{pmatrix} \tag{3.47}$$

The specific form of this rotation is used in section 3.5.3.1 to calibrate the system.

Minimization of a criterion based on the gradient of the image

Traditional estimation of algorithms of movement are based on the detection of outlines or corners and then put in correspondence with correlation algorithms. From the correspondence of these points, a criterion generally minimizing the distance between a point of the second image and the prediction of its position from the point of the first image is minimized. A more direct method, described in [ROB 95a], consists of directly maximizing the sum of the slope of the re-projected points. Indeed, the slope of a bend point is a maximum local and, in the case of estimation of a homography, for example, the slope of the re-projected point $\nabla(\mathbf{m_i})\nabla(\mathbf{m_i}) = \nabla(\mathbf{H}(\mathbf{x})\cdot\mathbf{M_i})$ is also a local maximum as represented in Figure 3.8. We then define a function of "cost" as a sum of the squares of the slope of the re-projected points:

$$\mathcal{C}(\mathbf{x}) = \sum_i \big\|\nabla\big(\mathbf{H}(\mathbf{x})\cdot\mathbf{M}_i\big)\big\|^2. \tag{3.48}$$

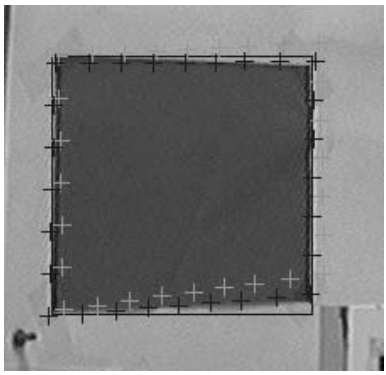


Figure 3.8. *Minimization of a criterion based on the slope of the image*

Figure 3.9. *Usage of a rotating table to generate singular movements in rotation*

3.5.3.1. *Rotation in pitch and calibration from a plane*

To illustrate the importance of the special cases, we consider a sequence of images obtained by using a rotating table. We have 8 images of a plane in rotation on an axis parallel to the y axis of the camera and the homography obtained must be in the form described by equation (3.47).

If α is different from 0, the ratio $\frac{\mathbf{H}_y[2,1]}{\mathbf{H}_y[3,1]}$ allows us to estimate v_0, as represented in Figure 3.10. Without any assumption on the angle of rotation, we can, from $\mathbf{H}_x$, also detect when the plane is fronto-parallel. Indeed, in this case, we have $\mathbf{H}[3,1] = \mathbf{H}[3,2] = 0$. Moreover, if we can control the rotation of the plane, either from a rotating table or from a turret, we can position the

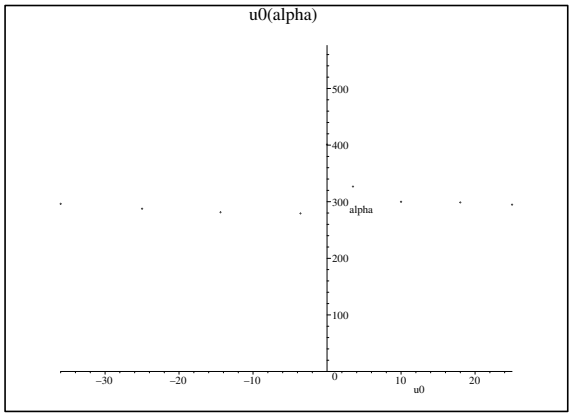


Figure 3.10. *Estimation of v_0 during a rotation*

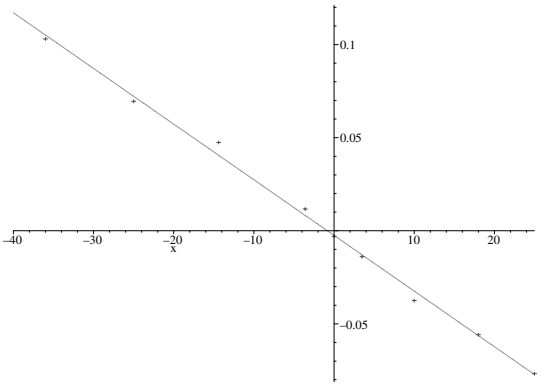


Figure 3.11. *Relation between the angle of rotation and* $\mathbf{H}[2,1]$

plane in such a way that it is fronto-parallel to the camera without knowing the intrinsic parameters (see Figure 3.11).

Moreover, Taylor series expansion of H[2,1] gives us:

$$\mathbf{H}[2,1] \equiv -v_0\alpha + O(\alpha^3)$$

Figure 3.11 represents the results obtained by interpolating a line to the least squares according to α. The coefficient of linear correlation obtained is .99826.

Finally, a decomposition of $\mathbf{H}[1,1]$ in Taylor series yields:

$$\mathbf{H}[1,1] \equiv f - u_0\alpha - \frac{1}{2}f\alpha^2 + O(\alpha^3).$$

This quadratic equation according to α leads to an extremum when $\alpha = -\frac{u_0}{f}$ (see Figure 3.12).

An interpolation in the least squares sense yields:

$$\mathbf{H}[1,1](\alpha) \approx -.0001527491602\alpha^2 - .003552077403\alpha + 1.005639344$$

After deriving this quadratic equation, we obtain $\alpha_{\max} \approx -10.6^0$. By considering a focal distance of approximately 2,300 of approximately 2,300 pixels, we obtain $u_0 = 425$.

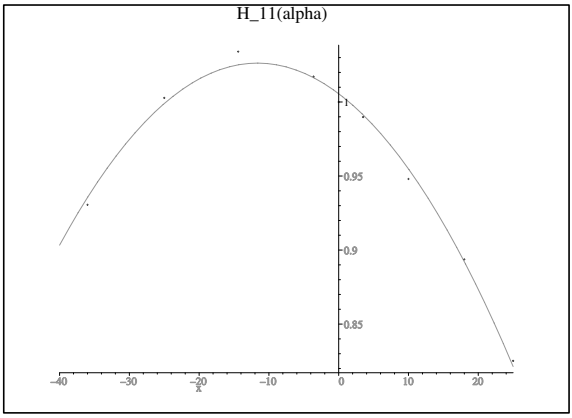


Figure 3.12. *Quadratic variations of* $\mathbf{H}[1,1](\alpha)$

Calibrated structural analysis with a zoom

In the case where the matrix of intrinsic parameters $\mathbf{A}$ as well as $\mathbf{H}_\infty$ is known, we can, in order to estimate the related homography, minimize a criterion based on three parameters: $\frac{\mathbf{t}_z}{d}$ and two angles θ and ϕ corresponding to the normal $\mathbf{n}$ expressed in spherical coordinates. Since, in the case of a zoom, the translation is small with respect to the distance between the plane and the camera d, such a criterion is non-linear and unstable at the same time. However, in the first stage, and in order to start the minimization of the criterion, we can estimate $\frac{\mathbf{t}_z}{d}$ by considering that the plane is fronto-parallel and then minimize the criterion according to θ and ϕ.

In order to experiment this minimization, we use a sequence during which the focal distance varies from 1,200 to 2,300 pixels (Figure 3.13). After an estimation of $\frac{\mathbf{t}_z}{d}$ for the two planes (manual selection), we obtain the results presented in Figure 3.13 by considering the first and the last image of the sequence.

Now, by minimizing the criterion compared to θ and ϕ, we get the following results for the fronto-parallel plane:

$$\theta = 0.607 \quad \text{and} \quad \phi = 0.101 \Longrightarrow \mathbf{n} = (.0830,\ .0577,\ .9949)$$

The normal thus estimated is very much along the optical axis and by considering the second plane, we obtain:

$$\theta = 0.172 \quad \text{and} \quad \phi = 0.430 \Longrightarrow \mathbf{n} = (.411,\ .0712,\ .909)$$

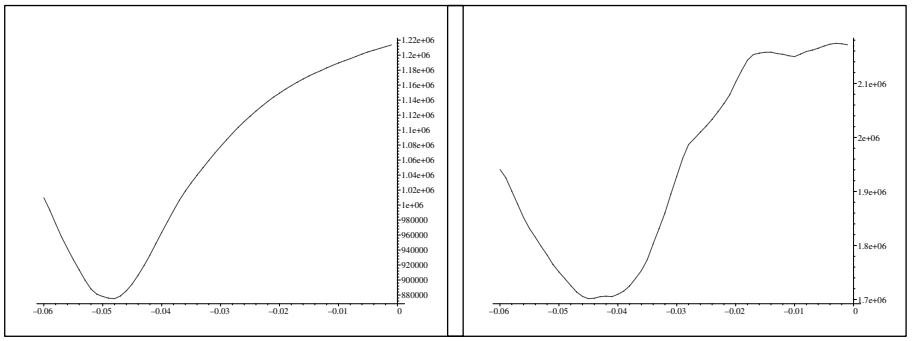


Figure 3.13. *Representation of the criterion according to* $\frac{\mathbf{t}_z}{d}$

As expected (see Figure 3.14), we note a significant increase in the first coefficient of the normal as compared to the fronto-parallel plane.

Figure 3.14. *Estimation of the proximity by local refining*

Moreover, in spite of the fact that this estimation becomes very unstable when we consider zooms of weak magnitude, it becomes stable again by considering the translations carried out by a robotized precision system.

Non-calibrated case

Now, we minimize a criterion as compared to the 8 coefficients of the homography $\mathbf{H}$ and use a sequence of images captured during a zoom of 2,250 pixels to 1,550 pixels on a scene containing several planes (Figure 3.15).

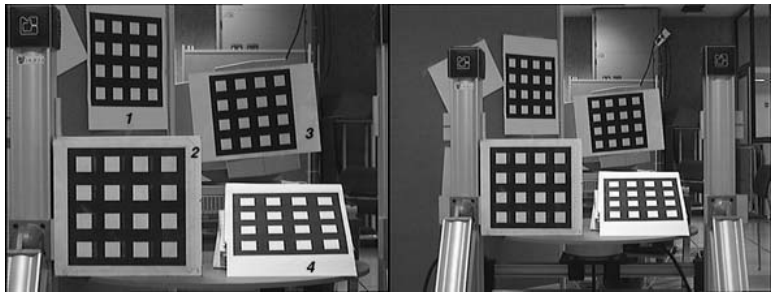


Figure 3.15. *Non-calibrated planes and zoom*

We obtain the following results.

Plane number	Homography
1	$\begin{pmatrix} .6860 & -0.0019 & 119.4010 \\ -0.0013 & 0.6891 & 87.7284 \\ -2.3880E-6 & -2.8038E-7 & 1 \end{pmatrix}$
2	$\begin{pmatrix} 0.6908 & -6.0502E-4 & 118.0713 \\ -3.1525E-4 & 0.6901 & 87.2886 \\ -1.4058E-6 & 4.3485E-8 & 1 \end{pmatrix}$
3	$\begin{pmatrix} 0.6870 & -0.0025 & 119.6178 \\ -0.0012 & 0.6880 & 88.0802 \\ -1.6964E-6 & -2.7720E-6 & 1 \end{pmatrix}$
4	$\begin{pmatrix} 0.6769 & 0.0149 & 119.2439 \\ -0.0088 & 0.7023 & 87.0714 \\ -2.5382E-5 & 3.3070E-5 & 1 \end{pmatrix}$

Table 3.3. *Results: uncalibrated case*

From equation (3.35), we can conclude that planes 3 and 2 are fronto-parallel and that plane number 3 is the farthest. The normal of plane number 4 has a significant component following the y axis. Finally, the comparison of $\mathbf{H}_1^{00}$ and $\mathbf{H}_1^{11}$ allows us to affirm that plane 1 is not fronto-parallel. In this case, the difference is smaller since the distance between the camera and the plane becomes significant and the coefficients connected to the normal are multiplied by $\frac{1}{d}$ as well as by the magnitude of

the translation in the homography. Hence, the minimization becomes unstable when we consider either small focal length variations or planes far away from the camera.

Depth and zoom

From the fronto-parallel plane of Figure 3.14, we estimate $\mathbf{H}_\infty$. Hence, we can, for each strong sloping point of the first image, minimize a criterion from $\delta_{\mathbf{m},\mathbf{H}_\infty}$. We assume that the translation is small as compared to the depth of the 3D point. When the user selects a point, we look for the point with the stronger slope in the neighborhood and finally minimize the non-linear criterion. Take for example the four points in Figure 3.14.

For the first two points, we obtain the results in Figure 3.16, representing the value of the criterion according to $\delta_{\mathbf{m},\mathbf{H}_\infty}$.

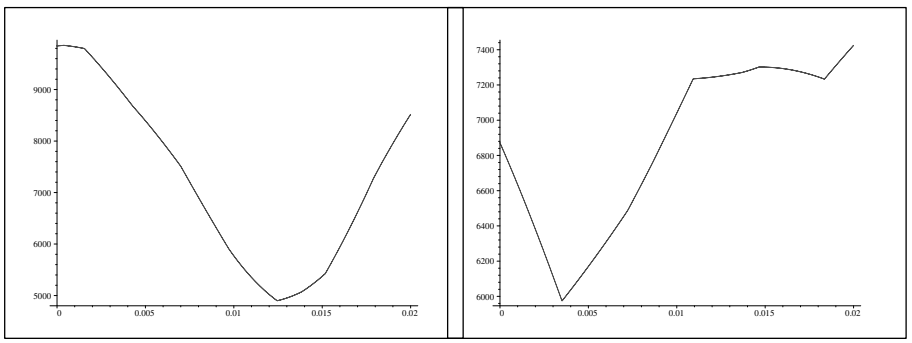


Figure 3.16. *Variation of criterion according to* $\delta_{\mathbf{m},\mathbf{H}_\infty}$ *for points* 1 *and* 2

As expected, the proximity of point number 1 is more significant than that of point number 2. Point number three has zero proximity and is thus considered to be at infinity, as described in Figure 3.17.

Finally, if we minimize the criterion for point number 4, we highlight a problem, which appears when the curve and the line defined by $\mathbf{H}_\infty$ $\mathbf{m}$ and $\mathbf{m}_0$ have the same direction. The points projected in the second image are then displaced along the curve and the minimum value of $\mathcal{C}(\delta_{\mathbf{m},\mathbf{H}_\infty})$ has no relation with the proximity of the point but has a relation with the structure of the curve. The calculation of the orientation of the slope makes it possible to take into account this "sliding" effect.

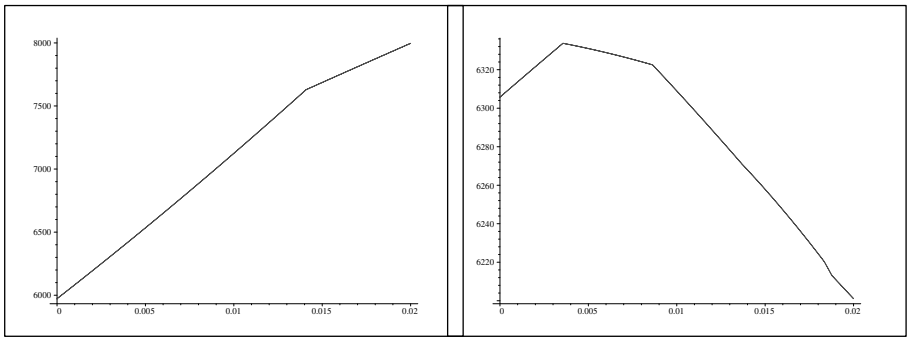


Figure 3.17. *Variation of criterion according to* $\delta_{\mathbf{m},\mathbf{H}_\infty}$ *for points number* 3 *and* 4

3.6. Conclusion

Within the framework of visual perception related to robotic problems, we are interested in the analysis of non-calibrated monocular sequences of images, by taking into account specific cases (camera models, evolution of internal parameters of the camera, displacement of objects in the scene of the camera, structure of the scene) leading to specific equations. Thus, we have seen that such singularities generally make it possible to recover the movement or the structure, and this is always done with better precision, as this type of approach reduces the number of parameters brought into play. A step further, it seems important to be able to detect and manage the cinematic and geometric properties that arise from it. An examination of these specific cases on two images, and then on a sequence of images, shows that the complexity of the problem requires an adapted processing.

3.7. Bibliography

[AGA 98] DE AGAPITO L., HAYMAN E. and REID I.L., "Self-calibration of a rotating camera with varying intrinsic parameters", *British Machine Vision Conference*, Southampton, UK, BMVA Press, September 1998.

[ALO 90] ALOIMONOS J., "Perspective Approximations", *Image and Vision Computing*, vol. 8, no. 3, p. 179–192, 1990.

[ARM 94] ARMSTRONG M., ZISSERMAN A. and BEARDSLEY P., "Euclidean structure from uncalibrated images", in HANCOCK E. (Ed.), *Proceedings of the 5th British Machine Vision Conference*, York, UK, BMVA Press, p. 508–518, September 1994.

[ARM 96] ARMSTRONG M., ZISSERMAN A. and HARTLEY R., "Self-calibration from Image Triplets", *Fourth European Conference on Computer Vision*, p. 3–16, April 1996.

[BEA 95] BEARDSLEY P.A., REID I.D., ZISSERMAN A. and MURRAY D.W., "Active visual navigation using non-metric structure", *Proceedings of the 5th International Conference on Computer Vision* [icc95], p. 58–64, June 1995.

[BOB 94] BOBET P., Tête stéréoscopique. Réflexes oculaires et Vision, PhD Thesis, Institut National Polytechnique de Grenoble, France, 1994.

[BOL 81] BOLLES R.C. and FISCHLER M.A., "A RANSAC-based approach to model fitting and its application to finding cylinders in range data", *International Joint Conference on Artificial Intelligence*, Vancouver, Canada, p. 637–643, August 1981.

[BOU 94] BOUFAMA B., WEINSHALL D. and WERMAN M., "Shape from motion algorithms: a comparative analysis of scaled orthography and perspective", in Eklundh [EKL 94], p. 199–204, May 1994.

[BOU 97] BOUGNOUX S. and ROBERT L., "TotalCalib: a fast and reliable system for off-line calibration of image sequences", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, June 1997, The Demo Session.

[BOU 98] BOUGNOUX S., "From Projective to Euclidean Space under any Practical Situation, a Criticism of Self-Calibration", IEEE *International Conference on Computer Vision*, p. 790–796, 1998.

[BRO 96] BROOKS M.J., DE AGAPITO L., HUYNG D.Q. and BAUMELA L., "Direct Methods for Self-Calibration of a Moving Stereo Head", *Fourth European Conference on Computer Vision*, vol. II, p. 415–426, April 1996.

[BUX 96] BUXTON B. (Ed.), *ECCV '96, Fourth European Conference on Computer Vision*, Cambridge, UK, April 1996.

[CHO 99] CHOJNACKI W., BROOKS M.J. and VAN DEN HENGEL A., "Rationalising Kanatani's method of renormalization in computer vision", *Statistical Methods for Image Processing*, Uppsala, Sweden, August, p. 61–63, 1999.

[CLA 97] CLARKE J.C., Applications of Sequence Geometry to Visual Motion, PhD Thesis, University of Oxford, 1997.

[DEM 89] DEMENTHON D. and DAVIS L.S., Exact and approximate solutions to the three-point perspective problem, Report no. CAR-TR-471, Computer Vision Laboratory, University of Maryland, 1989.

[EKL 94] EKLUNDH J.-O. (Ed.), vol. 800-801 of *Lecture Notes in Computer Science*, Stockholm, Sweden, Springer-Verlag, May 1994.

[ENC 94] ENCISO R., VIÉ VILLE T. and FAUGERAS O., "Approximation du Changement de Focale et de Mise au Point par une Transformation Affine à Trois Paramètres", *Traitement du Signal*, vol. 11, no. 5, p. 361–372, 1994.

[ENC 95] ENCISO R., Auto-Calibration des Capteurs Visuels Actifs. Reconstruction 3D Active., PhD Thesis, Paris XI Orsay University, December 1995.

[ENC 96] ENCISO R., ZISSERMAN A. and VIÉVILLE T., "An affine solution to the Euclidean calibration while using a zoom lens", *Workshop ALCATECH*, 21–27, Denmark, July 1996.

[FAU 93] FAUGERAS O., *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, 1993.

[GAS 96] GASPARD F. and VIÉVILLE T., Hierarchical Visual Perception without Calibration, RR no. 3002, INRIA Sophia-Antipolis, October 1996.

[GAS 97] GASPARD F., ZISSERMAN A. and VIÉVILLE T., "Le zoom comme outil de calibration affine d'une caméra", *Journées ORASIS'97*, p. 27–38, October 1997.

[GAS 00] GASPARD F. and VIÉVILLE T., "Non Linear Minimization and Visual Localization of a Plane", *The 6th International Conference on Information Systems, Analysis and Synthesis*, 2000.

[HAR 88] HARRIS C. and STEPHENS M., "A combined Corner and Edge Detector", *Proc. 4th Alvey Vision Conf.*, p. 189–192, 1988.

[HAR 94a] HARTLEY R., "Projective reconstruction and invariants from multiple images", *PAMI*, vol. 16, no. 10, p. 1036–1040, 1994.

[HAR 94b] HARTLEY R., "Self-Calibration from Multiple Views with a Rotating Camera", *Third European Conference on Computer Vision*, p. 471–478, May 1994.

[HAR 95] HARTLEY R.I., "A linear method for reconstruction from lines and points", *Proceedings of the 5th International Conference on Computer Vision* [icc95], p. 882–887, June 1995.

[HAY 96] HAYMAN E., RIED I.D. and MURRAY D.W., "Zooming while Tracking using Affine Transfer", *Proceedings of the British Machine Vision Conference*, 1996.

[HEY 97] HEYDEN A. and ASTROM K., "Euclidean reconstruction from image sequences with varying and unknown focal length and principal point", *Comp. Vision and Pattern Rec.*, IEEE Computer Society Press, p. 438–443, 1997.

[HOR 94] HORAUD R., CHRISTY S. and DORNAIKA F., Object Pose: The Link between Weak Perspective, Para-Perspective, and Full Perpective, Report no. 2356, INRIA, September 1994.

[HOR 97] HORAUD R., CHRISTY S. and MOHR R., "Euclidean Reconstruction and Affine Camera Calibration Using Controlled Robot Motions", *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Grenoble, France, September 1997.

[HOR 97] HORAUD R., DORNAIKA F., LAMIROY B. and CHRISTY S., "Object Pose: The Link between Weak Perspective, para-perspective, and Full Perspective", *IJCV*, vol. 22, no. 2, p. 173–189, 1997.

[HUB 81] HUBER P., *Robust Statistics*, John Wiley & Sons, New York, 1981. [icc95] Boston, MA, IEEE Computer Society Press, June 1995. [icp94] Jerusalem, Israel, Computer Society Press, October 1994.

[KAN 92] KANATANI K., *Geometric Computation for Machine Vision*, Oxford University Press, 1992.

[KAN 96] KANATANI K., "Automatic Singularity Test for Motion Analysis by an Information Criterion", Buxton [BUX 96], p. 697–708, April 1996.

[KOE 84] KOENDERINCK, "The structure of images", *Biol. Cybern.*, vol. 50, p. 363–370, 84.

[LAV 92] LAVEST J., Stéréovision axiale par zoom pour la robotique, PhD Thesis, University of Blaise Pascal, Clermont-Ferrand, France, 1992.

[LAV 93] LAVEST J., RIVES G. and DHOME M., "3-D reconstruction by zooming", *IEEE Trans. on Robotics and Automation*, vol. 9, no. 2, p. 196–207, April 1993.

[LEE 99] LEEDAN Y. and MEER P., "Heteroscedastic regression in computer vision: problems with bilinear constraint", *IJCV*, vol. 37, no. 2, p. 127–150, 1999.

[LI 96] LI M., LAVEST J.-M., "Some Aspects of Zoom Lens Camera Calibration", *PAMI*, vol. 18, no. 11, p. 1105–1110, November 1996.

[LIN 95] LINGRAND D. and VIÉVILLE T., Dynamic Foveal 3D Sensing Using Affine Models, Report no. RR-2687, INRIA Sophia-Antipolis, October 1995.

[LIN 96] LINGRAND D. and VIÉVILLE T., "Dynamic Foveal 3D Sensing Using Affine Models", *Proceedings of the International Conference on Pattern Recognition*, vol. 1, Vienna, Austria, Computer Society Press, p. 810–814, August 1996.

[LIN 99] LINGRAND D., Analyse Adaptative du Mouvement dans des Séquences Monoculaires non Calibrées, PhD Thesis, University of Nice - Sophia Antipolis, INRIA, Sophia Antipolis, France, July 1999.

[LUO 93] LUONG Q.-T., DERICHE R., FAUGERAS O. and PAPADOPOULO T., On Determining the Fundamental Matrix: Analysis of Different Methods and Experimental Results, Report no. 1894, INRIA, 1993.

[LUO 95] LUONG Q.-T. and FAUGERAS O., "The Fundamental matrix: theory, algorithms, and stability analysis", *The International Journal of Computer Vision*, vol. 17, no. 1, p. 43–76, January 1995.

[LUO 96] LUONG Q. and VIÉVILLE T., "Canonical representations for the geometries of multiple projective views", *Computer Vision and Image Understanding*, vol. 64, no. 2, p. 193–229, 1996.

[MAY 92] MAYBANK S.J. and FAUGERAS O.D., "A Theory of Self-Calibration of a Moving Camera", *The International Journal of Computer Vision*, vol. 8, no. 2, p. 123–152, 1992.

[MEE 91] MEER P., MINTZ D., ROSENFELD A. and KIM D., "Robust Regression Methods for Computer Vision: A Review", *The International Journal of Computer Vision*, vol. 6, no. 1, p. 59–70, 1991.

[PAH 92] PAHLAVAN K., EKHLUND J.-O. and UHLIN T., "Integrating Primary Ocular Processes", SANDINI G. (Ed.), *Proc 2nd ECCV*, Santa Margherita, Italy, Springer-Verlag, p. 526–541, May 1992.

[POE 93] POELMAN C.J. and KANADE T., A para-perspective Factorization method for Shape and Motion Recovery, Report no. CMU-CS-93-219, Carnegie Mellon University, School of Computer Science, December 1993.

[POE 94] POELMAN C.J. and KANADE T., "A para-perspective Factorization for Shape and Motion Recovery", in Eklundh [EKL 94], p. 97–108, May 1994.

[POL 95] POLLEFEYS M., VAN GOOL L. and MOONS T., "Euclidean 3D reconstruction from stereo sequences with variable focal lenghts", *Proceedings of the 2nd Asian Conference on Computer Vision*, vol. II, Singapore, page 6, December 95.

[POL 96] POLLEFEYS M., VAN GOOL L. and PROESMANS M., "Euclidean 3D reconstruction from stereo sequences with variable focal lenghts", in Buxton [BUX 96], p. 31–42, April 1996.

[POL 97] POLLEFEYS M., KOCH R. and VAN GOOL L., Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, Report no. KUL/ESAT/MI2/9707, Katholieke Universiteit Leuven, August 1997.

[QUA 96] QUAN L., "Self-calibration of an affine camera from multiple views", *IJCV*, vol. 19, no. 1, p. 93–105, May 1996.

[ROB 93] ROBERT L. and FAUGERAS O., "Relative 3D Positioning and 3D Convex Hull Computation from a Weakly Calibrated Stereo Pair", *Proceedings of the 4th International Conference on Computer Vision*, Berlin, Germany, IEEE Computer Society Press, p. 540–544, May 1993 and INRIA Technical Report 2349.

[ROB 95a] ROBERT L., "Camera Calibration Without Feature Extraction", *Computer Vision, Graphics, and Image Processing*, vol. 63, no. 2, p. 314–325, March 1995 and INRIA Technical Report 2204.

[ROB 95b] ROBERT L. and FAUGERAS O., "Relative 3-D Positioning and 3-D Convex Hull Computation From A Weakly Calibrated Stereo Pair", *Image and Vision Computing*, vol. 13, no. 3, p. 189–197, 1995 and INRIA Technical Report 2349.

[ROU 87] ROUSSEEUW P. and LEROY A., *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, 1987.

[SHA 93] SHAPIRO L. and BRADY M., Rejecting outliers and estimating errors in an orthogonal regression framework, Tech. Report OUEL no. 1974/93, Dept. Engineering Science, University of Oxford, February 1993.

[SHA 94a] SHASHUA A. and NAVAB N., "Relative Affine Structure: Theory and Application to 3D Reconstruction from Perspective Views", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Seattle, WA, IEEE, June 1994.

[SHA 94b] SHASHUA A., "Projective structure from uncalibrated images: structure from motion and recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, p. 778–790, 1994.

[SOA 95] SOATTO S. and PERONA P., "Dynamic Rigid Motion Estimation From Weak Perspective", *Proceedings of the 5th International Conference on Computer Vision* [icc95], p. 321–328, June 1995.

[STU 96] STURM P. and TRIGGS B., "A Factorization Based Algorithm for Multi-Image Projective Structure and Motion", in Buxton [BUX 96], p. 709–720, April 1996.

[STU 97a] STURM P., "Critical motion sequences for monocular self-calibration and uncalibrated Euclidean reconstruction", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Puerto Rico, USA, p. 1100–1105, 1997.

[STU 97b] STURM P., Vision 3D non calibrée. Contributions à la reconstruction projective et étude des mouvements critiques pour l'auto-calibrage., PhD thesis, INPG, Grenoble, France, December 1997.

[STU 99] STURM P.F., MAYBANK S.J., "On Plane-Based Camera Calibration: A General Algorithm, Singularities, Applications", *CVPR - IEE Conference on Computer Vision and Pattern Recognition*, vol. I, p. 432–437, June 1999.

[TAU 91] TAUBIN G., "Estimation of Planar Curves, Surfaces, and Nonplanar Space Curves Defined by Implicit Equations with Applications to Edge and Range Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 11, p. 1115–1138, November 1991.

[TOM 91] TOMASI C., KANADE T., "Factoring Image Sequences into Shape and Motion", *Proc. IEEE Workshop on Visual Motion*, Princeton, NJ, p. 21–28, October 1991.

[TOM 92] TOMASI C. and KANADE T., "Shape and motion from image streams under orthography: a factorization method", *The International Journal of Computer Vision*, vol. 9, no. 2, p. 137–154, 1992.

[TOR 95] TORR P., ZISSERMAN A. and MAYBANK S., "Robust Detection of Degenerate Configurations for the Fundamental Matrix", *Proceedings of the 5th International Conference mon Computer Vision*, p. 1037–1042, June 1995.

[TOR 97] TORR P.H.S., MURRAY D.W., "The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix", *IJCV*, vol. 24, no. 3, p. 271–300, 1997.

[TOR 98a] TORR P.H.S., "Geometric motion segmentation and model selection", *Phil. Trans. R. Soc. Lond. A*, vol. 356, p. 1321–1340, 1998.

[TOR 98b] TORR P., FITZGIBBON A.W. and ZISSERMAN A., "Maintaining Multiple Motion Model Hypotheses Over Many Views to Recover Matching and Structure", SOCIETY I.C. (Ed.), *ICCV'98*, Bombay, India, IEEE Computer Society, Narosa Publishing House, p. 485–491, January 1998.

[TRI 98a] TRIGGS B., "Optimal Estimation of Matching Constraints", in KOCH R. and GOOL L.V. (Eds.), *Workshop on 3D Structure from Multiple Images of Large-scale Environments SMILE'98*, Lecture Notes in Computer Science, 1998.

[TRI 98b] TRIGGS B., "Autocalibration from Planar Scenes", in Springer (Ed.), *ECCV'98*, vol. I, p. 89–105, June 1998.

[VAN 94] VAN GOOL L., MOONS T., PROESMANS M. and VAN DIEST M., "Affine reconstruction from perspective image pairs obtained by a translating camera", *Proceedings of the International Conference on Pattern Recognition*, October 1994.

[VIÉ 94a] VIÉVILLE T., "Autocalibration of Visual Sensor Parameters on a Robotic Head", *Image and Vision Computing*, vol. 12, 1994.

[VIÉ 94b] VIÉVILLE T., CLERGUE E., ENCISO R. and MATHIEU H., "Experimenting 3-D Vision on a Robotic Head", *Proceedings of the International Conference on Pattern Recognition*, p. 739–743, October 1994.

[VIÉ 95a] VIÉVILLE T., CLERGUE E., ENCISO R. and MATHIEU H., "Experimentating with 3-D vision on a robotic head", *Robotics and Autonomous Systems*, vol. 14, no. 1, p. 1–27, 1995.

[VIÉ 95b] VIÉVILLE T., FACAO P. and CLERGUE E., "Computation of ego-motion using the Vertical Cue", *Machine Vision and Applications*, vol. 8, no. 1, p. 41–52, 1995.

[VIÉ 96a] VIÉVILLE T., FAUGERAS O.D. and LUONG Q.-T., "Motion of Points and Lines in the Uncalibrated Case", *The International Journal of Computer Vision*, vol. 17, no. 1, p. 7–42, January 1996.

[VIÉ 96b] VIÉVILLE T. and LINGRAND D., "Using Singular Displacements for Uncalibrated Monocular Visual Systems", *4th ECCV*, vol. 2, p. 207–216, April 1996.

[VIÉ 96c] VIÉVILLE T., ZELLER C. and ROBERT L., "Using Collineations to Compute Motion and Structure in an Uncalibrated Image Sequence", *The International Journal of Computer Vision*, vol. 20, no. 3, p. 213–242, 1996.

[VIÉ 99] VIÉVILLE T. and LINGRAND D., "Using Specific Displacements to Analyze Motion without Calibration", *IJCV*, vol. 31, no. 1, p. 5–29, 1999.

[VIÉ 00] VIÉVILLE T., DROULEZ J. and PEH C.-H., "How do we perceive the eye intrinsic parameters?", *INRIA* 2000.

[VIÉ 01] VIEVILLE T., LINGRAND D. and GASPARD F., "Implementing a variant of the Kanatani's estimation method", *Int. J. Comp. Vision*, vol. 44, no. 1, p. 41–64, 2001.

[WIL 93] WILLSON R. and SHAFER S., "What is the center of the image?", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, New York, IEEE Computer Society, IEEE, p. 670–671, June 1993.

[WIL 94] WILLSON R.G., Modeling and Calibration of Automated Zoom Lenses, PhD Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1994, CMU-RI-TR-94-03.

[WIL 96] WILES C. and BRADY M., "Ground plane motion camera models", in Buxton [BUX 96], p. 234–247, April 1996.

[ZEL 94] ZELLER C. and FAUGERAS O., "Applications of Non-Metric Vision to Some Visual Guided Tasks", *Proceedings of the International Conference on Pattern Recognition* [icp94], p. 132–136, October 1994, long version in INRIA Tech Report RR2308.

[ZHA 94] ZHANG Z., DERICHE R., FAUGERAS O. and LUONG Q.-T., "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry", *Artificial Intelligence Journal*, vol. 78, no. 1-2, p. 87–119, 1994, Appeared in October 1995, also INRIA Research Report No.2273, May 1994.

[ZHA 95] ZHANG Z., DERICHE R., FAUGERAS O. and LUONG Q.-T., "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry", *Artificial Intelligence Journal*, vol. 78, p. 87–119, October 1995.

[ZHA 97] ZHANG Z., "Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting", *Image and Vision Computing Journal*, vol. 15, no. 1, p. 59–76, 1997.

[ZIS 98] ZISSERMAN A., LIEBOWITZ D. and ARMSTRONG M., "Resolving Ambiguities in Auto-Calibration", *Philosophical Transactions of the Royal Society of London, SERIES A*, vol. 356, no. 1740, p. 1193–1211, 1998.

Part 2

This page intentionally left blank

Chapter 4

Localization Tools

4.1. Introduction

This chapter is dedicated to the problem of localization of objects by computer vision. It refers to relocating the spatial position of the CAD model of an observed object by a video sensor in such a way that the latter is in conformity with the contents of the analyzed images. The reader will later on find the description of the techniques, which make it possible to process the case of localization of a rigid object by monocular vision. Its formalism will then be extended to understand cases as diverse as multi-ocular localization and hand-eye calibration, by conducting research on the posture of articulated objects such as robotic arms.

This very general problem can be addressed in various ways according to the context: the simplest case being monocular localization (usage of a single camera) of a rigid object. This point has been extensively discussed in other works during the last two decades.

Different approaches can be classified according to:

– the manipulated projection model (orthographic [KAN 81], scaled orthographic [DEM 92], para-perspective [HOR 95a], perspective [LOW 85, HOR 87, HOR 89, DHO 89, DHO 90]);

– the nature of paired primitive (points [HOR 89, DEM 92], lines [KAN 81, LOW 85, HOR 87, DHO 89], circles [DHO 90]);

Chapter written by Michel DHOME and Jean-Thierry LAPRESTÉ.

– the resolution method (analytical [KAN 81, HOR 89, DHO 89, DHO 90], iterative [LOW 85, DHO 89, DEM 92, HOR 95a], by collection of hypotheses [HOR 87]).

From this vast choice, we have adopted the method first proposed by Lowe [LOW 85] for its elegance and especially for the possibility to adapt its formalism to various contexts.

Indeed, we will see how to chronologically understand the monocular localization of a rigid object, the multi-ocular localization of the same object, the study of the posture of an articulated object of "*hand manipulator*" type and hand-eye calibration, which is very useful when the off-set camera is mounted on a robotic system. With the adopted generic approach based on the minimization of a non-linear criterion, the initialization phase will be the subject of one section. Finally, we will present how an estimation of errors associated with different realized measurements can be obtained with the help of a technique of propagation of uncertainties.

4.2. Geometric modeling of a video camera

This section succinctly tackles the problem of geometric modeling of CCD sensors according to the commonly adopted formalism in the field of computer vision [BRO 71, FAI 87, TSA 86, FAU 87, HOR 95b].

4.2.1. *Pinhole model*

The projection model generally used to understand the process of formation of video images refers to the pinhole model or "*eye of a needle*" which corresponds to an approximation of the optical model with fine lenses. It is based on the assumption that all optical rays pass through a single point, called the "*optical center*", before hitting the photosensitive plate of the sensor (see Figure 4.1).

The intrinsic parameters associated with this camera model are:

1) focal distance f (*orthogonal distance of the optical center to the image plane*);

2) co-ordinates (u_0, v_0) of the principal point (*intersection of the optical axis and the image plane assumed to be orthogonal*);

3) horizontal and vertical dimensions of the pixels of the CCD matrix (dx, dy) or in practice their ratio $(\kappa = dy/dx)$ since a perspective projection is always defined as a close scale factor.

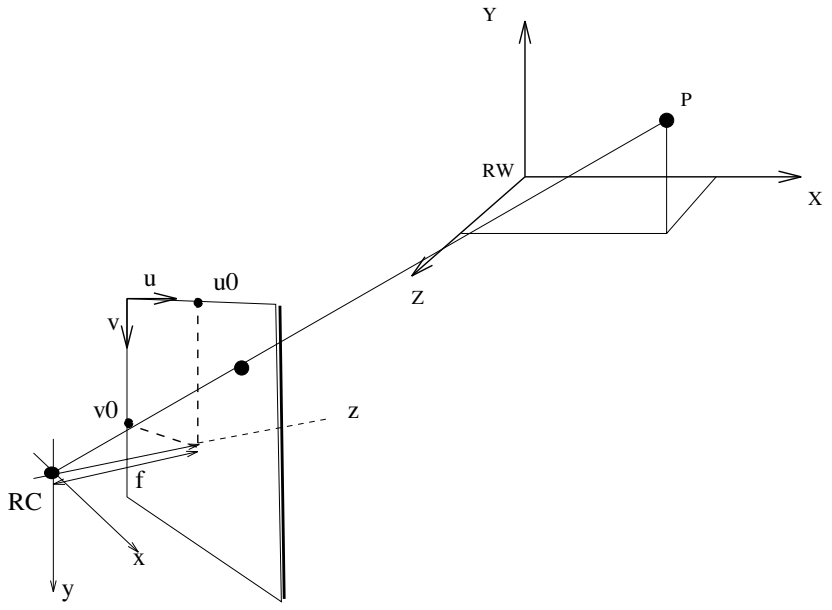


Figure 4.1. *The pinhole model and the system of co-ordinates used*

NOTE 4.1. The reader will note that to avoid manipulation of an inversed image, it is customary to represent the image in a plane situated in front of the optical center and distanced from the latter by a distance f.

4.2.2. *Perspective projection of a 3D point*

Let $P = (X, Y, Z)$ be a point defined in the world reference mark R_w; its counterpart in the camera reference mark R_c will be noted by $P' = (X', Y', Z')$. By assuming that the relative positioning of the reference mark R_w in reference mark R_c is represented by a parametric rotation matrix of Euler's three angles (α rotation on the x axis, β on the y axis and γ on the z axis) noted by $\mathbf{R}_{\alpha\beta\gamma}$ and a translation vector of components (u, v, w) noted by $\mathbf{T}_{uvw}$, the relation connecting P to P' is (these parameters are named as extrinsic parameters):

$$\begin{pmatrix} X' \\ Y' \\ Y' \end{pmatrix} = \left[\mathbf{R}_{\alpha\beta\gamma} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \mathbf{T}_{uvw} \right]$$

Let $p = (x, y, f)$ be the intersection of the image plane and the line passing through the optical center of the camera and the point P'. The respective co-ordinates of these points are connected by the following relations:

$$\begin{cases} x = f\dfrac{X'}{Z'} = f\dfrac{r_{11}X + r_{12}Y + r_{13}Z + u}{r_{31}X + r_{32}Y + r_{33}Z + w} \\ y = f\dfrac{Y'}{Z'} = f\dfrac{r_{21}X + r_{22}Y + r_{23}Z + v}{r_{31}X + r_{32}Y + r_{33}Z + w} \end{cases} \tag{4.1}$$

where the coefficients r_{ij} represent the different elements of the rotation matrix $\mathbf{R}_{\alpha\beta\gamma}$.

However, during their detection, the primitives are parametrized in the natural image reference mark R_n conventionally situated at the top-left corner of the latter (see Figure 4.1). Let p be a point in the image with co-ordinates (u, v) and (x, y, f) in R_n and R_c, respectively. The moving from one reference mark to another is defined by the following relations:

$$\begin{cases} x = (u - u_0) \\ y = (v - v_0)\kappa \end{cases} \tag{4.2}$$

Note that equations (4.1) and (4.2) impose a common unity, which is equal to the horizontal pixels of the image, to the co-ordinates (x, y, f) of point p.

Different intrinsic parameters of the pinhole model (f, u_0, v_0, κ) are obtained during the calibration stage of the sensor, which we will tackle again. For the moment, we will consider that the parametrization of primitive images is expressed in the camera reference mark with the help of system (4.2).

4.3. Localization of a voluminous object by monocular vision

4.3.1. *Introduction*

In this section, we describe a technique first proposed by Lowe [LOW 85] to calculate, from a video image, the relative position of a voluminous object as compared to the reference mark associated with the sensor, which observes it. We have referred to this approach for its formalism, which adapts to numerous situations, as seen later. Here, it refers to resolving an inverse geometric problem. Indeed, from a set of mappings between detected 2D

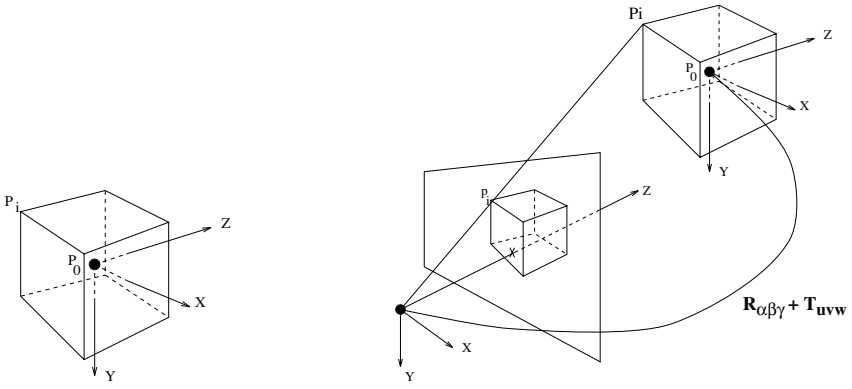


Figure 4.2. *Model defined in its reference mark R_m (left) and solution for the problem in reference mark R_c (right)*

primitives in an image and the corresponding elements of the voluminous model of the observed object, we look for the spatial attitude of the latter to be coherent with its pairs, i.e., the position of the object in the space, which realizes the coincidence between the perspective projections of 3D primitives with their counterparts in the image (see Figure 4.2).

4.3.2. *Mappings*

The initial algorithm proposed by Lowe [LOW 85] works from the pairs between line segments, but we will show that this approach changes easily to cases where the observed object is characterized by a set of interest points; in addition, these two types of pairs can be mixed within the same procedure for attitude research.

4.3.2.1. *Matching of lines*

Let us suppose an edge of the object to be delimited by points P_i^1 and P_i^2 in the reference mark of the model R_m. Let us assume that this edge created in the image, by perspective projection, a segment characterized by any point $p_i = (x_i, y_i, f)$ and a vector $\overrightarrow{v_i} = (a_i, b_i, 0)$. It is possible to define a specific plane, called an "*interpretation plane*", passing through the optical center of the camera and containing this image segment. This plane is characterized by its unitary normal $\overrightarrow{N_i}$, which can be easily estimated from the image elements

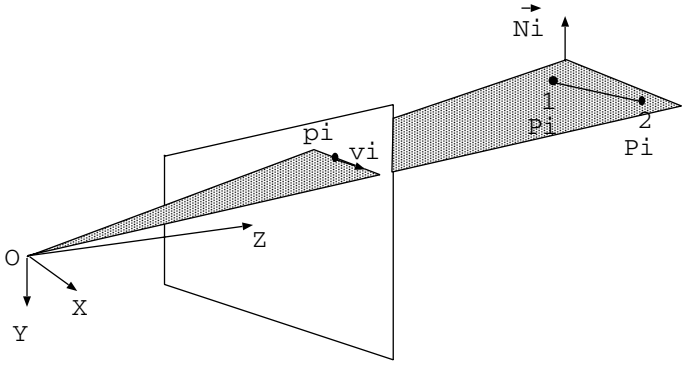


Figure 4.3. *Interpretation plane relative to matching of lines*

mentioned earlier:

$$\overrightarrow{N_i} = \frac{\overrightarrow{op_i} \wedge \overrightarrow{v_i}}{\|\overrightarrow{op_i} \wedge \overrightarrow{v_i}\|} \tag{4.3}$$

Since points P_i^1 and P_i^2 belong to this interpretation plane (see Figure 4.3), their co-ordinates have to verify the following relations:

$$\begin{cases} \overrightarrow{N_i} \cdot P_i^{1\prime} = \overrightarrow{N_i} \cdot \left[\mathbf{R}_{\alpha\beta\gamma} P_i^1 + \mathbf{T}_{uvw}\right] = 0 \\ \overrightarrow{N_i} \cdot P_i^{2\prime} = \overrightarrow{N_i} \cdot \left[\mathbf{R}_{\alpha\beta\gamma} P_i^2 + \mathbf{T}_{uvw}\right] = 0 \end{cases} \tag{4.4}$$

NOTE 4.2. In equations (4.3) and (4.4), the mathematical symbols "$\wedge$" and "." represent the vector product and the scalar product, respectively. Moreover, in order to simplify certain equations, we omit the vectorial symbol in the normals of the interpretation planes.

4.3.2.2. *Pairing of points*

When the object is characterized by interest points, the mapping stage defines the matching between a point P_i whose co-ordinates are expressed in the reference mark of the model R_m and the localization of its perspective projection in the image plane p_i. Hence, it is possible to go back to the previous case by defining for each p_i point, two interpretation planes qualified as "*quasi-horizontal*" and "*quasi-vertical*", of normals N_i^1 and N_i^2

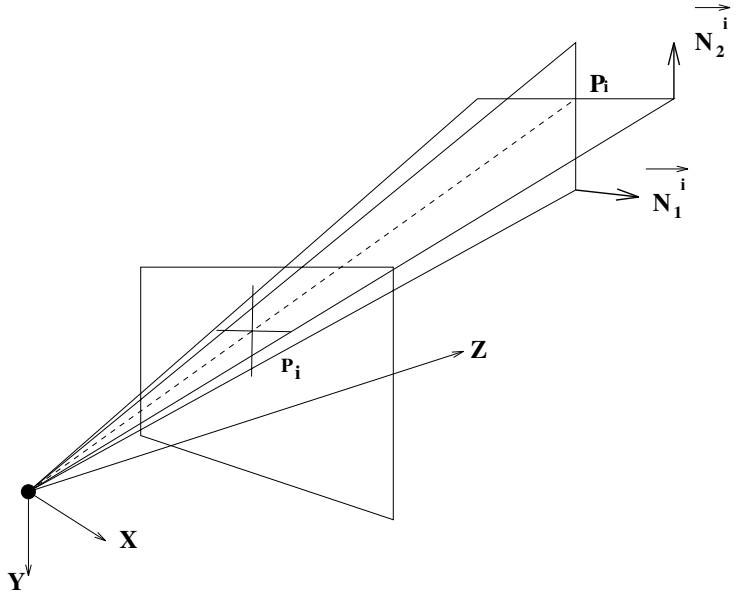


Figure 4.4. *Interpretation planes relative to pairing of points*

respectively (see Figure 4.4).

$$N_i^1 = \frac{\overrightarrow{op_i} \wedge \overrightarrow{Y}}{\left\| \overrightarrow{op_i} \wedge \overrightarrow{Y} \right\|}$$

$$N_i^2 = \frac{\overrightarrow{op_i} \wedge \overrightarrow{X}}{\left\| \overrightarrow{op_i} \wedge \overrightarrow{X} \right\|}$$

where $\overrightarrow{X} = (1, 0, 0)$ and $\overrightarrow{Y} = (0, 1, 0)$ are unitary vectors of axes X and Y, respectively, of the camera reference mark. As a solution, the two relations, which impose the alignment of points p_i and P_i' have to be verified:

$$\begin{cases} N_i^1 \cdot P_i' = N_i^1 \cdot \left[\mathbf{R}_{\alpha\beta\gamma} P_i + \mathbf{T}_{uvw} \right] = 0 \\ N_i^2 \cdot P_i' = N_i^2 \cdot \left[\mathbf{R}_{\alpha\beta\gamma} P_i + \mathbf{T}_{uvw} \right] = 0 \end{cases} \tag{4.5}$$

4.3.3. *Criterion to minimize*

In order to include within the same framework the calculations for the two types of pairing mentioned above (see equation systems (4.4) and (4.5)), we will consider the couples (P_i^j, N_i^j) corresponding to a 3D point of the object and to the normal (or to one) of the associated interpretation plane(s), respectively. With this notation, $N_i^1 = N_i^2$ for a matching between lines and $P_i^1 = P_i^2$ for a pairing between points.

As a solution to our localization problem, n points of the paired model will have to be found in their respective interpretation planes. However, following the detection of primitives, which are not precise in images, and sometimes following imperfections in the model of the observed object, this alignment cannot be perfectly carried out. Hence, it is necessary for us to define a criterion to minimize. With concerns of simplicity, we propose to consider the orthogonal distance of the 3D point to the associated interpretation plane in [DHO 89].

By definition, such a plane passes through the origin of the camera reference mark R_c and is parametrized by a unitary normal; the distance mentioned is equal to a simple scalar product:

$$\mathcal{D}\Big(N_i^j, P_i^{j\prime}\Big) = N_i^j \cdot P_i^{j\prime} = N_i^j \cdot \Big[\mathbf{R}_{\alpha\beta\gamma} P_i^j + \mathbf{T}_{uvw}\Big] \tag{4.6}$$

Let $\mathcal{F}(\alpha, \beta, \gamma, u, v, w, N_i^j, P_i^j)$ be the functional, which for any couple (P_i^j, N_i^j) associates the value defined in (4.6). The objective of the localization method is to then find the vector of parameters $(\alpha, \beta, \gamma, u, v, w)$, which minimize the following criterion:

$$\begin{aligned}\mathcal{E} &= \sum_{i=1}^{n}\sum_{j=1}^{2}\Big[\mathcal{D}\Big(N_i^j, P_i^{j\prime}\Big)\Big]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2}\mathcal{F}\Big(\alpha, \beta, \gamma, u, v, w, N_i^j, P_i^j\Big)^2\end{aligned}$$

This criterion is not linear. Hence, its minimization requires the implementation of an iterative process of the Newton-Raphson type or the Levenberg-Marquard type (see [PRE 92] for theoretical details concerning these techniques) in order to estimate the sought state vector.

4.3.4. *Solving the problem using the Newton-Raphson method*

Let A be the state vector corresponding to a rotation and any translation, and A_k be the state vector relative to the rotation and translation determined at stage k of the iterative process:

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_6 \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \cdot \\ \cdot \\ w \end{pmatrix}$$

$$A_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \cdot \\ \cdot \\ a_{6k} \end{pmatrix} = \begin{pmatrix} \alpha_k \\ \beta_k \\ \cdot \\ \cdot \\ w_k \end{pmatrix}$$

Let us write the development limited to order 1 of the functional $\mathcal{F}$ in the neighborhood of A_k:

$$\begin{aligned} \mathcal{F}\big(A, N_i^j, P_i^j\big) &\simeq \mathcal{F}\big(A_k, N_i^j, P_i^j\big) \\ &\quad + \frac{\partial \mathcal{F}\big(A, N_i^j, P_i^j\big)}{\partial A}\bigg|_{A=A_k} \big(A - A_k\big) \\ &\simeq \mathcal{F}\big(A_k, N_i^j, P_i^j\big) \\ &\quad + \sum_{l=1}^{6} \frac{\partial \mathcal{F}\big(A, N_i^j, P_i^j\big)}{\partial a_l}\bigg|_{A=A_k} \big(a_l - a_{lk}\big) \end{aligned}$$

By assuming that A is the expected solution, then $\mathcal{F}(A, N_i^j, P_i^j) = 0$ and we can write the following approximation:

$$-\mathcal{F}\big(A_k, N_i^j, P_i^j\big) \simeq \frac{\partial \mathcal{F}\big(A, N_i^j, P_i^j\big)}{\partial A}\bigg|_{A=A_k} \big(A - A_k\big)$$

By assuming that $\Delta a_l = a_l - a_{lk}$ for $l \in [1,6]$ and by considering the set of n pairs, we obtain the following linear system:

$$\begin{array}{cccc}
-\mathcal{F}\big(A_k, N_1^1, P_1^1\big) = \dfrac{\partial \mathcal{F}\big(A, N_1^1, P_1^1\big)}{\partial \alpha}\Big|_{A=A_k} \Delta\alpha + \cdots + \dfrac{\partial \mathcal{F}\big(A, N_1^1, P_1^1\big)}{\partial w}\Big|_{A=A_k} \Delta w \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
-\mathcal{F}\big(A_k, N_n^2, P_n^2\big) = \dfrac{\partial \mathcal{F}\big(A, N_n^2, P_n^2\big)}{\partial \alpha}\Big|_{A=A_k} \Delta\alpha + \cdots + \dfrac{\partial \mathcal{F}\big(A, N_n^2, P_n^2\big)}{\partial w}\Big|_{A=A_k} \Delta w
\end{array}$$

The latter having 6 unknown values, it is necessary to process a minimum of 3 pairs providing 2 equations each. It can be noted as matrices:

$$[F] = [J][\Delta A]$$

Its solution in the direction of least squares, for ($n > 3$), is given by:

$$[\Delta A] = \big({}^t[J][J]\big)^{-1}\,{}^t[J][F]$$

This amounts to calculating of the corrective vector ΔA, which minimizes the sum of the distances to the square of the points in the model of the interpretation planes. By knowing the corrective vector ΔA, it is possible to again calculate the state vector A_{k+1} using the following relation:

$$A_{k+1} = A_k + \Delta A$$

If the initial conditions A_0 are selected carefully, around 10 iterations are sufficient to realize the convergence, since the limit test can be based on the norm of the vector ΔA.

4.3.5. *Calculation of partial derivatives*

Obtaining the Jacobian matrix $[J]$ requires the estimation of a number of partial derivatives. We will see how to simplify these calculations, especially those associated with Euler angles. Remember that the functional $\mathcal{F}(A, N_i^j, P_i^j)$ is equal to the following scalar product:

$$\mathcal{F}\big(A, N_i^j, P_i^j\big) = N_i^j \cdot \Big[\mathbf{R}_{\alpha\beta\gamma} P_i^j + \mathbf{T}_{uvw}\Big]$$

where $\mathbf{R}_{\alpha\beta\gamma}$ is equal to the product of the three matrices of elementary rotation around each axis of the reference mark.

$$\mathbf{R}_{\alpha\beta\gamma} = \mathbf{R}_\gamma \mathbf{R}_\beta \mathbf{R}_\alpha$$

$$= \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}$$

If we consider, for example, the partial derivative of the functional $\mathcal{F}(A, N_i^j, P_i^j)$ in comparison to angle β, we have to calculate as follows:

$$\frac{\partial \mathcal{F}\big(A, N_i^j, P_i^j\big)}{\partial \beta} = N_i^j \cdot \frac{\partial \Big[\mathbf{R}_\gamma \mathbf{R}_\beta \mathbf{R}_\alpha P_i^j + \mathbf{T}_{uvw}\Big]}{\partial \beta} = N_i^j \cdot \left[\mathbf{R}_\gamma \frac{\partial \mathbf{R}_\beta}{\partial \beta} \mathbf{R}_\alpha P_i^j\right]$$

In order to simplify these expressions, at each stage of the iterative process, we modify the position of the model in its reference mark R_m. Let A_k be the state vector found at stage k. Then, we calculate the points $P_i^j|_k = \mathbf{R}_{\alpha_k\beta_k\gamma_k} P_i^j + \mathbf{T}_{u_k v_k w_k}$, which become the new characteristic points of the model.

Very simply, at each stage, this amounts to choosing a reference mark relative to the last calculated position in such a way that the calculation of all partial derivatives is done for a vector of zero parameter. The criterion to minimize at stage $k+1$ is thus:

$$\mathcal{E}_{k+1} = \sum_{i=1}^{n} \sum_{j=1}^{2} \left[\mathcal{F}\big(A_k = 0, N_i^j, P_i^j|_k\big) + \sum_{l=1}^{6} \frac{\partial \mathcal{F}\big(A, N_i^j, P_i^j|_k\big)}{\partial a_l}\bigg|_{A=0} \Delta_{a_l}\right]^2$$

Hence, the partial derivatives are trivial. Indeed, noting by $(N_{ix}^j, N_{iy}^j, N_{iz}^j)$ the components of the normal N_i^j, we obtain:

$$\mathbf{R}_\alpha = \mathbf{R}_\beta = \mathbf{R}_\gamma = I_{3\times 3}$$

$$\frac{\partial \mathbf{R}_\alpha}{\partial \alpha} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad \frac{\partial \mathbf{R}_\beta}{\partial \beta} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad \frac{\partial \mathbf{R}_\gamma}{\partial \gamma} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from which:

$$\left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \alpha}\right|_{A=0} = N_{iz}^j Y_i^k - N_{iy}^j Z_i^k \quad \left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial u}\right|_{A=0} = N_{ix}^j$$

$$\left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \beta}\right|_{A=0} = N_{ix}^j Z_i^k - N_{iz}^j X_i^k \quad \left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial v}\right|_{A=0} = N_{iy}^j$$

$$\left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \gamma}\right|_{A=0} = N_{iy}^j X_i^k - N_{ix}^j Y_i^k \quad \left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial w}\right|_{A=0} = N_{iw}^j$$

or even, in a condensed manner:

$$\left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \mathbf{R}_{\alpha\beta\gamma}}\right|_{A=0} = P_i^j|_k \wedge N_i^j \qquad \left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \mathbf{T}_{uvw}}\right|_{A=0} = N_i^j$$

The estimation of the state vector relative to stage $(k+1)$ is obtained by:

$$\begin{cases} \mathbf{R}_{\alpha_{k+1}\beta_{k+1}\gamma_{k+1}} = \left[\mathbf{R}_{\Delta\alpha\Delta\beta\Delta\gamma}\right] \cdot \left[\mathbf{R}_{\alpha_k\beta_k\gamma_k}\right] \\ \mathbf{T}_{u_{k+1}v_{k+1}w_{k+1}} = \left[\mathbf{R}_{\Delta\alpha\Delta\beta\Delta\gamma}\right] \cdot \mathbf{T}_{u_kv_kw_k} + \mathbf{T}_{\Delta u\Delta v\Delta w} \end{cases}$$

4.3.6. *Results*

In this section, we present an experiment realized with the localization algorithm described above, with the objective of giving an indication of its precision. In order to know the correct field, we will work with the synthesis data.

Let a parallelepiped be placed in a given attitude (see Figure 4.5 left). After regularly sampling its visible vertices, we carried out the following process 100 times:

1) for each vertex, the projection in the image plane of sampled points and the approximation in the direction of least squares of the line passing perfectly through the latter. This stage provides the pairs (segments 2D/ vertex 3D) necessary for the algorithm (see Figures 4.5 right and 4.6 left, respectively);

2) then, an estimation of the localization of the parallelogram (Figure 4.6 right).

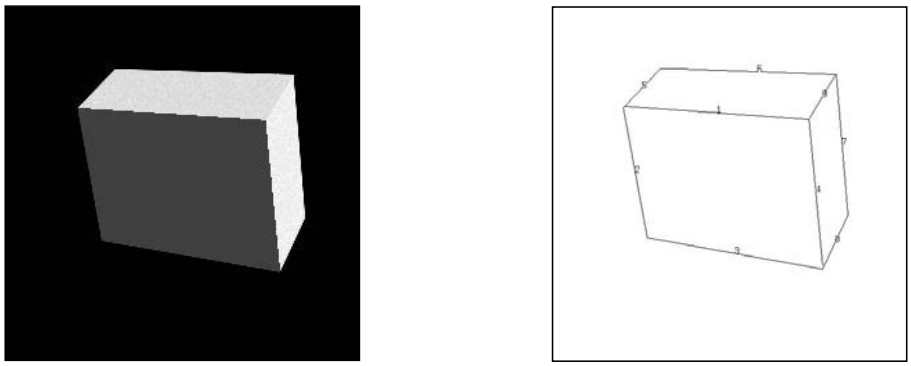


Figure 4.5. *Synthesis image (left) and paired image segments (right)*

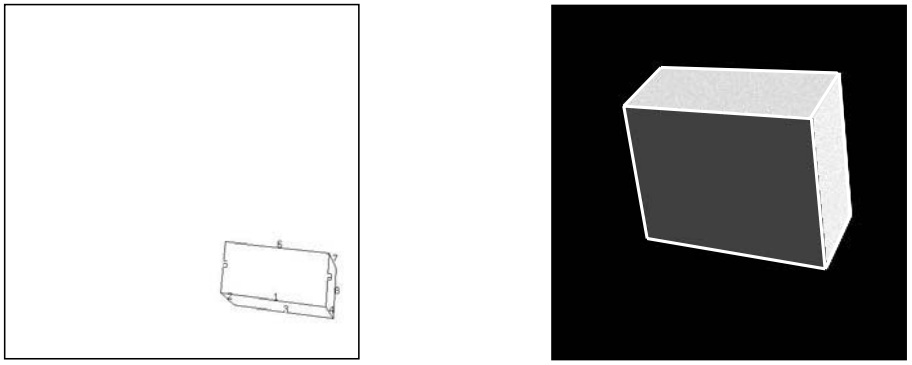


Figure 4.6. *Edges of the paired model in the initial position of the iterative calculation (left) and visualization of the result (right)*

However, before each localization experiment, the intrinsic parameters of the camera as well as the perspective projection of the tested points along the edge are disturbed by an additive Gaussian noise. Repetition of the experiment makes it possible to calculate the statistics of the standards to the true field for different parameters of localization $(\alpha, \beta, \gamma, u, v, w)$. Two perturbation cards are used (see Table 4.1); the first corresponds to the optimal case (finely calibrated camera and careful detection of primitive images). The results obtained are classified in Table 4.2 for the two parameter cards. It is necessary for us to add that, to be complete, the height of the object is 200 mm, that it is situated at 800 mm of the camera and that the focal distance f is taken to be equal to 1,000 pixels.

	Sampled projection points	f	u_0	v_0
card 1	0.1 pixel	10 pixels	1 pixel	1 pixel
card 2	1 pixel	100 pixels	10 pixels	10 pixels

Table 4.1. *Standard deviations for additive Gaussian noises*

	α	β	γ	u	v	w
set 1	0.1 degree	0.1 degree	0.03 degree	0.5 mm	0.5 mm	2.4 mm
set 2	0.9 degree	0.9 degree	0.1 degree	4.9 mm	4.9 mm	28.6 mm

Table 4.2. *Statistical results (standard deviation) on the precision of localization*

4.4. Localization of a voluminous object by multi-ocular vision

In this section, we will show that the previous method is easily generalized while using several cameras (see [BRA 94, BRA 96]), if the relative positioning between the different sensors is known in advance.

4.4.1. *Mathematical developments*

Let us consider the multi-ocular case represented in Figure 4.7. Without loss of generality, we accept the reference mark of the first camera R_{c1} as the reference mark and we will look for the localization of the object in this reference mark.

Let (P_i^j, N_i^j) be a pairing couple between a point of the model of the object and an interpretation plane defined in the reference mark associated with the lth camera R_{cl}. The only remarkable difference as compared to the monocular case is that, when we express these primitives in the reference mark R_{c1}, the interpretation plane associated with N_i^j no longer passes through the origin of the reference mark, which slightly modifies equation (4.6). If R_{l1} and T_{l1} represent the rotation and translation, respectively, which make it possible for the reference mark R_{cl} to pass to reference mark R_{c1}, then the previously defined functional is written as:

$$\mathcal{F}\big(\alpha, \beta, \gamma, u, v, w, N_i^j, P_i^j\big) = \mathbf{R}_{l1} N_i^j \cdot \Big[\mathbf{R}_{\alpha\beta\gamma} P_i^j + \mathbf{T}_{uvw}\Big] - \mathbf{R}_{l1} N_i^j \cdot \mathbf{T}_{l1}$$

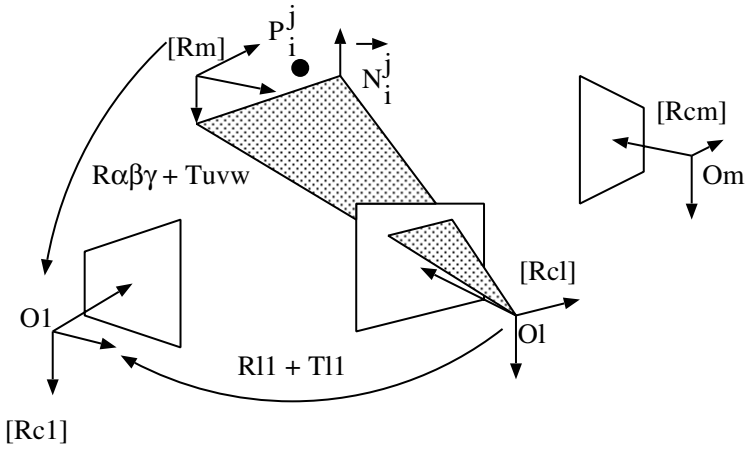


Figure 4.7. *Localization by multi-ocular vision*

where the symbol "·" denotes the scalar product, $\mathbf{R}_{l1}N_i^j$ corresponds to the normal of the interpretation plane expressed in the reference mark R_{c1} and the expression $\mathbf{R}_{l1}N_i^j \cdot \mathbf{T}_{l1}$ is just the orthogonal distance of the interpretation plane to the optical center of the first camera.

4.4.2. *Calculation of partial derivatives*

As for the calculation of partial derivatives, it proves to be only slightly modified:

$$\left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \mathbf{R}_{\alpha\beta\gamma}}\right|_{A=0} = P_i^j|_k \wedge \mathbf{R}_{l1}N_i^j \quad \left.\frac{\partial \mathcal{F}(A, N_i^j, P_i^j)}{\partial \mathbf{T}_{uvw}}\right|_{A=0} = \mathbf{R}_{l1}N_i^j$$

4.4.3. *Results*

Here, we present the results obtained with a trinocular system (see Figure 4.8). Figures 4.9 and 4.10 visualize the 2D/3D pairs that are created. Of course, let us point out the reverse of a method based on 3D reconstruction of primitive images by triangulation that a primitive paired in an image does not necessarily have to be visible and *a fortiori* be paired in other images. The result of the localization is represented in Figure 4.11.

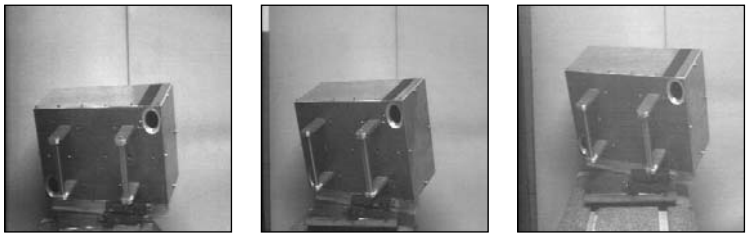


Figure 4.8. *Images obtained by a trinocular acquisition system*

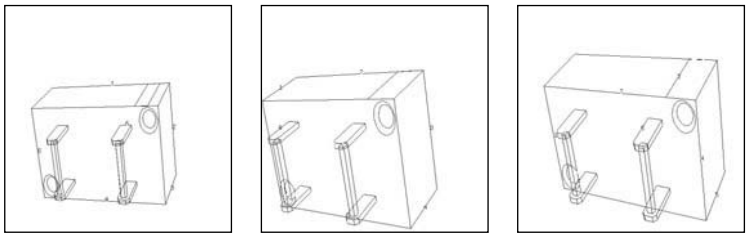


Figure 4.9. *Edges of paired models in each image*

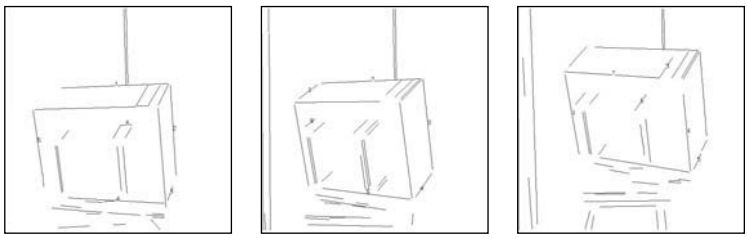


Figure 4.10. *Paired image segments*

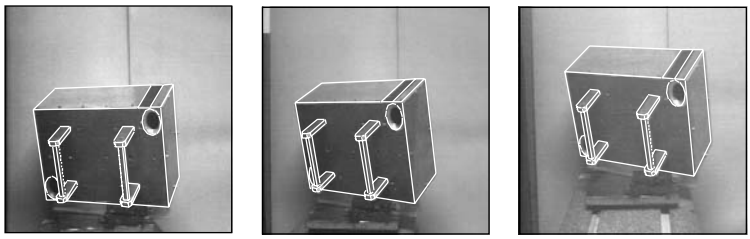


Figure 4.11. *Final localization*

In order to give some indications on the benefits of the multi-camera approach, we again take the experiments described in section 4.3.6 using on the one hand card 2 of the additive noise and, on the other hand, by successively considering 1, 2, or 3 cameras. The results of these experiments are presented in Table 4.3. The relative availability of the cameras is approximately identical to that of the real cases (Figure 4.8). These results illustrate the benefits of several cameras in terms of localization precision, the latter appearing mainly on the translation parameter w along the optical axis.

	α	β	γ	u	v	w
1 camera	0.8 degree	1.1 degree	0.1 degree	4.7 mm	3.2 mm	29.9 mm
2 cameras	0.9 degree	1.0 degree	0.3 degree	5.1 mm	2.8 mm	9.3 mm
3 cameras	0.7 degree	0.9 degree	0.3 degree	4.7 mm	2.6 mm	7.0 mm

Table 4.3. *Statistical results (standard deviation) on the localization precision according to the number of cameras used*

4.5. Localization of an articulated object

Now we address the problem of localization of objects whose configuration is dependant on the internal degrees of freedom (see [LOW 85, DHO 93]).

4.5.1. *Mathematical development*

For example, let us consider an articulated object of "*robotic hand*" type having q degrees of freedom able to correspond to rotoidal or prismatic articulations. The state vector characterizing the position of such an object is noted by $A = (\alpha, \beta, \gamma, u, v, w, \theta_1, \ldots, \theta_q)$. Let P_i^j be a point of the object expressed in the reference mark of definition R_m. Its position in the reference mark camera is obtained by the following calculation:

$$P_i^{j\prime} = \left[\mathbf{T}_{\theta_q}\right]^* \cdots \left[\mathbf{T}_{\theta_1}\right]^* \left[\mathbf{T}_{\alpha\beta\gamma uvw}\right] P_i^j$$

where different matrices $[\mathbf{T}]$ are of size 4×4 and characterize rigid displacements (use of homogenous co-ordinates for points P_i^j).

The matrix $[\mathbf{T}_{\alpha\beta\gamma uvw}]$ corresponds to the extrinsic parameters of positioning of the object in the reference mark camera. The matrices $[\mathbf{T}_{\theta_l}]$ are

relative to different internal parameters. The symbol * indicates that the matrix $[\mathbf{T}_{\theta_l}]$ is taken as equal to the identity matrix if the lth degree of freedom of the object has no influence on the position of point P_i^j. According to the nature of relations, we have two types of equations:

1) Rotoidal relations give the following equation:

$$\mathbf{T}_{\theta_l} = \mathbf{T}_{C_l} \cdot \mathbf{R}_{\theta_l} \cdot -\mathbf{T}_{C_l}$$

with (taking the help of quaternion formalism of parameter $\mathbf{R}_{\theta_l}$):

$$\mathbf{T}_{C_l} = \begin{bmatrix} 1 & 0 & 0 & X_l \\ 0 & 1 & 0 & Y_l \\ 0 & 0 & 1 & Z_l \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{R}_{\theta_l} = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2\big(q_1q_2 - q_0q_3\big) & 2\big(q_1q_3 + q_0q_2\big) & 0 \\ 2\big(q_1q_2 + q_0q_3\big) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2\big(q_2q_3 - q_0q_1\big) & 0 \\ 2\big(q_1q_3 - q_0q_2\big) & 2\big(q_2q_3 + q_0q_1\big) & q_0^2 - q_1^2 - q_2^2 + q_3^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where $C_l = (X_l, Y_l, Z_l)$ is a point on the axis of rotation, $q_0 = \cos(\theta_l/2)$, $q_1 = U_l \cdot \sin(\theta_l/2)$, $q_2 = V_l \cdot \sin(\theta_l/2)$, and $q_3 = W_l \cdot \sin(\theta_l/2)$ with θ_l as the angle of rotation and $\Delta_l = (U_l, V_l, W_l)$ as the unitary vector of the axis of rotation.

2) The prismatic relations give the following formula where θ_l represents the amplitude of the translation and $\Delta_l = (U_l, V_l, W_l)$ corresponds to the unitary vector of the translation direction:

$$\mathbf{T}_{\theta_l} = \begin{bmatrix} 1 & 0 & 0 & \theta_l \cdot U_l \\ 0 & 1 & 0 & \theta_l \cdot V_l \\ 0 & 0 & 1 & \theta_l \cdot W_l \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

4.5.2. *Calculation of partial derivatives for intrinsic parameters*

Elementary calculation of derivatives gives us the following matrices, for a rotoidal relation and a prismatic relation, respectively:

$$\left.\frac{\partial T_{\theta_l}}{\partial \theta_l}\right|_{\theta_l=0} = \begin{bmatrix} 0 & -W_l & V_l & 0 \\ W_l & 0 & -U_l & 0 \\ -V_l & U_l & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \left.\frac{\partial T_{\theta_l}}{\partial \theta_l}\right|_{\theta_l=0} = \begin{bmatrix} 0 & 0 & 0 & U_l \\ 0 & 0 & 0 & V_l \\ 0 & 0 & 0 & W_l \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

from which it is easy to show that the corresponding elements of the Jacobian matrix are equal to a mixed product and a scalar product:

$$\left.\frac{\partial \mathcal{F}\left(A, N_i^j, P_i^j\right)}{\partial \theta_l}\right|_{\theta_l=0} = N_i^j \cdot \left[\Delta_l \wedge P_i^j|_k\right] \quad \left.\frac{\partial \mathcal{F}\left(A, N_i^j, P_i^j\right)}{\partial \theta_l}\right|_{\theta_l=0} = N_i^j \cdot \Delta_l$$

4.5.3. *Results*

The following images are relative to the implementation of this algorithm in order to calculate the attitude of a robotic arm. Figure 4.12 represents the pairs realized between the CAD model of the object and the segments detected in the image. The various attitudes successively obtained during the convergence of the process are presented in Figure 4.13 (left). The right-hand side of the same figure illustrates the final estimated exposure.

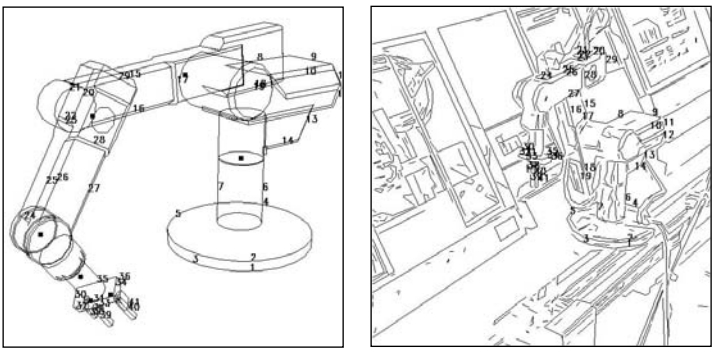


Figure 4.12. *Pairs realized: (left) model, (right) image*

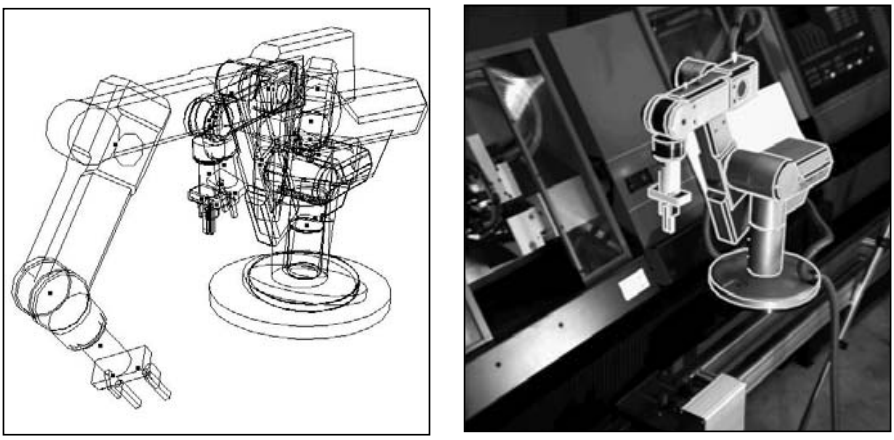


Figure 4.13. *Results: (left) convergence of the algorithm, (right) superposition of the video image and of the model in the expected attitude*

4.6. Hand-eye calibration

4.6.1. *Introduction*

In this section, we will address the hand-eye calibration problem. This amounts to, while using a video camera mounted on an articulated system of active turret or hand manipulator type, determining the relative position of the sensor compared to the terminal organ. This makes it possible to express the information extracted from the image (for example, localization of an object) in the reference mark of the articulated system reference and hence simplify its order.

4.6.2. *Presentation of the method*

Let us consider a robotic hand equipped with a mounted camera (see Figure 4.14). The analysis of the problem evoked leads us to define the following reference marks:

– R_r reference mark of articulated arm reference;

– R_e reference mark connected to the effector;

– R_c camera reference mark;

– R_m reference mark of the definition of the model of the observed object

whose relative positions are quantified by rigid transformations:

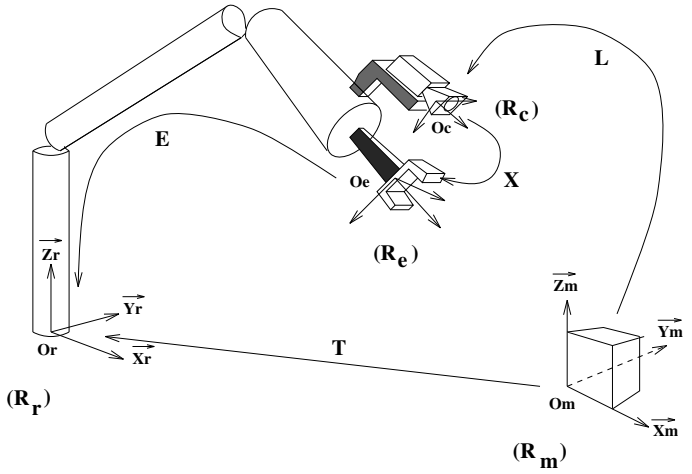


Figure 4.14. *Hand-eye calibration*

– $[\mathbf{L}]$ localization of the observed object in the camera reference mark (can be estimated with the help of the monocular localization technique explained earlier);

– $[\mathbf{X}]$ hand-eye transition matrix (which we are studying);

– $[\mathbf{E}]$ positioning of the effector reference mark in the absolute reference mark R_r (matrix provided by the geometric model of the robot and the reading of the top-encoders);

– $[\mathbf{L}]$ localization of the observed object in the reference mark R_r.

Almost all hand-eye calibration methods proposed in other works (see [SHI 89, TSA 89, WAN 92, HOR 93]) amount to resolving the matrix equation system of the form:

$$\mathbf{A}_{lm}\mathbf{X} = \mathbf{X}\mathbf{B}_{lm} \tag{4.7}$$

The experimental method is as follows. The robot moves in its workplace and occupies a series of n successive positions by acquiring each time an image of a reference test chart. For each image couple of indices l and m, respectively, it proves possible to form an equation of type (4.7). The transformation $\mathbf{B}_{lm}$ is obtained by composing the positions of the effector (matrices $[\mathbf{E}_l]$ and $[\mathbf{E}_m]$) to the two filming moments. As for transformation

$\mathbf{A}_{lm}$, it begins from the two localizations of the test chart in the referential of the camera (matrices $[\mathbf{L}_l]$ and $[\mathbf{L}_m]$). Considering the set of image pairs, we obtain an equation system of type (4.7), from which it is possible to extract the elements of the matrix $\mathbf{X}$ using different techniques (see previous references). However, let us note that solving the problem arising from this method requires the total estimation of $6 * n + 6$ parameters.

As a result we propose a new method [REM 98] working from identical data (n images of a reference test chart). However, we note that the position of the test chart sample in the camera reference mark can be expressed, for the lth image, by the composition of the following matrix: $[\mathbf{X}]^{-1}[\mathbf{E}_l]^{-1}[\mathbf{T}]$. Whatever the number of images acquired, the problem thus formulated possesses only 12 parameters, which correspond to the degrees of freedom of rigid transformations $[\mathbf{X}]$ and $[\mathbf{T}]$.

4.6.3. *Geometric constraint*

To estimate these 12 unknown values, we will, as in the earlier methods, find with the help of an iterative method how to minimize the distances to the interpretation planes. Let (P_i^j, N_i^j) be a pair couple; P_i^j being a point in the test chart and N_i^j defining an interpretation plane in the camera reference mark. The usual functional $\mathcal{F}$ is written by expressing the paired entities in the absolute reference mark of the robot R_r:

$$\mathcal{F}\big(\mathbf{X}, \mathbf{T}, \mathbf{E}_l, N_i^j, P_i^j\big) = \mathbf{R}_{E_l}\mathbf{R}_X N_i^j \cdot \Big[\Big[\mathbf{R}_T P_i^j + \mathbf{T}_T\Big] - \big[\mathbf{R}_{E_l}\mathbf{T}_X + \mathbf{T}_{E_l}\big]\Big]$$

where $\mathbf{R}_{E_l}\mathbf{R}_X N_i^j$ and $\mathbf{R}_{E_l}\mathbf{R}_X N_i^j \cdot [\mathbf{R}_{E_l}\mathbf{T}_X + \mathbf{T}_{E_l}]$ represent the normal of the interpretation plane and its distance from the origin expressed in the reference mark R_r, respectively.

4.6.4. *Results*

A comparative study [REM 98] conducted from synthetic data has shown that the recommended method assures a better resistance to noise.

On the real experimental plane, we carried out hand-eye calibration of a Cartesian robot at six degrees of freedom. For this experiment, we used a camera equipped with a large object angle of focal length 3.8 mm.

Without knowing the real values of the parameters of rigid transformations $\mathbf{X}$ and $\mathbf{T}$, we cannot quantify, *a priori*, the precision of the results but we have

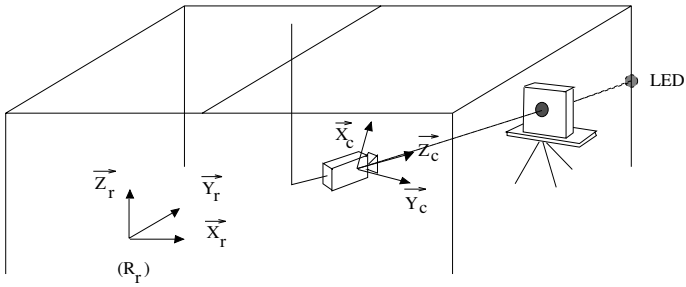


Figure 4.15. *Experimental device to validate the precision of the estimation*

imagined an experimental device making it possible to highlight the quality of the latter. The camera, fixed tightly on the hand manipulator, is initially placed approximately 50 cm from a plate pierced with a hole of 1.0 cm diameter. Then we position an "LED" such that we obtain the alignment of the latter with the optical center of the camera and the inner bore of the plate (see Figure 4.15).

Knowledge of the hand-eye transformation makes it possible for us to directly order the displacement of the camera in the space of the robot. If the estimation of X is precise, it is possible for us to apply a rotation to the camera reference mark without changing its origin; this transformation must preserve the achieved alignment. Since the camera is equipped with a short focal length objective, it is possible to apply rotations of quite strong amplitude (of order 80^o) while preserving the plate in the vision field of the camera.

This experiment is illustrated in Figure 4.16, which represents the images delivered by the camera mounted for several configurations of the robotic set, leaving the optical center of the camera unchanged. Figure 4.17 represents the partial enlargement of previous images showing good preservation of the initial alignment.

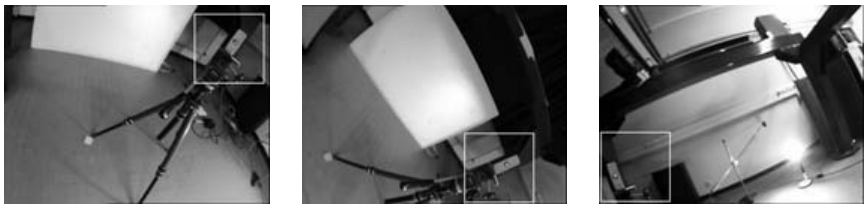


Figure 4.16. *Images delivered by a mounted camera*

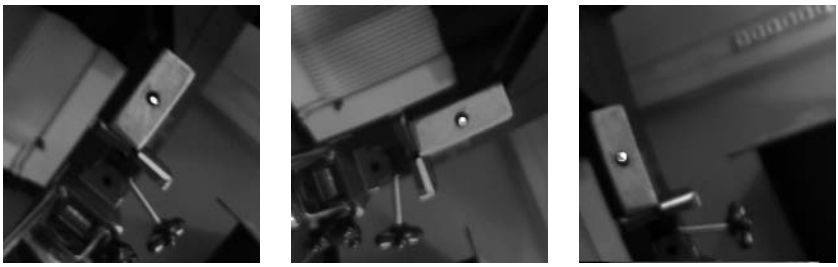


Figure 4.17. *Partial enlarging of images in Figure 4.16 validating the conservation of the initial alignment*

4.7. Initialization methods

All localization methods presented earlier use non-linear minimization algorithms and hence require an initialization of the value of the expected position, in the convergence field of the algorithm.

The method presented here has the advantage of being able to pass through the initialization conditions and of being very efficient, but has the disadvantage of generally providing results with slightly inferior accuracy. Thus, it becomes a perfect candidate for the initialization of earlier algorithms. Moreover, the quality of initialization provided makes it possible to strongly limit the number of iterations and thus their cost. The idea put forth by Dementhon is to try to go back steadily to the perspective projection model, starting from the hypothesis of the scaled orthographic projection. Indeed, these two projection models are differentiated by a factor $(1 + \epsilon)$ where ϵ represents the depth of the object as compared to its distance from the camera.

The advantage of this method is that it reduces the number of calculations to be carried out. It specifically helps to avoid a systematic inversion of the matrix (at each iteration of the method), since we work on a matrix that is pseudo-inversed and predetermined and depends only on three-dimensional points chosen in the model. Owing to this aspect, the Dementhon method proves to be of great interest for a process working in real-time.

4.7.1. *Initial hypotheses*

– Knowledge of the model of the observed object, defined in the reference mark R_m.

– Acquisition of an image of the object.

– Mapping between the detected points in the image and the characteristic points of the object.

4.7.2. *Objective*

To correlate the camera reference mark (R_c) and the model reference mark (R_m) and to find the rotation and translation to be applied to the model in order to place it in space, in an attitude conforming to the image.

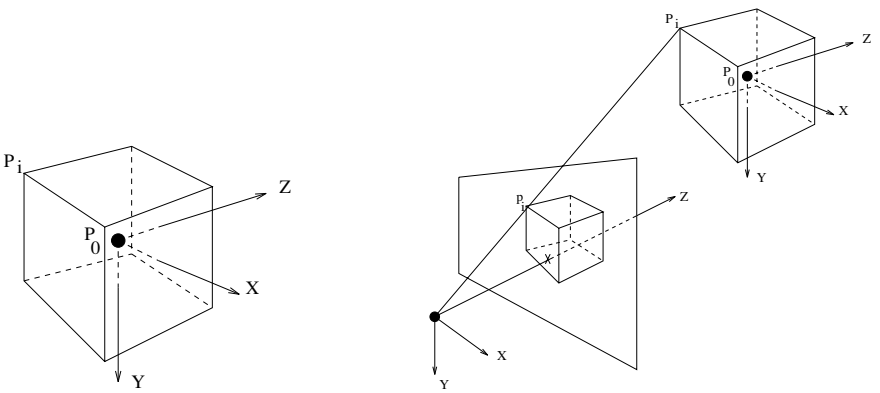


Figure 4.18. *a) Reference mark of the model (R_m); b) reference mark of the camera (R_c)*

Points p_i $i \in [1, n]$ are images of points P_i by perspective projection. We will note by (X_i, Y_i, Z_i) the co-ordinates of P_i in the reference mark (R_m) and by (X'_i, Y'_i, Z'_i) the co-ordinates of P_i after the sought rotation and translation.

As homogenous co-ordinates, we have:

$$\begin{pmatrix} X'_i \\ Y'_i \\ Z'_i \\ 1 \end{pmatrix} = \begin{pmatrix} R_{(3\times3)} & T_{(3\times1)} \\ 0_{(1\times3)} & 1 \end{pmatrix} \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} i_x & i_y & i_z & u \\ j_x & j_y & j_z & v \\ k_x & k_y & k_z & w \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix}$$

or even:

$$X_i' = \left(i_x, i_y, i_z, u\right) \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} = \left(I_{(1\times 3)}, u\right) \cdot \begin{pmatrix} P_{i(3\times 1)} \\ 1 \end{pmatrix}$$

$$Y_i' = \left(j_x, j_y, j_z, v\right) \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} = \left(J_{(1\times 3)}, v\right) \cdot \begin{pmatrix} P_{i(3\times 1)} \\ 1 \end{pmatrix}$$

$$Z_i' = \left(k_x, k_y, k_z, w\right) \cdot \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} = \left(K_{(1\times 3)}, w\right) \cdot \begin{pmatrix} P_{i(3\times 1)} \\ 1 \end{pmatrix}$$

I, J, K are director vectors of the camera reference mark (R_c) expressed in the reference mark of the model (R_m) in the position corresponding to the solution of the problem.

4.7.3. *Under the hypothesis of perspective projection*

With $p_i = (x_i, y_i, z_i(= f))$ correlated with $P_i = (X_i, Y_i, Z_i)$, we obtain:

$$x_i = \frac{X_i' f}{Z_i'} = \frac{f(I,u) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}}{(K,w) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}} \tag{4.8}$$

$$\Longrightarrow x_i (K,w) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} = f(I,u) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}$$

$$y_i = \frac{Y_i' f}{Z_i'} = \frac{f(J,v) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}}{(K,w) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}} \tag{4.9}$$

$$\Longrightarrow y_i (K,w) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} = f(J,v) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix}$$

$$x_i\left(\frac{K \cdot P_i}{w} + 1\right) = \frac{f}{w}(I, u) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} \tag{4.10}$$

$$y_i\left(\frac{K \cdot P_i}{w} + 1\right) = \frac{f}{w}(J, v) \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} \tag{4.11}$$

By supposing that:

$$\varepsilon_i = \frac{K \cdot P_i}{w} = \frac{k_x X_i + k_y Y_i + k_z Z_i}{w}$$

$$I^* = \frac{f}{w}(I, u) = \frac{f}{w}\big(i_x, i_y, i_z, u\big) = \big(I_1, I_2, I_3, I_4\big)$$

$$J^* = \frac{f}{w}(J, v) = \frac{f}{w}\big(j_x, j_y, j_z, v\big) = \big(J_1, J_2, J_3, J_4\big)$$

by considering all the correspondences, we obtain the following two linear systems:

$$x_i\big(1 + \varepsilon_i\big) = I^* \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} \Longrightarrow \big(P_i, 1\big) \cdot I^* = x_i\big(1 + \varepsilon_i\big) \tag{4.12}$$

$$i \in [1, n]$$

$$y_i\big(1 + \varepsilon_i\big) = J^* \cdot \begin{pmatrix} P_i \\ 1 \end{pmatrix} \Longrightarrow \big(P_i, 1\big) \cdot J^* = y_i\big(1 + \varepsilon_i\big) \tag{4.13}$$

These can be re-written as:

$$\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_n & Y_n & Z_n & 1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} = \begin{bmatrix} x_1\big(1 + \varepsilon_1\big) \\ \cdot \\ \cdot \\ x_n\big(1 + \varepsilon_n\big) \end{bmatrix} \tag{4.14}$$

$$\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_n & Y_n & Z_n & 1 \end{bmatrix} \begin{bmatrix} J_1 \\ J_2 \\ J_3 \\ J_4 \end{bmatrix} = \begin{bmatrix} y_1\big(1 + \varepsilon_1\big) \\ \cdot \\ \cdot \\ y_n\big(1 + \varepsilon_n\big) \end{bmatrix} \tag{4.15}$$

Only vectors I^*, J^* and coefficients ε_i are unknown in this system.

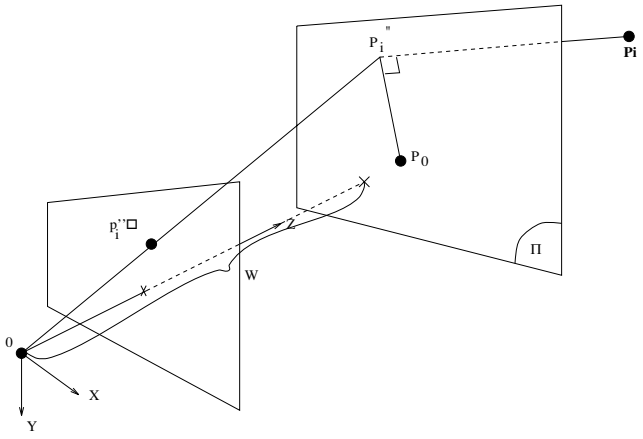


Figure 4.19. *Scaled orthographic projection*

4.7.4. *Under the hypothesis of scaled orthographic projection*

We consider plane Π passing through the origin P_0 of the reference mark of the model (R_m) and parallel to the image plane.

In scaled orthographic projection, we project the object in this plane and then we carry out a perspective projection on the image plane (changing the scale according to the distance).

The sought solution is that the distance between plane Π and the origin of the camera reference mark is equal to the component w of the translation.

Let P_i be a point of the model and P_i'' its orthographic projection on plane Π. In the camera reference mark, P_i'' has (X_i', Y_i', w) as co-ordinates and $p_i''(x_i'', y_i'')$ as image with:

$$\begin{gathered} x_i'' = \frac{X_i' f}{w} \quad \text{and} \quad y_i'' = \frac{Y_i' f}{w} \\ x_i'' = \frac{f}{w}(I, u)\begin{pmatrix} P_i \\ 1 \end{pmatrix} \quad \text{and} \quad y_i'' = \frac{f}{w}(J, v)\begin{pmatrix} P_i \\ 1 \end{pmatrix} \quad i \in [1, n] \end{gathered} \tag{4.16}$$

It appears that the term neglected between the two projection models is just $x_i\varepsilon_i$ or $y_i\varepsilon_i$ with $\varepsilon_i = (K \cdot P_i)/w$.

$K \cdot P_i$ is equal to the distance, according to the optical axis, between point P_i and the origin P_0 of (R_m). If this origin is judiciously selected (bar center of the object), neglecting ε_i amounts to neglecting the depth of the object (according to the optical axis) as compared to its distance from the optical center. This hypothesis is often realistic. The fact that ε_i is multiplied by x_i or y_i shows us that we will always have an interest to keep the object as centered as possible in the image, in order to minimize the error introduced by the scaled orthographic projection.

The originality of the Dementhon method remains in a progressive shift from the scaled orthographic projection model, to the perspective projection model.

4.7.5. *Development of the algorithm*

1) $t=0$ and $\varepsilon_i(t)=\varepsilon_i(0)=0$ for $i\in[1,n]$ (scaled orthographic projection).

2) Calculation of vectors I^* and J^* with the help of 2 linear systems (4.15, 4.16) (resolution in the direction of least squares):

$$A = \begin{bmatrix} X_1 & Y_1 & Z_1 & 1 \\ . & . & . & . \\ . & . & . & . \\ X_n & Y_n & Z_n & 1 \end{bmatrix},$$

$$S_x = \begin{bmatrix} x_1(1+\varepsilon_1(t)) \\ . \\ . \\ . \\ x_n(1+\varepsilon_n(t)) \end{bmatrix}, \tag{4.17}$$

$$S_y = \begin{bmatrix} cy_1(1+\varepsilon_1(t)) \\ . \\ . \\ . \\ y_n(1+\varepsilon_n(t)) \end{bmatrix},$$

$$AI^* = S_x \qquad AJ^* = S_y \tag{4.18}$$

If $n > 4$, the systems are over-determined and we use a minimization in the direction of least squares:

$$\begin{aligned} {}^tAAI^* = {}^tAS_x &\Longrightarrow I^* = ({}^tAA)^{-1}\,{}^tAS_x = A^+S_x \\ {}^tAAJ^* = {}^tAS_y &\Longrightarrow J^* = ({}^tAA)^{-1}\,{}^tAS_y = A^+S_y \end{aligned} \tag{4.19}$$

A^+ is called the *pseudo-inverse* matrix of matrix A. This matrix can be pre-calculated as it depends only on points P_i of the model paired to the image points:

$$N_I = \sqrt{I_1^2 + I_2^2 + I_3^2} \quad N_J = \sqrt{J_1^2 + J_2^2 + J_3^2}$$

3) $I = (I_1, I_2, I_3)/(N_I)\ J = (J_1, J_2, J_3)/(N_J) K = I \wedge J\ J = K \wedge I$

$$w = \frac{f}{N_I} = \frac{f}{N_J} \quad u = I_4 \cdot \frac{w}{f} \quad v = J_4 \cdot \frac{w}{f}$$

4) $t = t + 1$, calculation of new $\varepsilon_i(t) = \frac{K \cdot P_i}{w}$

If $|(\varepsilon_i(t) - \varepsilon_i(t-1))/n| \ <$ threshold, then I, J, K, u, v, w define the position sought, otherwise return to stage 2. Generally, the convergence needs 4 to 5 iterations without requiring the study of specific initial conditions.

4.7.6. *Specific case of a planar object*

In the case of perfectly flat objects, it is necessary to bring a modification to the Dementhon algorithm. Indeed, the matrix ${}^{\mathbf{t}}\mathbf{AA}$ is only in row 2. If we arrange it to express the co-ordinates of the model in such a way that all values of z are zero, we can only calculate the components I_1, I_2, I_4 and J_1, J_2, J_4 of vectors $\mathbf{I}^*$ and $\mathbf{J}^*$, respectively.

Let us recall that:

$$I^* = \big(I_1, I_2, I_3, I_4\big) = \frac{f}{w}\big(i_x, i_y, i_z, u\big)$$

$$J^* = \big(J_1, J_2, J_3, J_4\big) = \frac{f}{w}\big(j_x, j_y, j_z, v\big).$$

The vectors (i_x, i_y, i_z) and (j_x, j_y, j_z) being unitary, we can write relation (4.21) as:

$$\left(\frac{f}{w} i_x\right)^2 + \left(\frac{f}{w} i_y\right)^2 + \left(\frac{f}{w} i_z\right)^2 = \left(\frac{f}{w} j_x\right)^2 + \left(\frac{f}{w} j_y\right)^2 + \left(\frac{f}{w} j_z\right)^2 \tag{4.20}$$

By including as unknowns:

$$\begin{cases} x = \dfrac{f}{w} i_z \\ y = \dfrac{f}{w} j_z \end{cases} \tag{4.21}$$

we obtain the first expression:

$$x^2 - y^2 = \left(\frac{f}{w} j_x\right)^2 + \left(\frac{f}{w} j_y\right)^2 - \left(\frac{f}{w} i_x\right)^2 - \left(\frac{f}{w} i_y\right)^2 \tag{4.22}$$

Moreover, the two vectors, (i_x, i_y, i_z) and (j_x, j_y, j_z), are orthogonal and form a base with (k_x, k_y, k_z), from where relation (4.24) develops from the zero scalar product:

$$\left(\frac{f}{w} i_x\right)\left(\frac{f}{w} j_x\right) + \left(\frac{f}{w} i_y\right)\left(\frac{f}{w} j_y\right) + xy = 0 \tag{4.23}$$

This leads us to a system of two equations with two unknowns:

$$\begin{cases} x^2 - y^2 = A \\ xy = B \end{cases} \tag{4.24}$$

with:

$$\begin{cases} A = \left(\dfrac{f}{w} j_x\right)^2 + \left(\dfrac{f}{w} j_y\right)^2 - \left(\dfrac{f}{w} i_x\right)^2 - \left(\dfrac{f}{w} i_y\right)^2 \\ B = -\left[\left(\dfrac{f}{w} i_x\right)\left(\dfrac{f}{w} j_x\right) + \left(\dfrac{f}{w} i_y\right)\left(\dfrac{f}{w} j_y\right)\right] \end{cases} \tag{4.25}$$

from where one equation of the second degree in x^2 (4.27) has as solutions (4.28):

$$x^4 - Ax^2 - B^2 = 0 \tag{4.26}$$

$$x^2 = \frac{A \pm \sqrt{A^2 + 4B^2}}{2} \tag{4.27}$$

However, since $\sqrt{A^2 + 4B^2}$ is bigger than A, their difference is always negative and their sum is positive. Thus, there is only one positive solution in

x^2, which gives us two actual solutions for x:

$$x = \pm\sqrt{\frac{A + \sqrt{A^2 + 4B^2}}{2}} \tag{4.28}$$

Given relation (4.25) connecting y and x, and for a non-zero value of x:

$$y = \frac{B}{x} \tag{4.29}$$

For specific cases where $x = 0$, the value of y is extracted from expression (4.25).

$$y = \pm\sqrt{-A} \tag{4.30}$$

The Dementhon algorithm applied to specific cases of planar objects leads to two solution couples $(\mathbf{I_i^*}, \mathbf{J_i^*})$ for each iteration of the system and thus to two distinct attitudes of the object in space.

Within the framework of this study, we have tested two methods to select one of the two attitudes for each iteration. These two approaches differ from the criterion used to separate the two solutions. The first measures the remainders of the two linear systems, which are processed. The second calculates a criterion linked to the re-projection of 3D points in the image plane. These two approaches will be called *system-criterion and image-criterion*, respectively, for the rest of this chapter. The corresponding criteria are:

$$\mathrm{Arg}\min_{k\in[1,2]} \left(\left\|AI_k^* - S_x\right\| + \left\|AJ_k^* - S_y\right\|\right) \tag{4.31}$$

$$\mathrm{Arg}\min_{k\in[1,2]} \sum_{i=1}^{i=n} \sqrt{\left(x_i - \frac{fX_i^{k'}}{Z_i^{k'}}\right)^2 + \left(y_i - \frac{fY_i^{k'}}{Z_i^{k'}}\right)^2} \tag{4.32}$$

The experiments carried out provided identical results for the two criteria.

It is necessary to note a double modification of the Horaud and Christy method [HOR 95a] in the direction of:

– usage of the segment instead of points;

– replacement of orthographic projection by para-perspective.

4.8. Analytical calculations of localization errors

This section is devoted to the calculation of localization errors related to the utilization of the earlier methods. It is clear that the possibility of finding at the same time the physical magnitude and the estimation of uncertainties, which are attached to them, is crucial for the determination of the adequacy of the methods for the envisaged works.

The calculation chain must start from the data, *a priori*, on the uncertainties of localization of primitive images, to then end at the effective uncertainties on the localization parameters.

One of the major steps consists of obtaining the covariance matrices related to the normal parameters of the interpretation planes, which are used in our alignment equations.

4.8.1. *Uncertainties in the estimation of a line equation*

We start from the pixel data $(p_i = (x_i, y_i)_{i=1...n})$ provided by a curve detector. Following Ayache [AYA 89] we will assume that each curve point is affected by a Gaussian noise.

The Hessian equation for the support line of the points is:

$$-x \sin\theta + y \cos\theta + c = 0,$$

where θ represents the angle of the line with the abscissa axis, and c represents the distance marked from the line to the origin point.

Here, the Gaussian noise is not isotropic. Its orientation is orthogonal to that of the line support. This choice amounts to saying that in the direction of the line, the position of the point is perfectly known; this is obviously false but this knowledge does not help in the determination of the line. The advantage here is to reduce by a factor of two the internal estimation parameters (the "harmful" parameters).

In a reference mark formed from a unitary director vector of the line and of its normal, each point will have a covariance matrix C for which all coefficients will be zero, with the exception of $C_{22} = \sigma_p^2$.

If R is the rotation matrix, which transforms the natural reference mark of the image in that connected to the line, the covariance matrix of each point in the natural reference mark of the image will be: ${}^tRCR$.

The determination of a line equation from the points disturbed by a Gaussian noise is a traditional problem of linear least squares. Often, this problem (and the determination of the related covariance matrix) is treated by writing the line equation in the form $y = ax + b$ (or $x = ay + b$) and by minimizing the sum of type $\sum_{i=1}^{n}(y_i - ax_i - b)^2$.

Since this method can lead to a serious bias, we use the traditional result even though it is perhaps lesser known:

THEOREM 4.1 (Hessian minimization). *Consider the criterion:*

$$S = \sum_{i=1}^{n} \big(- x_i \sin\theta + y_i \cos\theta + c\big)^2.$$

If $\bar{x}$ and $\bar{y}$ are the mean values of (x_i), ((y_i), respectively) and $\mathrm{Var}(x)$ and $\mathrm{Var}(y)$ are their respective variances, and $\mathrm{Cov}(x, y)$ their covariance, then:

$$\tan(2\theta) = 2\,\mathrm{Cov}(x, y)/\big(\,\mathrm{Var}(x) - \mathrm{Var}(y)\big), \text{ and } c = \bar{x}\sin\theta - \bar{y}\cos\theta.$$

Two values of θ differing from $\pi/2$ seem to be the result of the calculation, but it is only illusory since $\mathrm{Cov}(x, y)$ and $(\mathrm{Var}(x) - \mathrm{Var}(y))$ have the signs of the sine and of the cosine respectively.

This result can be found in a number of works on probability and also in [PAV 77].

At present, the covariance matrix of the triplet (a, b, c) remains to be calculated, where $a = -\sin\theta$ and $b = \cos\theta$. With the help of Maple software, we obtain the following result:

Let $N = \sigma_p^2(b^2\,\mathrm{Var}(x) - 2ab\,\mathrm{Cov}(x, y) + a^2\,\mathrm{Var}(y))$ and $D = n(\mathrm{Var}(x) - \mathrm{Var}(y))^2 + 4\,\mathrm{Cov}^2(x, y)$.

The matrix is then:

$$\mathrm{Cov} = \begin{pmatrix} \dfrac{b^2 N}{D} & -\dfrac{abN}{D} & \dfrac{b(\bar{y}a - \bar{x}b)N}{D} \\ -\dfrac{abN}{D} & \dfrac{a^2 N}{D} & -\dfrac{a(\bar{y}a - \bar{x}b)N}{D} \\ \dfrac{b(\bar{y}a - \bar{x}b)N}{D} & -\dfrac{a(\bar{y}a - \bar{x}b)N}{D} & \dfrac{(\bar{y}a - \bar{x}b)^2 N}{D} + \dfrac{\sigma_p^2}{n} \end{pmatrix}.$$

It is to be noted that if we use primitive points and non-segments, the above results are completely applicable. Of course, it is necessary to consider the two interpretation planes (quasi-horizontal and quasi-vertical); the error relating to each of the two lines defining them is given by the horizontal and the vertical standard deviation of the considered point.

4.8.2. *Errors in normals*

For each segment, the uncertainty on the normal to the interpretation plane can be at present calculated using standard techniques.

We will perform two calculations, by specially detailing the first one:

– calculation of covariance for a normal;

– calculation of covariance for n normals.

The second calculation is very similar to the first one even though it is more complicated. It is not possible to restrict ourselves to the first one because of the common dependence of all equations towards the intrinsic parameters of the camera. Whatever it may be, the description of the first calculation will make it possible to perceive the method more easily.

Error in a normal. The line equation supporting the segment is: $ax+by+c=0$ with $a^2+b^2=1$. A unitary normal can then be written as:

$$\begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \begin{pmatrix} \dfrac{af_x}{D} \\ \dfrac{bf_y}{D} \\ \dfrac{c+au_0+bv_0}{D} \end{pmatrix}, \text{ where } D=\sqrt{a^2f_x^2+b^2f_y^2+\left(c+au_0+bv_0\right)^2}.$$

where we noted that $f_x = f$, $f_y = f/k$.

We will assume that we know the covariance matrix of the coefficients (a, b, c) and also:

– the position of the projection of the optical center (u_0, v_0). Here we have the hypothesis that these two variables are independent of the others and are also mutually independent; as a consequence, the corresponding matrix is diagonal with diagonal elements: $\sigma_{u_0}^2$ and $\sigma_{v_0}^2$;

– the focal parameters of the camera f_x and f_y. Here, we put forward the hypothesis that these two variables are independent of the others and are also mutually independent; as a consequence, the corresponding matrix is diagonal with diagonal elements: $\sigma^2_{f_x}$ and $\sigma^2_{f_y}$.

The global covariance matrix on the normal is then:

$$C(n_x, n_y, n_z) = J \begin{pmatrix} \sigma_a^2 & \mathrm{Cov}(a,b) & \mathrm{Cov}(a,c) & 0 & 0 & 0 & 0 \\ \mathrm{Cov}(a,b) & \sigma_b^2 & \mathrm{Cov}(b,c) & 0 & 0 & 0 & 0 \\ \mathrm{Cov}(a,c) & \mathrm{Cov}(b,c) & \sigma_c^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{u_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{v_0} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{f_x} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{f_y} \end{pmatrix} {}^tJ,$$

where J is the Jacobian matrix of the transformation connecting ${}^t(n_x, n_y, n_z)$ (the unitary normal) to the parameters $(a, b, c, u_0, v_0, f_x, f_y)$.

Errors in the calculations of n normals. Let C_i be the covariance matrix relative to the segment s_i of the image, for $i = 1 \ldots n$. The covariance matrix W_k on n normals N_i corresponding to n segments s_i is written as:

$$W_k = J \begin{pmatrix} C_1 & (0)_{3\times3} & (0)_{3\times3} & (0)_{3\times3} & (0)_{3\times3} & (0)_{3\times1} & (0)_{3\times1} & (0)_{3\times1} & (0)_{3\times1} \\ (0)_{3\times3} & C_2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & (0)_{3\times3} & \ddots & (0)_{3\times3} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & (0)_{3\times3} & C_{n-1} & (0)_{3\times3} & \vdots & \vdots & \vdots & \vdots \\ (0)_{3\times3} & (0)_{3\times3} & (0)_{3\times3} & (0)_{3\times3} & C_n & (0)_{3\times1} & (0)_{3\times1} & (0)_{3\times1} & (0)_{3\times1} \\ (0)_{1\times3} & \cdots & \cdots & \cdots & (0)_{1\times3} & \sigma^2_{u_0} & 0 & 0 & 0 \\ (0)_{1\times3} & \cdots & \cdots & \cdots & \vdots & 0 & \sigma^2_{v_0} & 0 & 0 \\ (0)_{1\times3} & \cdots & \cdots & \cdots & \vdots & 0 & 0 & \sigma^2_{f_x} & 0 \\ (0)_{1\times3} & \cdots & \cdots & \cdots & (0)_{1\times3} & 0 & 0 & 0 & \sigma^2_{f_y} \end{pmatrix} {}^tJ,$$

where J is the Jacobian matrix of the transformation connecting the unitary normals $(N_i)_{i=1...n}$ to the vector of parameters $((s_i)_{i=1...n}, u_0, v_0, f_x, f_y)$.

This Jacobian matrix is very similar to that described in the previous section. Its dimension is $3n \times (3n + 4)$; the column i contains only 7 non-zero components, corresponding to the partial derivatives of equation i relating to components n_{xi}, n_{y_i}, n_{zi} of N_i, and a_i, b_i, c_i, u_0, v_0, f_x, f_y.

4.8.3. *Uncertainties in final localization of polyhedral objects*

As we have just seen, the final localization of polyhedral objects is obtained by iteratively improving the attitude of the model by minimization of a criterion based on the adequacy of the edges of the model to the interpretation planes of the segments of the image.

4.8.3.1. *Covariance matrix associated with the localization parameters*

We localize the object by minimizing the distance of the extremities of the edges of the model to the interpretation planes of the corresponding image segments. Hence, each pairing between an edge of the model and an image segment leads to an elaboration of the two equations, which expresses the relation of these extremities to the corresponding plane, which is in turn expressed by the nullity of the scalar product between the normal vector to the plane and the co-ordinates of the point.

Let us consider the covariance matrix W_k calculated in the previous section. Once again, we pass the uncertainty in the normals represented by W_k to that in the localizations by calculating the new Jacobian matrix J_f associated with this latter transformation.

Let us recall that at each stage k of the iterative process permitting the resolution of our equations, the variation of the localization parameters is given by:

$$\Delta A_k = J_f^+ F_k$$

where J^+ is the pseudo-inverse of J.

Immediately, we deduce the covariance of ΔA_k, $\mathrm{Cov}_{\Delta A_k} = J^+ {W_k}^t J$.

We must express Cov_{A_k} in the correct reference mark. Indeed, each iteration (as we have explained earlier) changes the reference mark to facilitate the calculation of the partial derivatives.

The covariance matrix of our parameters, C, expressed in the camera reference mark will then be:

$$C = J_{Ak}\,\mathrm{Cov}_{\Delta A_k}\,{}^{t}J_{Ak}$$

where J_{Ak} is the Jacobian of the transformation connecting the localization parameters in the initial reference mark and those in the iteration reference mark k:

$$J_{A_k} = \begin{bmatrix} \dfrac{\cos\gamma_k}{\cos\beta_k} & \dfrac{\sin\gamma_k}{\cos\beta_k} & 0 & 0 & 0 & 0 \\ -\sin\gamma_k & \cos\gamma_k & 0 & 0 & 0 & 0 \\ \cos\gamma_k\tan\beta_k & \sin\gamma_k\tan\beta_k & 1 & 0 & 0 & 0 \\ 0 & t_{z_k} & -t_{y_k} & 1 & 0 & 0 \\ -t_{z_k} & 0 & t_{x_k} & 0 & 1 & 0 \\ t_{y_k} & -t_{x_k} & 0 & 0 & 0 & 1 \end{bmatrix}$$

	α	β	γ	u	v	w
1 camera	0.9 degree	1.2 degree	0.1 degree	5.3 mm	3.5 mm	28.6 mm
2 cameras	0.9 degree	1.1 degree	0.3 degree	5.2 mm	2.7 mm	9.7 mm
3 cameras	0.8 degree	0.9 degree	0.3 degree	4.9 mm	2.4 mm	6.9 mm

Table 4.4. *Analytical results (standard deviation) on the localization precision according to the number of cameras used*

	α	β	γ	u	v	w
1 camera	100	100	99	100	100	99
2 cameras	99	100	100	100	100	99
3 cameras	100	99	100	100	100	99

Table 4.5. *Percentage of analytical deviations vis-à-vis statistics verifying that the theoretical value of the parameter is at least 3σ from the estimated value*

Table 4.4 provides the standard deviation values calculated in an analytical manner taking into account a noise on the detection of primitive visuals and the intrinsic parameters of the camera corresponding to card 2. The values presented are compared to the standard deviation obtained statistically and already presented in Table 4.3. Table 4.5 indicates, out of 100 experiments carried out with synthetic data, the number for which the theoretical value of the parameter is included in the interval centered on the estimated value and the estimated breadth $+/-3\sigma$.

4.9. Conclusion

In this chapter, we have presented some techniques which allow the retiming of a CAD model on the contents of a video image. These techniques require the knowledge of pairs between the detected primitives in the image and the three-dimensional entities corresponding to the model of the observed object. They prove to be easy in their implementation and are relatively precise in their use.

Besides, we have shown that this precision can be improved, mainly for the translation parameter along the optical axis, by taking the help of a multi-ocular system.

Moreover, the formalism developed adapts very easily to the problem of looking for the posture of a hand manipulator type articulated object.

The problem of the estimation of hand-eye passage during the usage of a camera mounted on a robot system (turret site/azimuth, hand manipulator) can also be addressed in a similar manner; the proposed method restraining the number of parameters to be estimated as compared to the previous techniques proposed in other works.

Finally, we have shown, by propagation of uncertainties, how we can obtain an estimation of the precision of the measured parameters.

4.10. Bibliography

[AYA 89] AYACHE N., *Vision Stéréoscopique et Perception Multisensorielle: Applications à la Robotique Mobile*, InterEditions, Paris, 1989.

[BRA 94] BRAUD P., DHOME M., LAPRESTE J. and DAUCHER N., "Modelled Object Pose Estimation and Tracking by a Multi-Cameras System", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Seatle, Washington, p. 976–979, 1994.

[BRA 96] BRAUD P., LAPRESTE J. and DHOME M., "Recognition, Pose and Tracking of modelled Polyhedral Objects by Multi-ocular Vision", *Proc. European Conference on Computer Vision*, Cambridge, UK, p. 455–464, 1996.

[BRO 71] BROWN D., "Close-range Camera Calibration", *Photogrammetric Engineering*, vol. 8, no. 37, p. 855-866, 1971.

[DEM 92] DEMENTHON D. and DAVIS L., "Model-based object pose in 25 lines of code", *Proc. European Conference on Computer Vision*, Santa Margherita Ligure, Italy, p. 19–22, 1992.

[DHO 89] DHOME M., RICHETIN M., LAPRESTÉ J. and RIVES G., "Determination of the Attitude of 3-D Objects from a Single Perspective View", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, p. 1265–1278, 1989.

[DHO 90] DHOME M., LAPRESTÉ J., RIVES G. and RICHETIN M., "Spatial Localization of Modelled Objects of Revolution in Monocular Perspective Vision", *ECCV*, Antibes, France, p. 1–21, 1990.

[DHO 93] DHOME M., YASSINE A. and LAVEST J., "Determination of the Pose of an Articulated Object from a Single Perspective View", *BMVA*, Guildford, UK, p. 95–104, 1993.

[FAI 87] FAIG W., "Calibration of Close-Range Photogrammetric System: Mathematical Formulation", *Proc. of XIII Congress of Int. Society for Photogrammetry*, p. 1479–1486, Tokyo, Japan, 1987.

[FAU 87] FAUGERAS O. and TOSCANI G., "Camera calibration for 3D Computer Vision", *Proc. of CVPR*, Tokyo, Japan, 1987.

[HOR 87] HORAUD R., "New Methods for Matching 3D Object with Single Perspective View", *IEEE Trans. on PAMI*, vol. PAMI-9, no. 3, p. 401–412, 1987.

[HOR 89] HORAUD R., CONIO B., LEBOULLEUX O. and LACOLLE B., "An analytic solution for the perspective 4 points problem", *CVGIP*, vol. 47, p. 33–44, 1989.

[HOR 93] HORAUD R. and DORNAIKA F., "Hand-Eye Calibration", *Workshop on Computer Vision for Space Applications*, p. 369–379, Antibes, France, September 1993.

[HOR 95a] HORAUD R., CHRISTY S., DORNAIKA F. and LAMIROY B., "Object Pose: Links between Paraperspective and Perspective", *5th International Conference on Computer Vision*, p. 426–433, Cambridge, Massachusetts, USA, June 1995.

[HOR 95b] HORAUD R. and MONGA O., *Vision par ordinateur, outils fondamentaux*, Hermes, 2nd ed., 1995.

[KAN 81] KANADE T., "Recovery of the Three Dimensional Shape of an Object from a Single View", *Artificial Intelligence, Special volume on Computer Vision*, vol. 17, p. 1-13, August 1981.

[LOW 85] LOWE D., *Perceptual Organization and Visual Recognition*, Kluwer Academic, Dordrecht, 1985.

[PRE 92] PRESS W., TEUKOLSKY S., VETTERLING W. and FLANNERY B., *Numerical Recipies in C*, Cambridge University Press, 2nd ed., 1992.

[REM 98] REMY S., DHOME M., LAVEST J. and DAUCHER N., "Etalonnage bras-œil d'un bras manipulateur équipé d'une caméra vidéo", *Actes du 11ieme congrés AFCET/RFIA*, p. 45–56, Clermont-Fd, 1998.

[SHI 89] SHIU Y. C. and AHMAD S., "Calibration of Wrist-Mounted Robotic Sensors by Solving Homogeneous Transform Equations of the Form AX = XB", *IEEE Transactions on Robotics and Automation*, vol. 5, no. 1, p. 16–29, February 1989.

[TSA 86] TSAI R., "An efficient ansd accurate calibration technique for 3D machine vision", in *Proc. of CVPR*, p. 364–374, Miami, USA, 1986.

[TSA 89] TSAI R.Y. and LENZ R.K., "A New Technique for Fully Autonomous and Efficient 3D Robotics Hand/Eye Calibration", *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, p. 345–358, June 1989.

[WAN 92] WANG C.-C., "Extrinsic Calibration of a Vision Sensor Mounted on a Robot", *IEEE Transactions on Robotics and Automation*, vol. 8, no. 2, p. 161–175, April 1992.

This page intentionally left blank

Part 3

This page intentionally left blank

Chapter 5

Reconstruction of 3D Scenes from Multiple Views

5.1. Introduction

This chapter deals with the problem of reconstruction of a fixed voluminous scene observed from several points of view. The first part is concerned with the definition of the specific geometry associated with this observation method. The notions of homolog points, epipolar geometry, fundamental matrix, essential matrix, trifocal tensor, are introduced, as well as different approaches, which make their estimation possible. The second part, which is more applicative, tackles the problem of dense matching between stereoscopic image couples.

5.2. Geometry relating to the acquisition of multiple images

5.2.1. *Geometry of two images*

The study of the geometry relating to the acquisition of two images is fundamental in computer vision. It intervenes as soon as we expect to treat the problem of stereovision, which amounts to using the contents of the two video images, which can be acquired simultaneously by a system equipped with two cameras, or can be obtained sequentially by a single moving camera.

Chapter written by Long QUAN, Luce MORIN and Lionel OISEL.

The main objective is the 3D reconstruction of the observed structure or in the same manner, an estimation of the movement between the two places of filming [FAU 90, FAU 92a, FAU 93, LUO 92, LUO 93, HAR 92b, HAR 95, TOR 97a, ZHA 94].

This objective can be achieved when correspondence between *homologous points* arising from the two images is established. To arrive at this result, it is necessary to develop techniques, which make it possible to recover the couple points of the image arising from the observation, from the two points of view, from the same 3D entity of the observed scene. All this rests in the geometry called *epipolar*, which characterizes the two images taken into consideration. We present, in this chapter, an elegant algebraic representation of the epipolar geometry.

5.2.1.1. *Geometric aspect*

Geometrically speaking, at any point on the first image, it is possible to associate the line in the space based on the latter and the optical center of the corresponding camera; this line projects in the second image in the form of a line. The latter represents the position of all the points of this second image, which can be potentially *homologous* to the point of the considered first image (see Figure 5.1). The search for the homologous point to a point in the first image is thus reduced to an examination of a line in the second image. Another way of characterizing this geometric configuration is to note that two *homologous* points and their correspondence in space are coplanar.

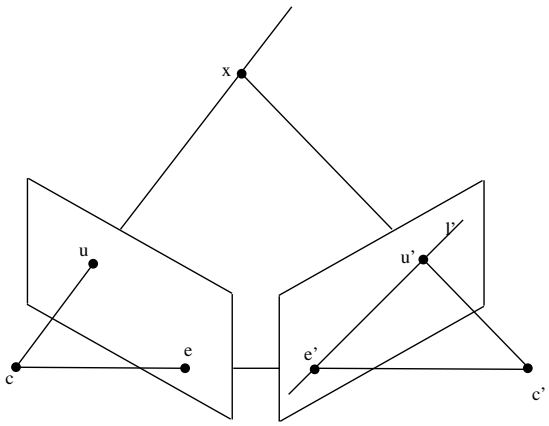


Figure 5.1. *Epipolar geometry of two images. The couple of homologous points is* $\mathbf{u} \Leftrightarrow \mathbf{u}'$. *The centers of the camera are* $\mathbf{c}$ *and* $\mathbf{c}'$. *The epipoles are* $\mathbf{e}$ *and* $\mathbf{e}'$ *in the first and second image. The line* $\mathbf{l}'$ *is the epipolar line of the point* $\mathbf{u}$

The entire geometry connecting the two images arises from this fact. Thus, any point in the first image corresponds to a line in the second, which is called the epipolar line. The geometry that follows from it is also called *epipolar geometry*. The points of intersection of the line joining the optical centers of the two cameras with the two image planes are called *epipoles*. The latter can be seen as the projection of the optical center of the camera in another image. All epipolar lines of an image pass through the corresponding epipole. The set of epipolar lines of an image thus form a beam of lines.

5.2.1.2. *Algebraic aspect*

Algebraically, this epipolar geometry is easily derived in the following manner (already tackled in Chapter 2). Without loss of generality, we can associate, with the two images, the following two projection matrices, respectively: $\mathbf{P}_{3\times4} = (\mathbf{I}, \mathbf{0})$ and $\mathbf{P}'_{3\times4} = (\mathbf{A}, \mathbf{a})$. Let $\mathbf{u} \Leftrightarrow \mathbf{u}'$ be a pair of homologous points between the two images created from the observation of a point in the space $\mathbf{x}$.

The line in space associated with point $\mathbf{u}$ is given by the optical center of the camera $\mathbf{c} = (\mathbf{0}, 1)^T$ and its *direction* $\mathbf{x}_\infty = (\mathbf{I}^{-1}\mathbf{u}, 0)^T$ is given by the infinite point of this line. The images of these two points in the second image are: $\mathbf{e}' = (\mathbf{A}, \mathbf{a})\mathbf{c} = \mathbf{a}$ and $\mathbf{u}'_\infty = (\mathbf{A}, \mathbf{a})\mathbf{x}_\infty = \mathbf{A}\mathbf{u}$, respectively. The epipolar line is then given by $\mathbf{l}' = \mathbf{e}' \times \mathbf{u}'_\infty$. By introducing the anti-symmetric matrix $[\mathbf{a}]_\times$ associated with vector $\mathbf{a}$ to represent the vectorial product denoted $\mathbf{F} = [\mathbf{a}]_\times \mathbf{A}$, we have:

$$\mathbf{l}' = \mathbf{F}\mathbf{u}.$$

As point $\mathbf{u}'$ is found on the line $\mathbf{l}'$, it verifies $\mathbf{u}'^T\mathbf{l}' = 0$ and from this we deduce the fundamental relation of the epipolar geometry between the two projections $\mathbf{u}$ and $\mathbf{u}'$:

$$\mathbf{u}'^T\mathbf{F}\mathbf{u} = 0.$$

The matrix $\mathbf{F}$ associated with the two images is called the *fundamental matrix*.

5.2.1.3. *Properties of* $\mathbf{F}$

Let us examine the properties of this matrix:

– The matrix $\mathbf{F}$ is singular and of row 2 since it contains an anti-symmetric matrix 3×3, of row 2. Geometrically, the set of epipolar lines passes through a common point, i.e. the epipole. They form a beam of lines of dimension 1.

– The core of $\mathbf{F}$ is the epipole in the first image. To find the other epipole, it is sufficient to transpose $\mathbf{F}$, i.e, $\mathbf{F}^T\mathbf{e}' = 0$, because if $\mathbf{u}'^T\mathbf{F}\mathbf{u} = 0$ gives the epipolar geometry between the first and the second, transposing the set gives the equation by considering the images in the reverse order.

– The number of degrees of freedom of $\mathbf{F}$ is 7. Algebraically, $\mathbf{F}$ has 9 (3×3) homogenous elements, which makes only 8 degrees of freedom. As it is singular, there remain only 7 degrees of freedom. Geometrically, each epipole accounts for 2 degrees of freedom and that is 4 in total for the two epipoles. The two beams of epipolar lines are in homographic correspondence to dimension 1, hence accounting for 3 degrees of freedom. This corresponds to a total number of degrees of freedom equal to 7.

– The epipolar line in the second image is given by $\mathbf{F}\mathbf{u}$ and reciprocally by $\mathbf{F}^T\mathbf{u}'$ in the first.

For calibrated cameras, the image points are subjected to an affine transformation defined by the intrinsic parameters of each camera. Given the intrinsic parameters $\mathbf{C}$ and $\mathbf{C}'$ of the two images and the *calibrated* image points $\mathbf{x} = \mathbf{C}^{-1}\mathbf{u}$ and $\mathbf{x}' = \mathbf{C}'^{-1}\mathbf{u}'$, we obtain

$$\mathbf{x}'^T\mathbf{E}\mathbf{x} = 0,$$

where $\mathbf{E} = \mathbf{C}'^T\mathbf{F}\mathbf{C}$. Matrix $\mathbf{E}$ is called the *essential matrix*. It possesses all the properties of the fundamental matrix $\mathbf{F}$ and one more of its own: two non-zero singular values are identical [HUA 89].

5.2.1.4. *Estimation of the fundamental matrix*

Each correspondence of points between two images $\mathbf{u} \Leftrightarrow \mathbf{u}'$ gives rise to a linear equation $\mathbf{u}'^T\mathbf{F}\mathbf{u} = 0$ relative to the parameters of the matrix $\mathbf{F}$, which can be written as:

$$(u'u, u'v, u', v'u, v'v, v', u, v, 1)(f_1, \ldots, f_9) = 0.$$

For N points in correspondence, we obtain a system of equations, which are linear and homogenous:

$$\mathbf{A}_{N\times 9}\mathbf{f}_9 = 0.$$

5.2.1.5. *7 point algorithm*

Since F has only 7 degrees of freedom, algebraically, it is sufficient to have at least 7 points in correspondence to be able to resolve this system,

which results from it. However, with 7 points in correspondence, it is impossible to obtain a single solution. By resolving the homogenous linear system $\mathbf{A}_{7\times 9}\mathbf{f}_9 = 0$, a set of solutions $\mathbf{f} = x\mathbf{a} + y\mathbf{b}$ is determined with two close homogenous parameters x and y. It is enough to add the corresponding constraint to the fact that the determinant of $\mathbf{F}$ is zero. By analytically developing this determinant, a cubic equation in x and y is obtained:

$$ax^3 + bx^2y + cxy^2 + dy^3 = 0.$$

This is equivalent to the Sturm method [STU 69], which determines the epipolar geometry from 7 points in correspondence: a method already well known in the last century. Even though the original version is more algebraic and complicated as it is based on the usage of canonic projective bases, this simplified version is easily implemented with numerical linear algebraic tools.

5.2.1.6. *8 point algorithm*

If we add one (or more) correspondence point and ignore the row constraint on the matrix $\mathbf{F}$, the solution is very easily given by the linear system $\mathbf{A}_{8\times 9}\mathbf{f}_9 = 0$. This algorithm of 8 points has also been known for a long time and references can be found in [LON 81, FAU 95a].

Unfortunately, this linear algorithm is numerically unstable and can be corrected by normalizing the data [HAR 95]. First of all, we transform the image points by the relations $\tilde{\mathbf{u}} = \mathbf{A}\mathbf{u}$ and $\tilde{\mathbf{u}'} = \mathbf{A}'\mathbf{u}'$ in such a way that the group of transformed points is centered and the mean distance of the new points to the center is equal to 1. We then apply the linear method to the transformed points $\tilde{\mathbf{A}}_{8\times 9}\tilde{\mathbf{f}}_9 = 0$. The solution $\tilde{\mathbf{f}}_9$ minimizing $\|\tilde{\mathbf{A}}_{8\times 9}\tilde{\mathbf{f}}_9\|$ under the constraint $\|\tilde{\mathbf{f}}_9\| = 1$ is obtained by taking the singular vector line corresponding to the smallest singular value. This solution still does not give a matrix $\tilde{\mathbf{D}}$ of row 2. It is enough to use the decomposition in the principal SVD components so that it becomes the matrix $\tilde{\mathbf{D}}$ of row 2. The original fundamental matrix is given by:

$$\mathbf{F} = \mathbf{A}'^{T}\tilde{\mathbf{F}}\mathbf{A}.$$

5.2.1.7. *Optimal algorithms*

To obtain a statistically optimal estimation, it is necessary to use a numerical optimization method on geometric errors. An obvious way is to

minimize the Euclidean distance between the point and the corresponding epipolar line, i.e.:

$$\min_{\mathbf{F}} \sum_i \left(d_i^2 + d_i'^2\right),$$

where $d_i^2 = \mathrm{dist}(\mathbf{u}_i, \mathbf{F}^T\mathbf{u}_i')$ and $d_i'^2 = \mathrm{dist}(\mathbf{u}_i', \mathbf{F}\mathbf{u}_i)$ by taking into account the symmetry in the two images.

5.2.1.8. *Robust algorithms which make it possible to eliminate false pairing between a couple of points*

When there are false mappings among all established pairs of homologous points, the methods given by the robust statistics, RANSAC or the least square median, based on a random sampling of a sub-set of homologous point pairs, prove to be efficient. The principle consists of randomly selecting 7 point correspondences and then estimating the fundamental matrix with this sample. We retain the sample which maximizes the consensus (RANSAC) or which minimizes the median error (LMS). Finally, an optimum estimation method is applied on all the correspondences retained by RANSAC or LMS. The only worry is the cost of this random sampling. With a sample size of 7, it is sufficient to carry out 382 draws for 50% false correspondences, which is perfectly reasonable.

An example of geometric estimation of two images is illustrated in Figure 5.2.

Figure 5.2. *Estimation of the epipolar geometry of two images of the INRIA building in Grenoble*

5.2.2. *Geometry of 3 images*

The study of the geometry of 3 images is a natural extension of that of 2 images. The objectives are always the same: namely, either the three-dimensional reconstruction of the observed scene or an estimation of the movement between the different image acquisitions or the dense mapping of points between images [SHA 95, SPE 90, HAR 97, QUA 95, TOR 97b, CAR 95, CAR 98, MOH 92].

Geometrically, it is sufficient to know the epipolar geometry between the first image and the third, and between the second and the third so that the correspondence point in the third image is completely determined, except in the case of singular spatial disposition of the cameras. The point in the third image is hence said to be *transferred* from the first to the second (see Figure 5.3).

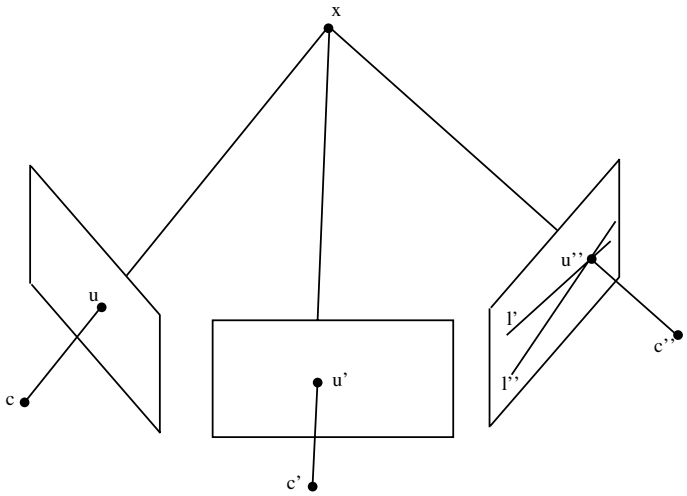


Figure 5.3. *Geometry of 3 images*

Algebraically, this transfer relation is deduced in the following manner. Let us take a triplet of correspondence points between the 3 images ($\mathbf{u} \Leftrightarrow \mathbf{u}' \Leftrightarrow \mathbf{u}''$), as well as the projection matrices for the 3 cameras:

$$\mathbf{P} = (\mathbf{I}, \mathbf{0}), \quad \mathbf{P}' = (\mathbf{A}, \mathbf{a}), \quad \mathbf{P}'' = (\mathbf{B}, \mathbf{b}) \text{ where } \mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \mathbf{a}_3^T \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \mathbf{b}_3^T \end{pmatrix}.$$

For the point in space $\mathbf{x} = (x, y, z, t)$, using the first image, we obtain $\mathbf{x} = (\mathbf{u}, t)$. By projecting $(\mathbf{u}, t)$ on the second image, we have:

$$\lambda' \mathbf{u}' = (\mathbf{A}, \mathbf{a}) \begin{pmatrix} \mathbf{u} \\ t \end{pmatrix}$$

From this vectorial equation, we find two possibilities for scalar t:

$$t_1 = \frac{\mathbf{a}_1^T \mathbf{u} - u' \mathbf{a}_3^T \mathbf{u}}{u' a_3 - a_1}$$

or

$$t_2 = \frac{\mathbf{a}_2^T \mathbf{u} - v' \mathbf{a}_3^T \mathbf{u}}{v' a_3 - a_2}.$$

With one of the values of t, we can project the point in the space, which is now known, in the third image. By correctly arranging the terms and by introducing the vectors:

$$\mathbf{t}_{ij} = a_i \mathbf{b}_j^T - b_j \mathbf{a}_i^T,$$

we obtain two equation systems with the two values of t:

$$\lambda'' \mathbf{u}'' = \begin{pmatrix} u' \mathbf{t}_{31}^T - \mathbf{t}_{11}^T \\ u' \mathbf{t}_{32}^T - \mathbf{t}_{12}^T \\ u' \mathbf{t}_{33}^T - \mathbf{t}_{13}^T \end{pmatrix} \mathbf{u} = \begin{pmatrix} v' \mathbf{t}_{31}^T - \mathbf{t}_{21}^T \\ v' \mathbf{t}_{32}^T - \mathbf{t}_{22}^T \\ v' \mathbf{t}_{33}^T - \mathbf{t}_{23}^T \end{pmatrix} \mathbf{u}$$

These two blocks of 3 equations are homogenous; by eliminating the scale factor, we obtain 4 scalar equations on 9 $\mathbf{t}_{ij}$ vectors. It is very interesting to observe that by continuing the index play of the t_{ij} vectors and by noting the kth element of the $\mathbf{t}_{ij}$ vector with a third index k, the set of elements indexed by i, j and k form a tensor with 3 indices T_k^{ij} varying from 1 to 3 that we call the trifocal tensor, which plays exactly the same role as the fundamental matrix for two images. If we have to scrupulously respect the conventions on the variant/covariant indices, this tensor has to be correctly presented as T_k^{ij}.

We can note that the geometry of the lines drawn from the 3 images is also governed by the same tensor. This has an important significance: the estimation of the geometry between 3 images can be indifferently estimated from homologous points or lines.

In spite of a number of mathematically elegant properties of the trifocal tensor, it does not provide an ideal parametrization to characterize the geometry between 3 images and moreover, its numerical estimation still remains precarious. The linear algorithm based on the usage of at least 7 homologous points or other combinations of points and lines provided by estimation is still very approximate. Finally, the implementation of an optimum estimation based on a numerical minimization is delicate because of its $8 = 27 - 1 - 18$ algebraic constraints connecting the various components of the tensor.

On the contrary, the projective structure in space provides a minimum parametrization [QUA 95, TOR 97b]. Indeed, in the case of 3 images, the projective structure can be directly determined from the mapping of 6 points. Let us assume that the projective and canonic coordinates of the sixth point in space are (X, Y, Z, T) if we choose for the first five the canonic projective base of the projective space of dimension 3. It is the same for the points in each of the images; the fifth and sixth points can be transformed in (u_5, v_5, w_5) and (u_6, v_6, w_6) if we choose the canonic projective base for the first 4 in the image plane. The image points and the points in space are connected by an unknown projection matrix. By eliminating all entries of the projection matrix with the image-space correspondence of 5 points, we obtain the following equation:

$$w_6(u_5 - v_5)XY + v_6(w_5 - u_5)XZ + u_5(v_6 - w_6)XT + u_6(v_5 - w_5)YZ + v_5(w_6 - u_6)YT + w_5(u_6 - v_6)ZT = 0.$$

This is an algebraic equation of unknowns X, Y, Z and T using only one single image. Given 3 images, we will obtain a system of 3 equations. With algebraic manipulations, we return to a cubic equation on one of the unknowns:

$$aX^3 + bX^2T + cXT^2 + dT^3 = 0.$$

It is not by chance that this algorithm and the Sturm method, both lead to resolving a polynomial of degree 3: an intrinsic duality is explained by Carlson in [CAR 95, CAR 98].

To conclude, the trifocal tensor has interesting properties, but it does not prove to be a good parametrization of the geometry between 3 images. A robust method based on the algorithm of 6 points, followed by a numerical optimization (of bundle adjustment type) probably corresponds to the most satisfying solution at the moment.

Figure 5.4. *One of the three images of the small house made of wood used for the estimation of the geometry of three images. The points used are circled in white*

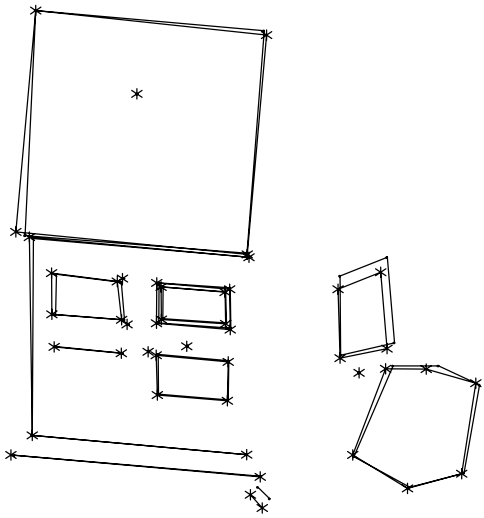


Figure 5.5. *The projective reconstruction rectified from three non-calibrated images: the points are shown as stars and the line segments as continuous lines. The Euclidean reconstruction from 5 images: the segments are shown as dotted lines. The 5 reference markers are* $\{2, 5, 8, 10, 11\}$ *(see Figure 5.4)*

5.2.3. *Geometry beyond 3 images*

A legitimate and natural question to be asked is whether there are constraints of similar type on more than 3 images. In fact, all these constraints can be systematically derived in the following manner [TRI 95, FAU 95b].

Given that a point $\mathbf{x}$ is seen in N images:

$$\begin{cases} \lambda\mathbf{u} = \mathbf{P}\mathbf{x}, \\ \lambda'\mathbf{u}' = \mathbf{P}'\mathbf{x}, \\ \vdots \\ \lambda^{(N)}\mathbf{u}^{(N)} = \mathbf{P}^{(N)}\mathbf{x}. \end{cases}$$

these equations can be written in matrix form:

$$\underbrace{\begin{pmatrix} \mathbf{P} & \mathbf{u} & 0 & 0 & 0 \\ \mathbf{P}' & 0 & \mathbf{u}' & 0 & 0 \\ \mathbf{P}'' & 0 & 0 & \cdots & 0 \\ \mathbf{P}'' & 0 & 0 & 0 & \mathbf{u}^{(N)} \end{pmatrix}}_{\mathbf{M}} \begin{pmatrix} \mathbf{x} \\ -\lambda \\ -\lambda' \\ \cdots \\ -\lambda^{(N)} \end{pmatrix} = 0,$$

The vector $(\mathbf{x}, -\lambda, -\lambda', \ldots, -\lambda^{(N)})^T$ cannot be zero, hence the row of the $\mathbf{M}$ matrix cannot exceed $N + 4$. This implies that all its minors $(N + 4) \times (N + 4)$ are cancelled. Expansion of all these minors gives all the imaginable geometric constraints between the multiple images.

Of course, these minors can be formed in different ways. When the minors are formed only from elements involving two images, the expansion gives the constraints of geometry between these 2 images. Similarly, when the minors are formed from elements intervening in three images, the expansion gives the constraints between these 3 images. Just as $\mathbf{M}$ has $N + 4$ columns and $3N$ lines, it can only have a maximum of 4 projection matrices intervening in the formation of the minors. Hence, there are geometric constraints for 4 images, which we call quadrifocal constraints. Unfortunately, quadrifocal constraints are not algebraically independent, i.e., they are easily decomposable in those of 3 images and of 2 images. Moreover, these constraints are very redundant because of Grassman intrinsic quadratic relations. The numerical exploitation of these constraints is not yet completely elucidated.

5.3. Matching

Now we present different algorithmic techniques, making it possible to find the correspondence mapping between two views of the same rigid scene. This stage called tracking or matching consists of identifying the points in the images, which are projected from the same 3D physical point and which are called correspondent or homologous points.

Our objective is to obtain a dense mapping, i.e., to find for each pixels of the first image, all the homologous points for this pixel in the second image. The simplification, which is generally carried out considers that luminance of the two points in correspondence coming from the projection of the same 3D point is identical (or very close). If the mapping problem seems simple to express it is because it has been the subject of many works for a long time. Indeed, a number of difficulties are faced following the type of scenes processed. It is especially so in the case of:

– homogenous zones. In these regions, the set of pixels possess the same value of luminance. The matching problem then consists of finding a good pixel in the set possessing the same characteristics;

– occultation zones (see Figure 5.6). In this case, a pixel of an image does not have a correspondence in the second image;

– contraction zones (see Figure 5.6). This corresponds to the case where a projected object from face or profile does not cover the same surface in the image. This means that a pixel in an image can correspond to several pixels in the second image. A pixel having several pixels as correspondents defines a contraction. In such a case, is it necessary to authorize a pixel to have several correspondents or not? Symmetrically, is it necessary to authorize several pixels to have the same correspondent in the second image?

A number of tracking algorithms exist (see state of the art in [ZHA 93] and [DHO 89]). We do not propose an exhaustive description of the different existing methods. However, we present the major classes of approach used to achieve a dense mapping.

5.3.1. *State of the art elements*

As we have seen, the rigidity hypothesis associated with the perspective projection model induces the possibility of using the epipolar geometry in order to restrict the matching. This constraint is used in the majority of algorithms presented.

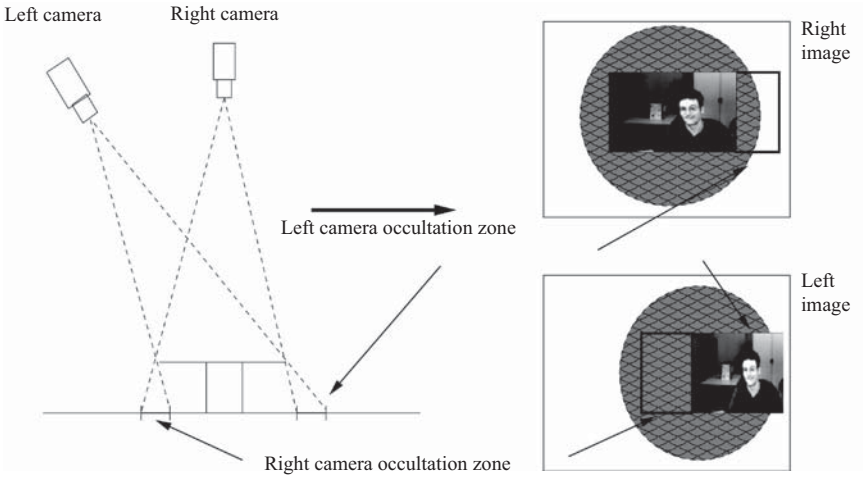


Figure 5.6. *Example of occultations and contractions*

The result of dense mapping is given in the form of a displacement vector field, or a disparity field: it is an image whose value in each pixel is a vector with two dimensions indicating the apparent movement of the pixel between the two images; the movement vector applied to a pixel of the first image gives the position of its correspondent in the second image. The disparity vector has coordinates of integer values, which produces a pixel to pixel correspondence (local methods, block-matching, and dynamic programming), or of real coordinates, which allows a more precise estimation of the disparity (optical flow, energy modeling). In cases where the epipolar constraint is used, a single coordinate is sufficient to define the disparity vector.

5.3.1.1. *Correlation*

The correlation methods produce displacements of integer value components: for a pixel in the right image, we study its correspondent(s) in the left image. The method consists of using a resemblance criterion for comparing a region enclosing the processed pixel with all the regions centered on the associated epipolar line in the other image. The resemblance criteria are based on a correlation coefficient [ASC 92]. The pixel in the left image driving the strongest correlation coefficient is associated with the current right pixel. It is clear that this type of technique does not solve the problems posed by homogenous or disturbed zones. This is why less local methods, taking into account the notion of continuity constraint in a region, have been developed.

5.3.1.2. *Block-matching*

The first method due to its frequent utilization, is a technique called block-matching tackled in [TZI 94]. This algorithm has a multitude of declinations and basically consists of cutting the image into blocks. For each block that is generated, we search for the best block of same size in the second image. Hence, it is possible to perform motion estimation in each pixel by gradually decreasing the size of the blocks while beginning the search for the displacement vector from the father block. This algorithm used within the general framework of the motion estimation can be made more robust by the usage of epipolar geometry. For a given block, the study of its correspondent is done only in a zone close to the epipolar line associated with its center [TAM 92].

5.3.1.3. *Dynamic programming*

The second method, specific to stereovision, consists of matching two homologous epipolar lines (i.e., two lines whose points correspond). This method calls for dynamic programming techniques while using, here also, simple correlation coefficients to match the pixels [BLA 98]. The algorithm begins to calculate the cost of the correlation resulting from the pairing of the two first points. The table is then prepared following a horizontal, vertical, or diagonal displacement. A cost is put on each case of the table corresponding to a given pair. This cost is calculated by adding the minimum cost to arrive up to this case and the cost of pairing.

For each line couple, we study the optimum path in the graph of possible mappings (see Figure 5.7). This method presents an interest to introduce a smoothing along the given epipolar line. It also makes it possible to respect the constraint order along the epipolar line. On the other hand, its major disadvantage is the absence of smoothing between neighboring epipolar lines. The resulting disparity map presents significant disparity differences, line by line, especially on the border of the occultation zones. Here also, a number of declinations of this algorithm exist allowing two-dimensional smoothing (for example, in close epipolar lines [COX 92]). Blanc, in his thesis, proposes a state of the art for these different variations around dynamic programming [BLA 98].

5.3.1.4. *Association of the optical flow and epipolar geometry*

In [DEM 96], an approach by optical flow calculation applied to the cases of static image scenes is proposed. The invariance hypothesis of the brightness

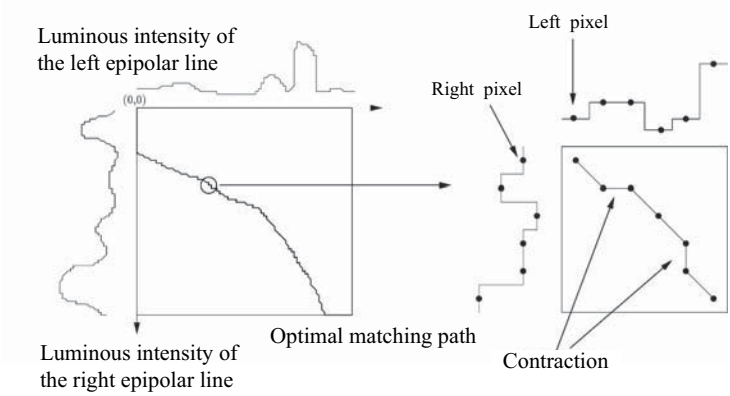


Figure 5.7. *Dynamic programming*

associated with the projection of a point $\mathbf{m}$ is given by:

$$\frac{dI(\mathbf{m},t)}{dt} = 0$$

A Taylor expansion of this expression gives:

$$\nabla I \cdot \left(\mathbf{m}_{t+1} - \mathbf{m}_t\right) + I_t = 0$$

where $\mathbf{m}_i$ represents the point $\mathbf{m}$ at time i, ∇ is the spatial gradient operator and I_t is the temporal gradient. From this, we find that the point $\mathbf{m}_{t+1}$ belongs to the projective line of coordinates:

$$\tilde{\delta} = \begin{bmatrix} I_x \\ I_y \\ I_t - \nabla I \cdot \mathbf{m}_t \end{bmatrix}.$$

Similarly, by knowing the epipolar geometry (by calibration in the article), it is possible to calculate the epipolar line associated with point $\mathbf{m}_t$ in the second image. Since point $\mathbf{m}_{t+1}$ must belong to this line and to "$\tilde{\delta}$", it is found at their intersection.

In order to make the calculation of the flow more robust to noise, a smoothing constraint is introduced by assuming that the disparity vector is constant in a 7×7 block centered at point $\mathbf{m}_t$. For this to be valid, the hypotheses of the parallel epipolar lines (pure translation between images,

for example) and low variations in depth must be verified. Each point in the neighborhood hence provides a line $\tilde{\delta}_i$, which intersects the epipolar line. The selected homologous point is the barycenter of the intersection points (a weighting function may be used to increase the role of the points close to $\mathbf{m}_t$).

5.3.1.5. *Energy modeling*

The energy approaches that we present here are approaches to continuous disparity values (the correspondent of a pixel can have real coordinates). Here, the modeling of the problem is done in a global manner and gives a global solution (it is the configuration for which the energy is minimum). The energy function decomposes into two terms:

– the observation term that represents the constraints, which we impose in each pixel independently and which must be verified by the displacement vector;

– the smoothing term, which represents the inter-pixel dependence constraints between neighboring displacements.

Here, we briefly present some methods requiring energy modeling:

– Method using the epipolar constraint: Ouali *et al*. in [OUA 96] proposed such a modeling in order to track two stereoscopic images while detecting the occultation zones. In their case, observation term is based on a resemblance function (between two pixels) similar to a calculation of the correlation coefficients. This coefficient is then weighed by the indicator of presence of an occultation zone in the neighborhood of the considered pixel. The observation term thus imposes on the data that two neighboring pixels have very close disparity values. In order to take into account the possible discontinuities of depth, this term is cancelled on luminance edges. Finally, a third energy term is added to limit the number of occultation zones. An stochastic optimization process constrained by the epipolar geometry assures that we obtain an optimum solution, but at a very high calculation cost.

– Multi-grid multi-resolution method: Weng proposed a method, which is inspired more directly from algorithms of optical flow calculation. As emphasized in [WEN 92], these techniques present some problems in the case of estimation of disparity maps (discontinuities of depth, significant disparity values). In order to overcome these difficulties, Weng proposed a multi-grid multi-resolution algorithm using information of intensity, curves and singular points while modeling the occultations and discontinuities. In general, the

results show that large displacements (of the order of 80 pixels) are estimated. However, some problems arise in the case of weak textured images or in cases where false matching of angles lead to a drift in the residual values calculation.

– Explicit 3D reconstruction method: Robert *et al*. in [ROB 96] studied the problem of dense 3D reconstruction from a pair of stereoscopic images. Here, the mapping stages and the calculation of depth are simultaneous. This implies a prior knowledge of the calibration parameters of cameras. Knowing the intrinsic and extrinsic parameters, they showed that finding the correspondent $\mathbf{p}_2$ of a pixel $\mathbf{p}_1$ of the left image amounts to calculating the depth $Z(\mathbf{p}_1)$ of the corresponding 3D point. This problem is then formalized under the form of partial differential equations (PDE) resolved using an iterative method.

5.3.2. *Dense estimation algorithm based on optical flow*

We now propose to describe in more detail a method for estimating a dense field constrained by epipolar geometry. This method is based on energy modeling.

5.3.2.1. *Hypothesis for the conservation of brightness*

The algorithm developed is inspired from the works carried out by Mémin and Perez [MÉM 96b, MEM 96a]. Let I_1 and I_2 be the right and left images, respectively. To facilitate easy writing, we will note by $I_i(s)$ the value of brightness of the image i, at point s.

The hypothesis of brightness conservation between the two projections of the same point in space coupled with taking into account of the epipolar constraint gives the following equation:

$$DFD(s) = I_1(s) - I_2(s + ds) = I_1(s) - I_2\big(s + \vec{N}_s + \lambda_s \vec{V}_s\big) = 0 \quad (5.1)$$

where ds is the displacement in the image, $\vec{N}_s$ is the normal vector of the right epipolar line associated with s, and $\vec{V}_s$ is the vector director of the right epipolar line (see Figure 5.8). For details concerning the validity conditions of this hypothesis known as *DFD* (*Displaced Frame Difference*), see [FRA 91].

Under the hypothesis that $\lambda_s \vec{V}_s$ is small before $s + \vec{N}_s$, a linearization around the position $s + \vec{N}_s$ can be realized. Relation (5.1) is then written as $I_2(s + \vec{N}_s) = \tilde{I}_2(s)$:

$$DFD(s) = \lambda_s \vec{V}_s \cdot \nabla \tilde{I}_2(s) + \tilde{I}_2(s) - I_1(s)$$

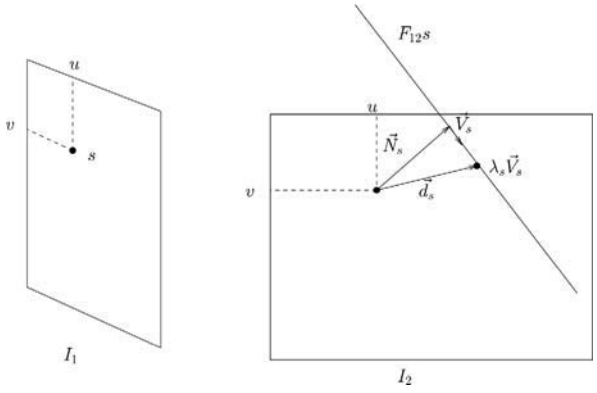


Figure 5.8. *Epipolar constraint*

where terms of equality permute, i.e. $\tilde{I}_2(s) = \tilde{I}_2(s + \vec{N}_s)$ and ∇ is the vector representing the spatial gradient.

5.3.2.2. *Energy modeling*

We now place ourselves within the Markovian framework (details are given in [PÉR 98]). An estimation of the best disparity field agreeing with the Bayesian criterion of MAP (maximum *a posteriori*) amounts to a problem of global minimization of the following energy function:

$$H(\lambda) = H_1\big(\lambda_s\big) + \alpha H_2\big(\lambda_s\big) \tag{5.2}$$

with α as the fixed weighting coefficient. H_1 is a term related to the observations in the images (linearized DFD):

$$H_1 = \sum_{s\in S} \rho\Big[\lambda_s\vec{V}_s \cdot \nabla\tilde{I}_2(s) + \tilde{I}_2(s) - I_1(s)\Big]^2$$

H_2 enforces smoothing in the pixel neighborhood:

$$H_2 = \alpha \sum_{<s,r>} \rho\big\|\lambda_s\vec{V}_s + \vec{N}_s - \lambda_r\vec{V}_r - \vec{N}_r\big\|^2$$

where $\langle s, r\rangle$ is the set of couples formed by s and its 4 neighbors. H_2 tends to minimize the difference between the neighboring disparity vectors $\vec{d}_s$ and $\vec{d}_r$ where $\vec{d}_s = \lambda_s\vec{V}_s + \vec{N}_s$ and $\vec{d}_r = \lambda_r\vec{V}_r + \vec{N}_r$.

These two terms also use robust estimators ρ, which make it possible to authorize deviations either in comparison to the model (occultation zones) or in comparison to smoothing (discontinuities of depth). Under certain assumptions, the robust function amounts to a specific sum of quadratic terms: a strong deviation in comparison to the model or to smoothing leads to a low weighting of the associated energy. In addition, more details on the functioning of robust estimators can be found in [HUB 81, PÉR 98].

5.3.2.3. *Multi-resolution minimization diagram*

In the majority of cases, the disparities reach values which are too large for the Taylor expansion of DFD to be valid. This problem is solved using a multi-resolution process. The fundamental matrix associated with a level of resolution k is calculated from the fundamental matrix obtained at the finest resolution level by a change of reference frame given by matrix $\mathbf{M}$:

$$\mathbf{F}^k = \mathbf{M}^{kT}\mathbf{F}\mathbf{M}^k \quad \text{with} \quad \mathbf{M} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $\mathbf{M} = \mathrm{diag}(2, 2, 1)$.

Thus, $\mathbf{F}^k$ matrix makes it possible to calculate the vectors $\vec{N}_s^k$ and $\vec{V}_s^k$ for each position s. The field of disparity vectors obtained at the resolution level k is then projected at the resolution level $k-1$ in order to start the process of energy minimization.

For each resolution level k, we then look to carry out a global minimization of $H(\lambda^k)$, namely, to find the set of λ^k for which the value of H is minimum. Since the energy function is convex, $H(\lambda_s^k)$ admits a minimum, which is the zero derivative point:

$$\frac{\partial H(\lambda_s)}{\partial \lambda_s} = 0$$

For this, a multiresolution scheme based on Gauss-Seidel minimization associated with an incremental formulation is implemented: in each point s, we look for the minimum of $H(\lambda_s^k)$ by freezing the other variables (the function being quadratic, we obtain a closed form solution for this minimum). After convergence, the weights associated with each pixel are evaluated. These two stages are repeated until the final convergence.

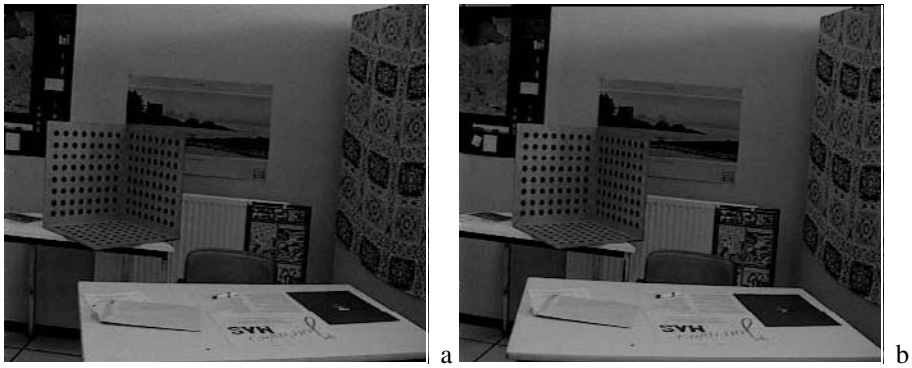


Figure 5.9. *Examples of image pairs (Armel sequences: image 1 (a) and image 50 (b))*

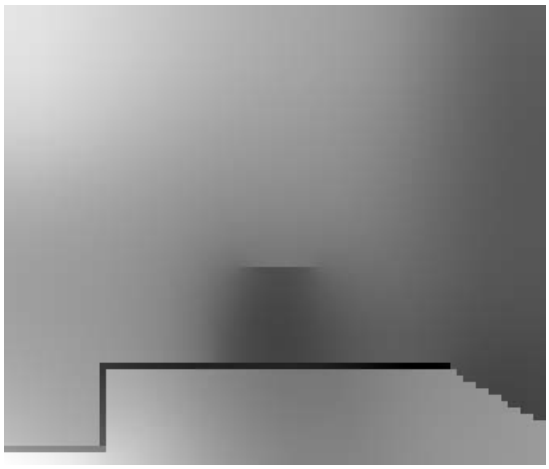


Figure 5.10. *Disparity field obtained on images 1 and 50; the posted value is the norm of the disparity vector*

5.4. 3D reconstruction

Once the disparity field is estimated, the 3D reconstruction consists of recovering the 3D coordinates of the points from the coordinates of projections in the images. This requires a calibration (possibly weak calibration) of the camera parameters. From now on we will focus we assume to have two images for which dense mapping has been done and we are on restoring the 3D information.

5.4.1. *Reconstruction principle: retro-projection*

We assume that the mapping carried out and the projection matrices P_1 and P_2 are known for each of the images. From these matrices, it is easy to show the 3D point $\mathbf{x}$ corresponding to a point $\mathbf{p}_1$ of the first image knowing its correspondent $\mathbf{p}_2$ in the second image. It is sufficient to resolve the system:

$$\begin{cases} \tilde{\mathbf{p}}_1 = \mathbf{P}_1\mathbf{x} \\ \tilde{\mathbf{p}}_2 = \mathbf{P}_2\mathbf{x} \end{cases} \tag{5.3}$$

This system provides four equations for three unknowns (we fix the last homogenous coordinate of $\mathbf{x}$ at 1). The solution amounts to looking for the intersection of the view lines $(\mathcal{C}_1, \mathbf{p_1})$ and $(\mathcal{C}_2, \mathbf{p_2})$. We talk of reconstruction by triangulation or retro-projection.

If the mapping exactly verifies the epipolar constraint (i.e., $\tilde{\mathbf{p}}_1^t\mathbf{F}\tilde{\mathbf{p}}_2 = 0$), and the projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are coherent with the fundamental matrix $\mathbf{F}$, the view lines intersect: the equations are interdependent and the system amounts to a system of 3 equations with 3 unknowns for which $\mathbf{x}$ is the only solution.

In the practical case where $\mathbf{F}$ and/or $\mathbf{P}_1$ and $\mathbf{P}_2$ are estimated from the image data, estimation errors prevent us from having a perfect coherence between $\mathbf{F}$, $\mathbf{P}_1$, $\mathbf{P}_2$ and the mapping. Hence, we can:

– recalculate the epipolar geometry obtained from $\mathbf{P}_1$ and $\mathbf{P}_2$ and impose its respect by projecting the correspondents on epipolar lines;

– resolve the system of least squares by going back to a system $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{x}$ is the vector containing the coordinates of the 3D point; this minimizes the Euclidean distance of the point to the 4 planes (on the condition of normalizing the equations), but not the distance from the point to the viewed lines;

– determine the 3D point which minimizes the distance to the view lines.

5.4.2. *Projective reconstruction*

In cases where we do not have a calibration of the camera parameters, it is possible to produce geometric 3D information in the form of a projective reconstruction. Here, the 3D points obtained correspond to a transformation of the Euclidean scene by a 3D homography (general linear transformation of homogenous 3D coordinates).

The principle of the projective reconstruction is to find a couple of projection matrices $\mathbf{P}_1$, $\mathbf{P}_2$ compatible with the image data and to use them for the retro-projection as described earlier. We re-write projection equations (5.3) in the following manner:

$$s_1\tilde{\mathbf{p}}_1 = s_2\mathbf{A}_1\mathbf{R}\mathbf{A}_2^{-1}\tilde{\mathbf{p}}_2 + \mathbf{A}_1\mathbf{t} \tag{5.4}$$

where $\mathbf{p}_1$ and $\mathbf{p}_2$ are the projections in images 1 and 2 from a point in the space, s_i is a multiplicative factor associated with the point of the image i, $\tilde{\mathbf{p}}_i$ is the vector of homogenous coordinates associated with the point $\mathbf{p}_i$, $\mathbf{A}_i$ is the matrix of intrinsic parameters associated with the image i, $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation matrices of the reference frame of the first camera towards that of the second camera.

For each pair of matched pixels, we then obtain three equations and two additional unknowns (s_1 and s_2). By assuming that we have a significant number of points in correspondence (dense disparity field calculated previously), it is possible to know the matrices $\mathbf{A}_1\mathbf{R}\mathbf{A}_2^{-1}$ and $\mathbf{A}_1\mathbf{t}$. The latter, however, will be known only up to a scale factor. If $\mathbf{H}$ is a 4×4 reversible matrix, then $\mathbf{A'}_1 = \mathbf{A}_1\mathbf{H}$, $\mathbf{A}'_2 = \mathbf{A}_2\mathbf{H}$, and $\mathbf{R}' = \mathbf{H}^{-1}\mathbf{R}\mathbf{H}$ are also solutions of system (5.4).

From two views of the same scene, we find an infinite number of matrices with intrinsic and extrinsic parameters compatible with the disparity information. Each solution obtained is deduced from the real calibration matrices by a projective transformation. Such a transformation is expressed in matrix form by a 4×4 matrix defined up to a scale factor.

Thus, by knowing the fundamental matrix, Faugeras and Hartley showed that we can calculate the second projection matrix $\mathbf{P}_2$ after having arbitrarily fixed the first one as a canonic projection matrix (noted by $\mathbf{P}_1$) [BEA 94, FAU 92b, HAR 92b, HAR 92a]. The second matrix (noted by $\mathbf{P}_2$) then contains all the known geometric information:

$$\begin{cases} \mathbf{P}_1 = \left[I_{3\times3} \mid 0\right] \\ \mathbf{P}_2 = \left[\mathbf{N}_2 + \mathbf{e}_2\mathbf{b}^T \mid c\mathbf{e}_2\right] \end{cases} \tag{5.5}$$

where $\mathbf{I}_{3\times3}$ is the identity matrix, $\mathbf{e}_2$ is the epipole in the second image, $\mathbf{b}$ is any 3×1 vector and c is a real arbitrary value.

$\mathbf{N}_2$ is a solution of $\mathbf{F}_{12} = \begin{bmatrix} 0 & -e_{2z} & e_{2y} \\ e_{2z} & 0 & -e_{2x} \\ -e_{2y} & e_{2x} & 0 \end{bmatrix} \cdot N_2$.

A specific solution for this system is given by the following formula:

$$N_2 = \begin{bmatrix} 0 & -e_{2z} & e_{2y} \\ e_{2z} & 0 & -e_{2x} \\ -e_{2y} & e_{2x} & 0 \end{bmatrix} \cdot \mathbf{F}_{12}$$

By knowing the fundamental matrix, it is possible to obtain a reconstruction of the scene for a set of projection matrices which are compatible with epipolar geometry [BOB 96, DEV 96].

Since each pair of matrices is derived from the other by a projective transformation, the reconstruction will be called projective. This, in fact, amounts to fixing a 3D projective reference frame associated with the first camera (which can neither be orthogonal nor standardized). The aspect of 3D reconstruction in this reference mark will not correspond to the expected form of the scene. In order to find an aspect which conforms more to reality, it is advisable to change from a projective reconstruction to a Euclidean reconstruction by fixing the projective transformation matrix applied to the 3D points expressed in homogenous coordinates.

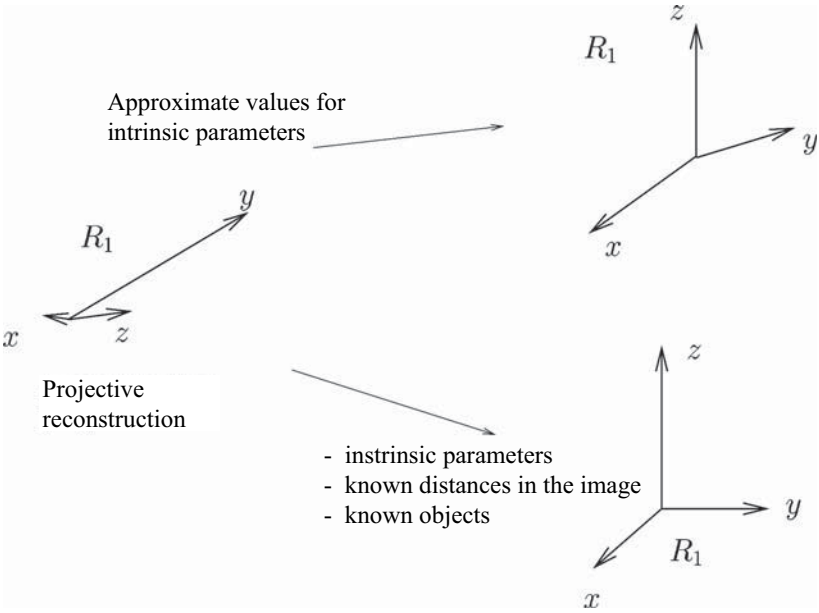


Figure 5.11. *From a projective reference frame to a Euclidean reference frame*

5.4.3. *Euclidean reconstruction*

The Euclidean reconstruction of the scene can be carried out in the following three cases:

– intrinsic and extrinsic parameters are known for the two cameras (calibrated cameras);

– only intrinsic parameters are known for each of the cameras;

– we have the metric data of the scene.

5.4.3.1. *Calibrated cameras*

In this case, the projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are directly calculated from the camera parameters by:

$$\mathbf{P}_1 = A_1\big[R_1 \mid t_1\big]$$

$$\mathbf{P}_2 = A_2\big[R_2 \mid t_2\big]$$

and then we use the retro-projection.

5.4.3.2. *Known intrinsic parameters*

In this case, the intrinsic parameters can be estimated by self-calibration, or we know their rough value. In the latter case, we will talk of a quasi-Euclidean reconstruction.

Hence, it is possible to recover the movement of the camera between the two views (R, t) from the fundamental matrix. This calculation is done by the intermediary of the essential matrix $\mathbf{E}$ obtained by the equation:

$$\mathbf{E} = {\mathbf{A}_2}^T \mathbf{F} \mathbf{A}_1 = \hat{\mathbf{t}}\mathbf{R}.$$

Rotation R and translation t are then obtained by decomposing matrix E. Luong compares the different methods in his thesis [LUO 92]. It is held as the best decomposition making it possible to take into account the constraints related to the nature of the essential matrix proposed by Tsai and Huang in [TSA 84].

Here, the rotation is obtained only up to a Π, while the translation is determined up to a scale factor. However, the latter ambiguity can be overcome from a pair of points in correspondence in the two images. For this, it is sufficient to reconstruct the corresponding 3D point and to modify

the angle of rotation and the sign of translation until the reconstructed point is situated in front of the two cameras. The translation is then known only up to positive scale factor.

5.4.3.3. *Known metric data in the scene*

This method consists of using known metric data in the scene such as points, lines, angles, distances, orthogonal relations, etc. to identify the 3D homography, which makes it possible to shift from a projective reconstruction to a Euclidean reconstruction of the 3D scene. Here, we again refer to the algorithm developed by Boufama [BOU 94]. The hypothesis, *a priori*, is the knowledge of a set of five non-coplanar points of the scene whose relative positions are known. The global diagram of the algorithm is as follows:

1) a sequence of images (at least two) of a rigid scene is taken with a camera;

2) interest points are extracted and tracked in the sequence;

3) 5 known points containing no subset of 4 coplanar points the 5 known coplanar 4×4 points are chosen as basis for a relative reference frame and are given coordinates;

4) the set of observations (image coordinates) is translated into a system of non-linear equations whose resolution directly gives the three-dimensional structure of the scene and the projection matrices.

This will then make it possible to obtain a three-dimensional reconstruction of the scene [BOU 93].

5.4.3.3.1. Notes

– If the five points taken as a basis are not points whose 3D coordinates are known, we find a projective reconstruction relative to the chosen base.

– If the ratios of the distance between the basis points are known, the reconstruction obtained is said to be affine (to a close affine transformation). This has the property to conserve the ratios of the length as well as parallelism.

This algorithm constitutes an example of using *a priori* knowledge on the scene. From a more general point of view, the more the knowledge on the real scene will be Euclidean (direction $<$ distance $<$ 3D coordinates), the more Euclidean the reconstruction will be [BOU 94].

5.5. 3D modeling

At the end of 3D reconstruction, a group of 3D points is available to represent the scene. Each 3D point corresponds to a pixel of the first image, which provides the color of the 3D point. 3D information can be kept in this form or a more easily used representation can be envisaged according to the expected application (visualization, synthesis of views, etc.).

5.5.1. *Implicit model*

The first method consists of keeping 3D information in the form of images and the correspondence information without going up to the stage of 3D reconstruction. This implicit representation of 3D information can be used to generate another view by transfer, for example, using the method of epipolar intersection.

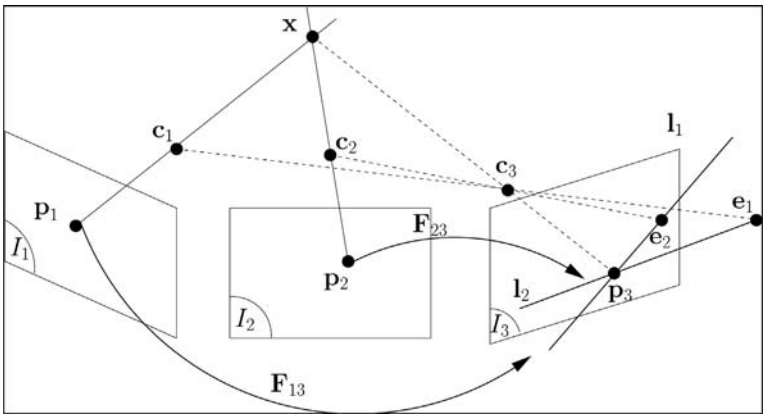


Figure 5.12. *Reconstruction by epipolar intersection*

This method simply comes from the definition of epipolar geometry. It uses epipolar lines in order to reconstruct a pixel by pixel image. This technique introduced in [FAU 94] and referred again in [FAU 98] uses the following idea: by knowing the matching of a pixel $\mathbf{p}_1$ of the first image with a pixel $\mathbf{p}_2$ of the second, the knowledge of the epipolar geometry between the first two images and the virtual image is sufficient to determine the corresponding point in the third view. This point is the intersection in the third image of the epipolar line associated with $\mathbf{p}_1$ with that associated with $\mathbf{p}_2$. The reconstruction of the complete image is then carried out using the set of couples in correspondence in the two image sources.

This technique is degenerated when the 3D point belongs to a trifocal plane (plane formed by three optical centers, since in this case, the two epipolar lines to intersect are identical). Blanc highlights examples of degenerated reconstructions in this specific case [BLA 94]. The use of a trifocal tensor connecting the 3 images makes it possible to avoid degeneration due to the trifocal plane [SHA 95].

Other problems have also been faced when the 3D point to be reconstructed is not the same as in the first two images (occultation). A possible solution is obtained by considering the virtual image. For each pixel $\mathbf{p}_3$ of this image, we look for the couple of points $(\mathbf{p}_1, \mathbf{p}_2)$ of the first two images corresponding to it. For this, we trace the curve of the points of I_2 corresponding to the points of the image I_1 belonging to the epipolar line associated with $\mathbf{p}_3$. Point $\mathbf{p}_2$ is then the first intersection point (point that is closest to the epipole) between the epipolar line associated with $\mathbf{p}_3$ in the image I_2 and this curve (see Figure 5.13).

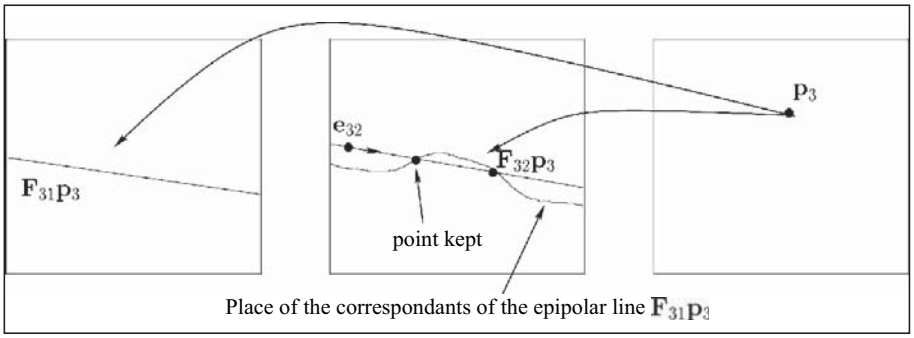


Figure 5.13. *Processing of occultation zones*

The use of transfer methods seems attractive in the case of projective reconstruction since only projective information (fundamental matrices) and mappings are necessary. However, to specify the fundamental matrices associated with the view to reconstruct, we have to either provide the position of the virtual camera expressed in the same projective space as the model, or know the position of the projected points in this image or have points, which are already matched. In the two cases, it is generally necessary to introduce Euclidean information and hence the method loses interest as compared to a Euclidean reconstruction.

5.5.2. *Point sets*

In this approach, we keep the group of reconstructed 3D points. This information is generally represented in the form of a depth chart, i.e., an image containing, in each pixel, its coordinate z expressed in the camera reference frame. Hence, it is also necessary to keep the corresponding image and the absolute position of the camera. The objective is to keep and deteriorate the initial image information to the minimum possible. The disadvantage of this approach is the problems of re-calibration if we wish to project the group of points in another viewpoint. In a *forward* projection where each point of the group is projected on a pixel of a new image, the expansion or occultation zones give rise to "holes", which must be overcome by a process of interpolation. If we want to realize a *backward* projection to avoid this problem, it is a question of finding, for each pixel of the image, its correspondent in the starting image. Hence, it is necessary to determine the intersection of the view line with the surface implicitly defined by the group of points, by a process similar to that seen in the previous section, or to interpolate the group of 3D points in a surface, which amounts to realizing a model with facets.

5.5.3. *Triangular mesh*

Triangular mesh facets constitute a 3D representation very often used as an output of the 3D reconstruction. Indeed, they present several advantages:

– The reconstructed scene is represented in the form of a continuous surface, which projects itself in a continuous image; we thus avoid the unknown sample problems faced during the projection of a group of dots.

– This representation is well adapted to polyhedral scenes, for which a small number of facet planes make it possible to represent the scene.

– Textured triangular meshes constitute a standard representation of synthetic scenes in computer graphics. We can thus benefit from the software and the material available for the visualization.

– In a similar way, 3D triangular meshes are now included in the compression-transmission norms of video data (MPEG4, SNHC); a representation such as this will be efficiently compressed.

Triangular meshes are constituted from a set of 3D triangles forming a continuous surface. They are defined by a set of vertex points and a set of arcs connecting the vertices. In the case of 3D reconstruction, the vertices can be defined in several ways.

5.5.3.1. *Interactive designation of mesh vertices*

The vertices of the mesh are interactively designed. This allows an "intelligent" description of the scene, with a minimum number of triangles and a segmentation of the scene into semantic objects. An interactive definition of the vertices also makes it possible to lighten the mapping phase; only the vertices are tracked and reconstructed. The 3D position of the other points of the scene is given by the planes of triangular facets.

5.5.3.2. *Microfacets*

On the contrary, the representation by microfacets consists of taking the set of constructed 3D group of dots as mesh vertices. Thus, we do not miss out on precision as compared to the group of points, even while having a continuous surface [LEM 97]. On the other hand, the disadvantage of this representation is that it is very voluminous.

5.5.3.3. *Triangulation of the points of interest*

To avoid an interactive procedure, some characteristic points detected and tracked automatically can be used as mesh vertices. However, there is no assurance that the triangular facets thus created correspond effectively to the planar facets of the scene. The 3D surface thus generated does not reflect that of the real 3D scene.

5.5.3.4. *Adaptive triangulation*

To assure the pertinence of triangular facets, an adaptive triangulation determines the mesh vertices in such a way so as to respect a planarity criterion of the points contained in each triangle.

Such a division is obtained by a recursive division: from an initial arbitrary triangulation (division of the image into 2 triangles), we find the planarity of each triangle; the triangles not respecting the planarity criterion are progressively divided to obtain a set of triangles, which correspond effectively to the planar zones in the 3D scene.

The planarity criterion can be constructed either from 3D data obtained from the reconstruction, or from image data. A 3D planarity criterion is based on the distance of the facet points to the plane passing through the 3 vertices. Hence, it is imperative to have a Euclidean reconstruction. We can also propose a planarity criterion based only on image data. Indeed, for a set of coplanar points, their projections in the two images are linked by a 2D

homography. The homography associated with the plane passing through the 3 summits of a facet can be estimated in the following manner.

Let $\mathbf{H}(3\times 3)$ be the homogenous matrix representing this homography. For each of the 3 vertices of the projected facets, $\mathbf{p}_1^i$ and $\mathbf{p}_2^i$, $i = 1\ldots 3$, we have:

$$\mathbf{H}\tilde{\mathbf{p}}_1^i = \lambda\tilde{\mathbf{p}}_2^i \tag{5.6}$$

These points verify the epipolar constraint, from where:

$$\mathbf{p}_1^{iT}\mathbf{H}^T\mathbf{F}\mathbf{p}_1^i = 0$$

By calculating the transpose of the left member and factorizing with the coordinates of the points, we obtain:

$$\mathbf{p}_1^{iT}(\mathbf{F}^T\cdot\mathbf{H} + \mathbf{H}^T\mathbf{F})\mathbf{p}_1^i = 0$$

Thus, any point p_1^i located in the facet plane belongs to the core of $(\mathbf{F}^T.\mathbf{H} + \mathbf{H}^T\mathbf{F})$ irrespective of the point belonging to the plane. Three points of the plane, which are non-linearly dependent, belong to the core of this matrix. As a result, this matrix is zero:

$$\mathbf{F}^T\cdot\mathbf{H} + \mathbf{H}^T\mathbf{F} = 0 \tag{5.7}$$

Owing to the symmetry of the matrix, we obtain 6 homogenous equations. This relation is valid for all the homographies consistent with F. Hence, we have to specify the homography to estimate by introducing some points belonging to the associated plane. For this, Robert *et al*. [ROB 95] showed that if the epipolar constraint and equation (5.7) are verified, equation (5.6) is equivalent to:

$$\left[\mathbf{p}_2, \mathbf{F}\mathbf{p}_1, \mathbf{H}\mathbf{p}_1\right] = 0 \tag{5.8}$$

where $[\mathbf{a},\mathbf{b},\mathbf{c}] = \mathbf{a}^T.\mathbf{b}\times\mathbf{c}$ represents the mixed product.

Each pair of points gives an equation. Hence, 3 pairs of points, which are non-linearly dependent, are necessary to determine a single matrix $\mathbf{H}$. By knowing the epipolar geometry and the disparity field, we can thus calculate the homography associated with the plane defined by any 3 points chosen in image 1.

The planarity criterion then measures the adequacy between the homography $\mathbf{H}$ and the disparity field for the interior points of the triangular facet. The distance criterion for a given triangle T is thus expressed in the form of a Euclidean distance in the image by the expression:

$$D_T = \frac{1}{n} \sum_{\mathbf{p}_1 \in T} \left\| \mathbf{H}\, \mathbf{p}_1 - \mathbf{p}_2 \right\| + \left\| \mathbf{H}^{-1}\, \mathbf{p}_2 - \mathbf{p}_1 \right\| \tag{5.9}$$

where $\| \ \|$ represents the Euclidean norm on the Cartesian coordinates (distance in pixels in the image) and n is the number of pixels of the triangle.

5.5.3.5. *Regular triangulation*

If the compactness of the grid is an important criterion for the expected application, it is interesting to use a triangulation from among the following:

– tips positioned on a regular grid in the image, whose representation is zero in cost (predefined summits and arcs);

– the hierarchical regular grid, with systematic subdivision, easily represented by a tree;

– the Delaunay grid, entirely defined by the position of the tips.

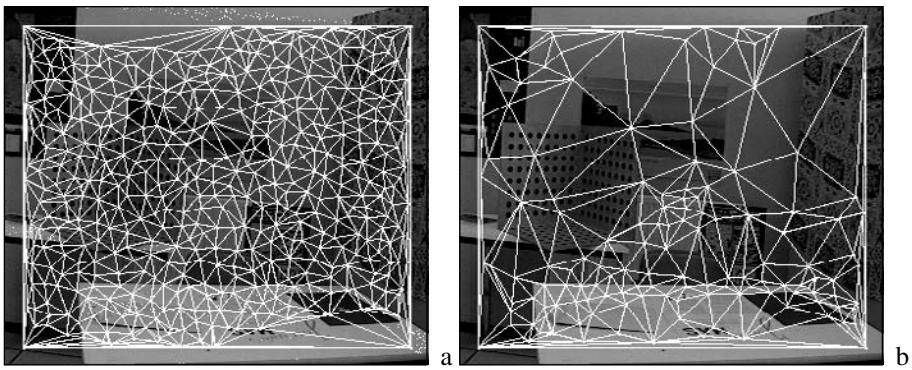


Figure 5.14. *Final triangulation obtained by the estimation of the homography of each triangle (a) and by the estimation of the plane of each facet (b)*

5.6. Examples of applications

5.6.1. *Virtual view rendering*

Virtual views are views of the 3D scene reconstructed from a viewpoint different from those which have been acquired.

This can be realized by transfer from images and depth charts, or by visualization of a reconstructed 3D model.

5.6.2. *VRML models*

We present below an example of a VRML model reconstructed from 2 images of the Armel sequence.

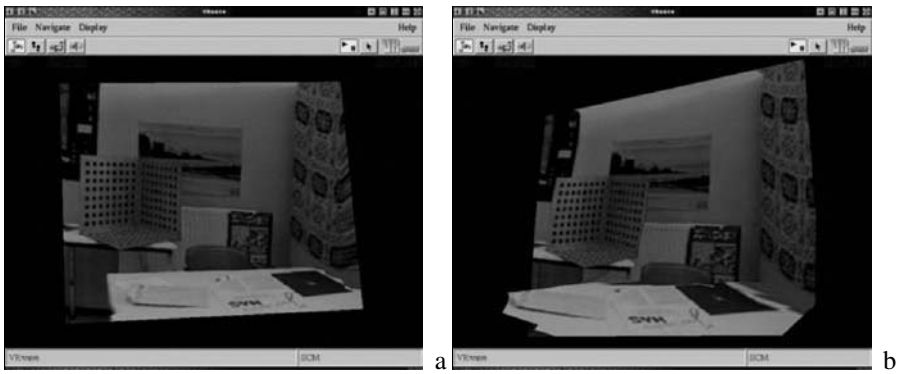


Figure 5.15. *Final VRML model (by a 3D cutting criterion – close to the original viewpoint (a) and relatively distant (b))*

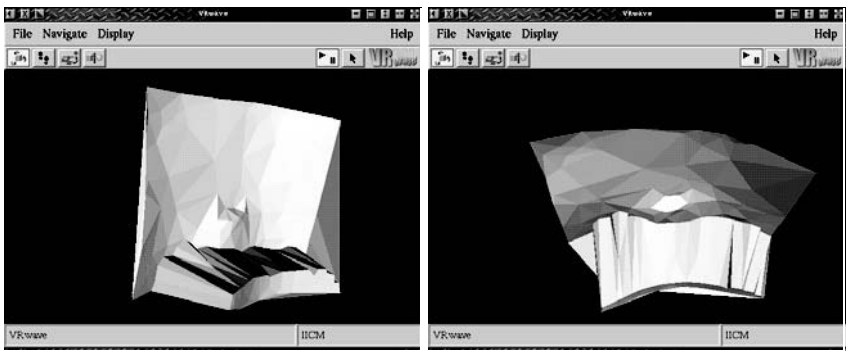


Figure 5.16. *Final VRML model – general idea of the geometric model*

5.7. Conclusion

In this chapter, we have presented the basic techniques to resolve the problem of reconstruction of a fixed voluminous 3D scene from several points of view. At first, we presented the geometric notions related to this

observation mode: homologous points, epipolar geometry, fundamental matrix, essential matrix, trifocal tensor. Then we presented the practical resolution of the problem of dense matching between stereoscopic image couples.

5.8. Bibliography

[ASC 92] ASCHWANDEN P. and GUGGENBÜHL W., "Experimental Results from a Comparative Study on Correlation-Type Registration Algorithms", *ISPRS Workshop*, Bonn, Germany, 1992.

[BEA 94] BEARDSLEY P., ZISSERMAN A. and MURRAY D., "Navigation using Affine Structure from Motion", in EKLUNDH J.-O. (Ed.), *Proceedings of the 3rd European Conference on Computer Vision*, vol. 2 of *Lecture Notes in Computer Science*, Stockholm, Sweden, Springer-Verlag, p. 85–96, May 1994.

[BLA 94] BLANC J., 3-D Reconstruction for Image Synthesis, Master's Thesis, University of Joseph Fourier, Grenoble, July 1994, in French.

[BLA 98] BLANC J., Synthèse de nouvelles vues d'une scène 3D à partir d'images existantes, PhD Thesis, Institut National Polytechnique, Grenoble, 1998.

[BOB 96] BOBET P., BLANC J. and MOHR R., "Aspects cachés de la tri-linéarité", *Reconnaissance des Formes et Intelligence Artificielle*, p. 137–146, January 1996.

[BOU 93] BOUFAMA B., MOHR R. and VEILLON F., Euclidean Constraints for Uncalibrated Reconstruction, Report no. RT96 IMAG 17 LIFIA, LIFIA, INSTITUT IMAG, March 1993.

[BOU 94] BOUFAMA B., Reconstruction tridimensionnelle en vision par ordinateur: cas des caméras non étalonnées, PhD Thesis, Institut National Polytechnique, Grenoble, 1994.

[CAR 95] CARLSSON S., "Duality of the reconstruction and positioning from projective views", *Workshop on Representation of Visual Scenes*, Cambridge, Massachusetts, USA, p. 85–92, June 1995.

[CAR 98] CARLSSON S. and WEINSHALL D., "Dual computation of projective shape and camera positions from multiple images", *The International Journal of Computer Vision*, vol. 27, no. 3, p. 227–241, 1998.

[COX 92] COX I.J., HINGORANI S.N.B., MAGGS B.M. and RAO S.B., "Stereo without disparity gradient smoothing: a bayesian sensor fusion solution", *Proceedings of British Machine Vision Conference*, p. 337–346, 1992.

[DEM 96] DEMARTY J. and SCHMITT F., "Reconstruction 3D dense à partir de séquences d'images", *Journées Orasis*, Clermont-Ferrand, GDR-PRC CHM, p. 159–163, May 1996.

[DEV 96] DEVERNAY F. and FAUGERAS O., "From Projective to Euclidean Reconstruction", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, IEEE, p. 264–269, June 1996.

[DHO 89] DHOND U.R. and AGGARWAL J., "Structure from Stereo – A Review", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 6, p. 1489–1510, 1989.

[FAU 90] FAUGERAS O. and MAYBANK S., "Motion from point matches: multiplicity of solutions", *The International Journal of Computer Vision*, vol. 4, no. 3, p. 225–246, 1990.

[FAU 92a] FAUGERAS O., "What can be seen in three dimensions with an uncalibrated stereo rig?", *Proceedings of the 2nd ECCV*, p. 563–578, May 1992.

[FAU 92b] FAUGERAS O., LUONG T. and MAYBANK S., "Camera self-calibration: theory and experiments", Sandini [SAN 92], p. 321–334, May 1992.

[FAU 93] FAUGERAS O., *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, 1993.

[FAU 94] FAUGERAS O. and LAVEAU S., "Representing Three-Dimensional Data as a Collection of Images and Fundamental Matrices for Image Synthesis", *Proceedings of the International Conference on Pattern Recognition*, Jerusalem, Israel, Computer Society Press, p. 689–691, October 1994.

[FAU 95a] FAUGERAS O., "Stratification of 3-D vision: projective, affine, and metric representations", *Journal of the Optical Society of America A*, vol. 12, no. 3, p. 465–484, March 1995.

[FAU 95b] FAUGERAS O. and MOURRAIN B., "On the geometry and algebra of the point and line correspondences between n images", *Proceedings of the 5th International Conference on Computer Vision*, Boston, MA, p. 951–956, IEEE Computer Society Press, June 1995.

[FAU 98] FAUGERAS O., "De la géométrie au calcul variationnel: théorie et applications de la vision tridimensionnelle", *11ème Congrès RFIA'98*, January 1998.

[FRA 91] FRANÇOIS E., Interprétation qualitative du mouvement à partir d'une séquence d'images, PhD Thesis, University of Rennes I, 1991.

[HAR 92a] HARTLEY R.I., "Estimation of Relative Camera Positions for Uncalibrated Cameras", Sandini [SAN 92], p. 579–587, May 1992.

[HAR 92b] HARTLEY R., GUPTA R. and CHANG T., "Stereo from Uncalibrated Cameras", *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Urbana Champaign, IL, IEEE, p. 761–764, June 1992.

[HAR 95] HARTLEY R., "In defence of the 8-point algorithm", *Proceedings of the 5th International Conference on Computer Vision*, Boston, MA, p. 1064–1070, IEEE Computer Society Press, June 1995.

[HAR 97] HARTLEY R.I., "Lines and points in three views and the trifocal tensor", *The International Journal of Computer Vision*, vol. 22, no. 2, p. 125–140, March 1997.

[HUA 89] HUANG T.S. and FAUGERAS O.D., "Some Properties of the E Matrix in Two-View Motion Estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, p. 1310–1312, December 1989.

[HUB 81] HUBER P., *Robust Statistics*, John Wiley & Sons, New York, 1981.

[LEM 97] LE MESTRE G., Analyse de séquences d'images pour la reconstruction de vues intermédiaires, PhD Thesis, University of Rennes I, 1997.

[LON 81] LONGUET-HIGGINS H., "A Computer Algorithm for Reconstructing a Scene from Two Projections", *Nature*, vol. 293, p. 133–135, 1981.

[LUO 92] LUONG Q.-T., Matrice Fondamentale et Calibration Visuelle sur l'Environnement-Vers une plus grande autonomie des systèmes robotiques, PhD Thesis, Paris-Sud University, Center d'Orsay, December 1992.

[LUO 93] LUONG Q.-T., *Handbook of Pattern Recognition and Computer Vision*, Chapter "Color vision", World scientific, 1993.

[MEM 96a] MEMIN E. and PEREZ P., "Robust Discontinuity-Preserving Model For Estimating Optical Flow", *ICPR*, vol. A, p. 920–924, 1996.

[MÉM 96b] MEMIN E., PÉREZ P. and MACHECOURT D., Dense estimation and object-oriented segmentation of the optical flow with robust techniques, Report no. 991, IRISA, March 1996.

[MOH 92] MOHR R., QUAN L., VEILLON F. and BOUFAMA B., Relative 3D reconstruction using multiple uncalibrated images, Report no. RT84-IMAG12, LIFIA, June 1992.

[OUA 96] OUALI M., LANGE H. and LAURGEAU C., "An Energy Minimization Approach to Dense Stereovision", *International Conference on Image Processing*, p. 841–846, September 1996.

[PÉR 98] PÉREZ P., Markov random field and images, Research Report no. 1196, IRISA Rennes, July 1998.

[QUA 95] QUAN L., "Invariant of a pair of non-coplanar conics in space", in MOHR R. and WU C. (Eds.), *Proceedings of Europe-China Workshop on Geometric Modeling and Invariants for Computer Vision*, Xi'an, China, Xidian University Press, p. 190–197, April 1995.

[ROB 95] ROBERT L. and FAUGERAS O., "Relative 3-D Positioning and 3-D Convex Hull Computation From A Weakly Calibrated Stereo Pair", *Image and Vision Computing*, vol. 13, no. 3, p. 189–197, 1995 and INRIA Technical Report 2349.

[ROB 96] ROBERT L. and DERICHE R., "Dense Depth Map Reconstruction: A Minimization and Regularization Approach which Preserves Discontinuities", in BUXTON B. (Ed.), *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, April 1996.

[SAN 92] SANDINI G. (Ed.), *EECV '92, Second European Conference on Computer Vision*, Santa Margherita, Italy, Springer-Verlag, May 1992.

[SHA 95] SHASHUA A., "Algebraic functions for recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, p. 779–789, 1995.

[SPE 90] SPETSAKIS M. E. and ALOIMONOS J., "Structure from Motion Using Line Correspondences", *The International Journal of Computer Vision*, vol. 4, p. 171–183, 1990.

[STU 69] STURM R., "Das Problem der Projektivität und seine Anwendung auf die Flächen zweiten Grades", *Math. Ann.*, vol. 1, p. 533–574, 1869.

[TAM 92] TAMTAOUI A., Coopération stéréovision-mouvement en vue de la compression de séquences d'images stéréoscopiques. Application à la télévision en relief (TV3D), PhD Thesis, University of Rennes I, 1992.

[TOR 97a] TORR P.H.S. and MURRAY D.W., "The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix", *IJCV*, vol. 24, no. 3, p. 271–300, 1997.

[TOR 97b] TORR P. and ZISSERMAN A., "Robust parametrization and computation of the trifocal tensor", *Image and Vision Computing*, vol. 15, p. 591–605, 1997.

[TRI 95] TRIGGS B., "Matching Constraints and the Joint Image", *Proceedings of the 5th International Conference on Computer Vision*, Boston, MA, p. 338–343, IEEE Computer Society Press, June 1995.

[TSA 84] TSAI R. and HUANG T., "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, p. 13–26, January 1984.

[TZI 94] TZIRITAS G. and LABIT C., *Motion Analysis for Image Sequence Coding*, Elsevier, Paris, p. 366–384, 1994.

[WEN 92] WENG J., AHUJA N. and HUANG T., "Matching Two Perspective Views", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, p. 806–825, 1992.

[ZHA 93] ZHANG Z., Le problème de la mise en correspondance: l'état de l'art, Research report no. 2146, INRIA, December 1993.

[ZHA 94] ZHANG Z., DERICHE R., FAUGERAS O. and LUONG Q.-T., "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry", *Artificial Intelligence Journal*, vol. 78, no. 1–2, p. 87–119, 1994, Appeared in October 1995 and INRIA Research Report No. 2273, May 1994.

Chapter 6

3D Reconstruction by Active Dynamic Vision

6.1. Introduction: active vision

Techniques of active vision, which were created about 15 years ago, draw their origin from a simulation attempt of the animal and human visual system trying to recreate its adaptabilities. From a methodological point of view, active vision, where perceived information is used within a feedback loop, tries to improve the quality of perception when compared to the passive traditional approach, where we restrict ourselves to observing, measuring and interpreting data from the sensor. In fact, active vision consists of elaborating the intelligent perception strategies, by controlling the parameters of the sensor (position, speed, development, etc.). It can be defined as an intelligent acquisition process of data in order to solve the problems raised while designing a system of computer vision, i.e. their sensitivity to noise, their weak precision and particularly their lack of reactivity.

If the general definition of active vision is simple, the motivations of various researchers who took part in the introduction of these new techniques are, however, very different. In fact, what we traditionally call active vision can be divided into several sub-classes[1].

Chapter written by Éric MARCHAND and François CHAUMETTE.

1. The terms "active vision", "active perception", "animate vision" and "purposive vision", which were retained in this typology draw their origin from the titles of the first articles published in each of these fields (respectively known as *"Active Vision"* [ALO 87], *"Active Perception"* [BAJ 88], *"Animate Vision"* [BAL 91] and *"Purposive and qualitative active vision"* [ALO 90]).

– *active vision* as explained by Aloimonos [ALO 87] is a theoretical analysis of the vision process whose principal motivation is the optimization of the visual task with the help of an active sensor;

– *active perception* is a concept introduced by Bajcsy [BAJ 88] attempting to elaborate the perception strategies in order to improve the knowledge of environment;

– *animate vision* [BAL 91] is based on the analysis of human perception. Depending on the use of active binocular heads, it supports the aspects of fixing and focusing on areas of *gaze control*. Ballard's aim is to decrease the algorithmic complexity of the perception process;

– *purposive vision* [ALO 90] aims to extract only relevant information according to a given task.

As we can see, Aloimonos and Bajcsy (considering only the principal authors) have radically different objectives. One seeks to simplify the fundamental problems of vision by using an active observer and the other seeks to improve knowledge on the environment observed by defining adequate perception strategies. Though these two approaches have the use of one or several mobile sensors as a common point, the motivations, methods and tools used are very different.

These various aspects are, however, strongly interdependent. Thus, within the framework of the problem of reconstruction of scenes, various aspects of active vision can be implemented to carry out this task efficiently. More precisely, these aspects appear in the two levels that we consider in this chapter, i.e., a local level dedicated to 3D reconstruction of a particular primitive, and a global level dedicated to complete reconstruction of the scene:

– at local level, the movements of the camera are constrained in such a manner as to optimize the quality of the reconstruction results of particular geometric primitives. In this level, we place ourselves in Aloimonos' vision. Moreover, the constraints necessary for a reliable, non-biased estimate of the parameters of 3D primitives are, as we will see thereafter, the constraints on the position of a primitive in the image and on the trajectory of the camera as compared to this primitive. Automatic generation of these movements is done using visual servoing. This aspect is similar to the concept of purposive vision since only necessary information to the estimate of parameters of this primitive is acquired and used. Moreover, the command laws used consider *only* camera movements which make it possible to acquire 3D information. Thus, any movement that does not bring information is not carried out;

– the use of active vision at a local level is quite constraining since the camera movements are strongly constrained. To mitigate this problem, a second and more global level, based on the development of the *perception strategies* was proposed: movements of the camera are then controlled in a manner which enables the reconstruction of scenes made up of several unknown objects. Here, it is a question of developing strategies of camera movement making it possible to carry out complete exploration and focusing on particular zones of the scene. Accordingly, we instead join the aspect of "active perception" by Bajcsy and/or the purposive vision aspect where the explicit intention is the reconstruction and 3D exploration.

As we can note, dissociating various aspects of active vision is neither always possible nor desirable. Active vision thus consists of defining the set of *perception strategies* making it possible to ensure the realization of a given task. This relates to low level strategies (*local and continuous*), which make it possible to achieve stable and robust methods of data acquisition, as well as high level strategies (*global and event-driven*), which aim to define the set of necessary actions in the realization of the nominal task. These two aspects are strongly interdependent. The high level strategies must depend on reliable bases and use data as precise as possible. On the other hand, data resulting from low level strategies are not directly usable because they give only a partial and incomplete vision of the environment. Thus, they are only of interest if we consider a vast context, which is given by the strategy aspect of high level perception.

In this chapter, we will first present a reconstruction method by dynamic vision and then we will proceed to show that a process of active vision can largely improve the quality of reconstruction. Finally, we will describe the essential stages in the complete reconstruction of an unknown environment.

6.2. Reconstruction of 3D primitives

6.2.1. *Reconstruction by dynamic vision: a rapid state of the art*

The objective of reconstruction techniques by dynamic vision is to carry out the 3D reconstruction of a scene from a sequence of images acquired by a camera. We can classify the approaches treating this problem into two principal categories: discrete techniques and continuous techniques:

– Discrete techniques are based on the displacement measurements of the camera and the displacement corresponding to particular primitives in images [AGG 87, CHI 89, VIA 92, CRO 92, WEN 92, WEL 89, ZHA 94, ZHA 95].

These techniques depend on three fundamental stages: extraction of a set of sufficiently relevant 2D primitives in the sequence of images, matching of the selected primitives of image with image and reconstruction by triangulation.

– Continuous approaches are based on a formulation in terms of speed [ADI 85, ADI 89, VER 90, WAX 87, ESP 87, BOU 93, NEG 87, XIE 89a]. The images are then acquired with a pace closer to video pace. From an image processing point of view, this approach depends either on an estimate of the apparent velocity field (optical flow) – which generally proves to be not very reliable and unstable, particularly around the occluding contours [ADI 89, WAX 87] – or on the sequence of objects of *interest* (*token trackers*) [ESP 87, XIE 89b, BOU 93]. These approaches are generally very sensitive to noise because of the weak movement that they impose between two successive images. However, we will see that these defects can be removed using active vision.

In this chapter, only the continuous techniques are considered. Discrete techniques depend on the same diagram as that of stereovision. Since these methods are described in Chapter 5, we will not discuss them again.

Similar to the problem of matching the primitives between two images, the determination of the apparent velocity field is a difficult problem. Indeed, noise present in the images leads to errors and instability in the estimate of velocity fields, which imply errors, sometimes major ones, in the estimate of the 3D structure of the scene. To carry out an estimate of the structure of a scene, the movement of the camera must be known or estimated. If this information is not available or not measurable, it is first necessary to proceed with an estimation stage of this movement [AGG 88, SUB 87]. Among the studies carried out in this vision, we find the works of [NEG 87, ADI 85].

Research by Negahdaripour and Horn [NEG 87] aims to determine, in addition to camera movement, the orientation of a plane surface using not the apparent velocity field, but the space and temporal derivatives of light intensity f (i.e., $\nabla f = (\frac{\partial f}{\partial X}, \frac{\partial f}{\partial Y})^T$ and $\frac{\partial f}{\partial t}$). The resolution of a non-linear system is necessary to obtain the solution of the problem. Adiv [ADI 85] considered it to be a non-static environment, since it takes into account the potential presence of several mobile objects. The apparent velocity field is initially segmented in elementary areas in which the movement is coherent with that of a plane surface. The movement parameters of each of these areas are calculated and then, if required, the areas having close movements are joined. Each of these areas is supposed to correspond to the same rigid object.

Since the segmentation of the velocity field is realized, the relative positions between each object and the camera can be determined.

These methods consider that the movement of the camera is completely unknown. However, if we place ourselves in a robotic context, this movement is, if not controllable (or controlled), at least measurable with some precision. By considering particular movements of the camera, Vernon and Tistarelli [VER 90] strongly simplify the calculation of the apparent movement. Using the velocity field thus calculated and the movement parameters of the camera, a depth chart is then made.

Another approach was developed by Arbogast [ARB 91] in order to reconstruct non-polyhedric objects. This is based on the concept of spatio-temporal surface: it is about the surface described by a set of contours, which move on the image plane in the course of time. The differential properties of this spatio-temporal surface are analyzed, after having chosen an adequate parametrization of this surface. For this, it uses a spherical modeling of the camera by reconsidering Blake and Cipolla's modeling [BLA 90]. This leads to a very flexible parametrization and allows us to simplify the equations obtained. Although the method is attractive from the theoretical point of view, it presents the disadvantage of being very sensitive to noise because of its differential aspect. Thus, the precision required on the motion parameters is significant. In addition, this method does not allow us to reconstruct surfaces that do not present any apparent contour (plane or concave surfaces).

Finally, like discrete approaches, a class of method uses mapping of particular primitives (*token trackers*). The problematic determination of optical flow is thus not achieved and, as the continuous approaches imply weak displacements of the camera, tracking is simplified and brings back the problem of 2D sequence of primitives. Three similar methods were proposed in this direction [RIV 87, ESP 87, XIE 89b, BOU 93]. Espiau and Rives [RIV 87, ESP 87] considered the case of point type primitives. The instantaneous measurement of 2D movement of a point in the image sequence is used to refine the estimate of its 3D parameters at the time of sensor displacement in the scene using a Kalman filter type recursive filtering. Xie [XIE 89a, XIE 89b] extended this method to segments. Finally, this approach is generalized to any type of parameterable geometric primitive in [BOU 93, CHA 96]. This method will be described now.

6.2.2. *General principle*

The camera is modeled in a traditional way by a perspective projection. Without loss of generality, the focal distance of the camera is fixed to be equal to 1. A point $\mathbf{x} = (x, y, z)$ coordinates in the camera frame is projected in $\mathbf{X} = (X, Y, 1)$ coordinates with:

$$\mathbf{X} = \frac{1}{z}\mathbf{x}. \tag{6.1}$$

Let $\mathcal{P}_s$ be a parameterable geometric primitive described by an equation of the form:

$$h(\mathbf{x}, \mathbf{p}) = 0, \quad \forall \mathbf{x} \in \mathcal{P}_s \tag{6.2}$$

where h defines the nature of the primitive (point, line, sphere, etc.) and $\mathbf{p}$ defines its parameter vector. The objective of the reconstruction is to estimate the value of parameters $\mathbf{p}$ in order to reconstruct and localize primitive $\mathcal{P}_s$ defined by h. Let $\mathcal{P}_i$ be the projection in the image of $\mathcal{P}_s$. Primitive $\mathcal{P}_i$ can be written as:

$$g(\mathbf{X}, \mathbf{P}) = 0, \quad \forall \mathbf{X} \in \mathcal{P}_i \tag{6.3}$$

where g defines the primitive nature and where the value of parameters $\mathbf{P}$, function of $\mathbf{p}$, describes its configuration in the image.

In addition, using perspective projection equations (6.1), equation (6.2) becomes:

$$h'(\mathbf{X}, 1/z, \mathbf{p}) = 0. \tag{6.4}$$

If we exclude degenerated cases (such as $\frac{\partial h'}{\partial z} = 0$, which occur, for example, when a circle is projected in the image in the form of a segment instead of an ellipse), the implicit functions theorem ensures the existence of a function μ such that:

$$1/z = \mu(\mathbf{X}, \mathbf{p}_L) \tag{6.5}$$

where $\mathbf{p}_L$ is function of $\mathbf{p}$.

For plane primitives (circle, etc.), function μ represents the plane in which the primitive is placed. In a more general case of voluminous primitives (cylinder (see Figure 6.1), sphere, torus, etc.), function $g(\mathbf{X}, \mathbf{P})$ represents the

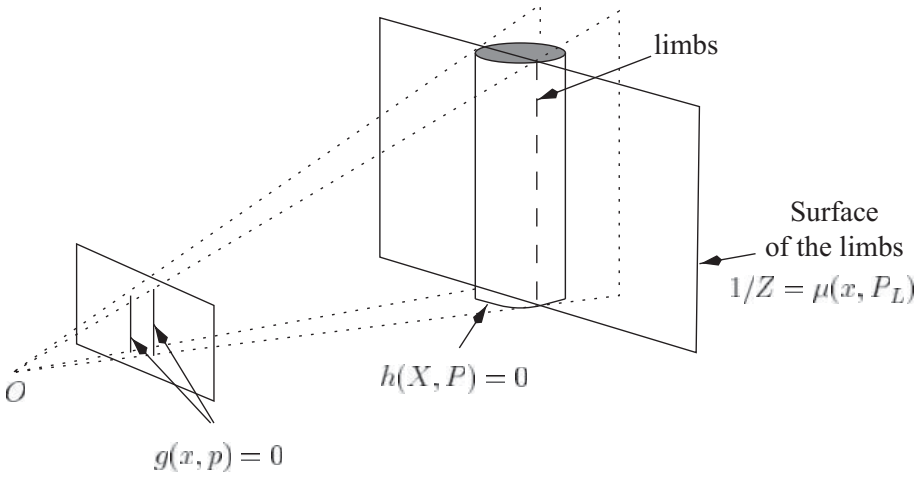


Figure 6.1. *Projection of primitive* (h) *in the image* (g) *and surface of the limbs* (μ) *in the case of a cylinder*

projection in the image of limbs of the primitives and function μ, therefore called the *limb surface*, expresses the relation between the points of $\mathcal{P}_i$ and the corresponding points of $\mathcal{P}_s$. Parameters $\mathbf{p}_L$, depending only on $\mathbf{p}$, describe the configuration of this surface in the pointer of the camera.

Let $\mathbf{T_c} = (\mathbf{V}, \mathbf{\Omega})$ be the kinematic torque of the camera where $\mathbf{V} = (V_x, V_y, V_z)$ represents the translation speed of the camera and $\mathbf{\Omega} = (\Omega_x, \Omega_y, \Omega_z)$ represents its rotation speed. The variation of $\mathbf{p}$, which connects the projected primitive movement in the image to the camera movement $\mathbf{T_c}$ can be explicitly calculated and expressed by:

$$\dot{\mathbf{P}} = \mathbf{L}_{\mathbf{P}}^{\mathbf{T}}(\mathbf{P}, \mathbf{p_L})\, \mathbf{T_c} \tag{6.6}$$

where $\mathbf{L}_{\mathbf{P}}^{T}(\mathbf{P}, \mathbf{p_L})$, known as the interaction matrix associated with $\mathbf{P}$, characterizes the interactions between the sensor and the primitive considered [CHA 90, ESP 92].

The estimation of parameters $\mathbf{p}$ is carried out in two stages [BOU 93, CHA 96]: first, parameters $\mathbf{p_L}$ characterizing the limb surface μ are obtained from equation (6.6) using the measurement of $\mathbf{T_c}, \mathbf{P}$ and $\dot{\mathbf{P}}$:

$$\mathbf{p_L} = \mathbf{p_L}(\mathbf{T_c}, \mathbf{P}, \dot{\mathbf{P}}) \tag{6.7}$$

With the help of geometric constraints on the type of primitive to be reconstructed (line, cylinder, etc.), which are characterized by the equation $h(\mathbf{x}, \mathbf{p}) = \mathbf{0}$, it is then possible to show, by intersection of the limb surface with the center cone, the optical center of camera O and the generator, $g(\mathbf{X}, \mathbf{P}) = 0$, with $\mathbf{p}$ parameters which characterize the studied primitive:

$$\mathbf{p} = \mathbf{p}\big(\mathbf{p_L}, \mathbf{P}\big). \tag{6.8}$$

6.2.3. *Some specific cases*

6.2.3.1. *Point*

The estimation of position (x_0, y_0, z_0) of a point can indeed be obtained from this approach. In this simple case, the equation h of the primitive is given by:

$$h(\mathbf{x}, \mathbf{p}) = \begin{cases} x - x_0 = 0 \\ y - y_0 = 0 \\ z - z_0 = 0 \end{cases} \tag{6.9}$$

Projection of the point in the image allows us to write the function g:

$$g(\mathbf{X}, \mathbf{P}) = \begin{cases} X - X_0 = 0 \\ Y - Y_0 = 0 \end{cases} \quad \text{with} \quad \begin{cases} X_0 = \dfrac{x_0}{z_0} \\ Y_0 = \dfrac{y_0}{z_0} \end{cases} \tag{6.10}$$

Finally, surface μ is defined by:

$$1/z = \mu\big(\mathbf{X}, \mathbf{p_L}\big) = 1/z_0. \tag{6.11}$$

Depth z_0 is obtained by way of equation (6.6) for point (see Appendix A):

$$\begin{pmatrix} \dot{X}_0 \\ \dot{Y}_0 \end{pmatrix} = \begin{pmatrix} -1/z_0 & 0 & X_0/z_0 & X_0 Y_0 & -(1 + X_0^2) & Y_0 \\ 0 & -1/z_0 & Y_0/z_0 & 1 + Y_0^2 & -X_0 Y_0 & -X_0 \end{pmatrix} \mathbf{T_c} \tag{6.12}$$

which can be written in the form:

$$A \frac{1}{z_0} = B \tag{6.13}$$

with:

$$A = \begin{pmatrix} V_x - X_0 V_z \\ V_y - Y_0 V_z \end{pmatrix}$$

and:

$$B = \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \begin{pmatrix} X_0 Y_0 \Omega_x - \big(1 + X_0\big)^2 \Omega_y + Y_0 \Omega_z - \dot{X}_0 \\ \big(1 + Y_0\big)^2 \Omega_x - X_0 Y_0 \Omega_y - X_0 \Omega_z - \dot{Y}_0 \end{pmatrix}$$

The solution of the system is then given by:

$$1/z_0 = \frac{\alpha_x \big(X_0 V_z - V_x\big) + \alpha_y \big(Y_0 V_z - V_y\big)}{\big(X_0 V_z - V_x\big)^2 + \big(Y_0 V_z - V_y\big)^2} \tag{6.14}$$

Parameters x_0 and y_0 are then easily determined by:

$$x_0 = X_0 z_0, \quad y_0 = Y_0 z_0$$

It is to be noted that a movement in the direction of point $(V_x = X_0 V_z,\ V_y = Y_0 V_z)$ does not make it possible to obtain its 3D position.

6.2.3.2. *Line*

A line can be represented by the intersection of two orthogonal planes:

$$h\big(\mathbf{x}, \mathbf{p}\big) = \begin{cases} A_1 x + B_1 y + C_1 z = 0 \\ A_2 x + B_2 y + C_2 z + D_2 = 0 \end{cases}$$
$$\text{with} \begin{cases} A_1^2 + B_1^2 + C_1^2 = 1 \\ A_2^2 + B_2^2 + C_2^2 = 1 \\ A_1 A_2 + B_1 B_2 + C_1 C_2 = 0. \end{cases} \tag{6.15}$$

A minimal and complete representation of a corresponding 2D line is given by:

$$g(\mathbf{X}, \mathbf{P}) = X \cos\theta + Y \sin\theta - \rho = 0$$
$$\text{with} \begin{cases} \cos\theta &= A_1 / \sqrt{A_1^2 + B_1^2} \\ \sin\theta &= B_1 / \sqrt{A_1^2 + B_1^2} \\ \rho &= -C_1 / \sqrt{A_1^2 + B_1^2}. \end{cases} \tag{6.16}$$

Moreover, function μ is easily obtained from (6.15):

$$1/z = \mu\big(\mathbf{X}, \mathbf{p_L}\big) = AX + BY + C \text{ with } \begin{cases} A = -A_2/D_2 \\ B = -B_2/D_2 \\ C = -C_2/D_2 \end{cases} \tag{6.17}$$

Finally, the relation between the movement of the line in the image (defined by $(\dot{\rho}, \dot{\theta})$) and the speed of camera $\mathbf{T_c}$ is given by the interaction matrix associated with parameters (ρ, θ) [ESP 92]:

$$\begin{pmatrix} \dot{\rho} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \lambda_\rho \cos\theta & \lambda_\rho \sin\theta & -\lambda_\rho \rho & \big(1+\rho^2\big)\sin\theta & -\big(1+\rho^2\big)\cos\theta & 0 \\ \lambda_\theta \cos\theta & \lambda_\theta \sin\theta & -\lambda_\theta \rho & -\rho\cos\theta & -\rho\sin\theta & -1 \end{pmatrix} T_c \tag{6.18}$$

with $\lambda_\rho = -A\rho\cos\theta - B\rho\sin\theta - C$, and $\lambda_\theta = B\cos\theta - A\sin\theta$.

From the values measured from $\rho, \theta, \dot{\rho}, \dot{\theta}$ and T_c, it is now necessary to estimate the parameters of the two planes, which define the line considered.

Parameters A_1, B_1 and C_1 are immediately deduced from ρ and θ. Then, as explained earlier, the parameters describing the function μ are determined by using the measurement of the camera speed and the apparent velocity of the line in the image which results from it. Thus, parameters A, B, C are given by the resolution of the following linear system:

$$\begin{cases} -A\rho\cos\theta - B\rho\sin\theta - C = \lambda_\rho \\ -A\sin\theta + B\cos\theta = \lambda_\theta \\ A\cos\theta + B\sin\theta - C\rho = 0 \end{cases} \tag{6.19}$$

where λ_ρ and λ_θ are obtained from (6.18):

$$\begin{cases} \lambda_\rho = \dfrac{\dot{\rho} + \big(1+\rho^2\big)\big(\Omega_Y\cos\theta - \Omega_X\sin\theta\big)}{V_X\cos\theta + V_Y\sin\theta - \rho V_Z} \\ \lambda_\theta = \dfrac{\dot{\theta} + \rho\big(\Omega_X\cos\theta + \Omega_Y\sin\theta\big) + \Omega_Z}{V_X\cos\theta + V_Y\sin\theta - \rho V_Z}. \end{cases} \tag{6.20}$$

Finally, we obtain $D_2 = 1/\sqrt{A^2 + B^2 + C^2}$, $A_2 = -AD_2$, $B_2 = -BD_2$, and $C_2 = -CD_2$.

6.2.3.3. *Cylinder*

A cylinder is characterized by the following equation:

$$\begin{aligned} h(\mathbf{x}, \mathbf{p}) = &\left(x - x_0\right)^2 + \left(y - y_0\right)^2 + \left(z - z_0\right)^2 \\ &- \left(ux + vy + wz\right)^2 - r^2 = 0 \end{aligned} \tag{6.21}$$

where r is the radius of the cylinder, (u, v, w) represents the directing vector of the axis of the cylinder and (x_0, y_0, z_0) are the coordinates of the point of the axis of the cylinder nearest to the optical center O of the camera. The function relative to the limb surface μ is given by:

$$\begin{aligned} 1/z = \mu(\mathbf{x}, \mathbf{P_L}) = AX + BY + C \\ \text{with} \begin{cases} A = x_0/\left(x_0^2 + y_0^2 + z_0^2 - r^2\right) \\ B = y_0/\left(x_0^2 + y_0^2 + z_0^2 - r^2\right) \\ C = z_0/\left(x_0^2 + y_0^2 + z_0^2 - r^2\right) \end{cases} \end{aligned} \tag{6.22}$$

where A, B, and C are the normal vector components in the limb surface of the cylinder. The image of a cylinder is made up of two lines whose parameters ρ and θ can be expressed according to $\mathbf{P}$ parameters. The interaction matrix associated with each line is given by equation (6.18), the only difference as compared to the previous case being the value of $\mathbf{p_L}$ parameters intervening in λ_ρ and λ_θ. Parameters $\mathbf{p_L}$ are estimated by the resolution of the linear system constructed from the interaction matrix associated with the two lines. It is then possible to achieve parameters $\mathbf{p}$. Let us note that the cylinder can also be reconstructed using the projection of only one of its limbs [BOU 93].

Similar results can be obtained for other parameterable geometric primitives. Circles and spheres, for example, are described in [BOU 93].

6.2.4. *3D reconstruction by active vision*

The previously described method proves ineffective to precisely reconstruct a geometric primitive. From its continuous aspect, it is confronted with discretization errors and is, moreover, very sensitive to measurement noise. Traditionally, in order to try to solve these problems and increase the robustness of solutions, the processes of recursive filtering [MAT 89, BOU 89, VIA 92, CRO 92] or the non-linear optimization processes [WEN 92] are implemented. Another way of solving this problem is to use the paradigm of the active vision [ALO 90, BAJ 88, SWA 93].

6.2.4.1. *3D reconstruction by active vision: state of the art*

Contrary to dynamic vision where it is only necessary to observe camera movement, in the case of active vision, this movement is controlled. Although works relating to active vision have multiplied during the past few years, few researchers tried to know with regard to the problem of reconstruction from movement whether there were trajectories of the observer, which are more beneficial than others. However, this aspect is covered in the following chapter.

Sandini and Tistarelli proposed two approaches, one relying on a co-operation between stereovision and movement [SAN 86] and the other based on the use of a moving monocular system [SAN 90]. The first approach, aiming at a robust estimate of a depth chart [SAN 86], uses a moving stereoscopic sensor. The objective is to define displacement strategies of the sensor, which enable the stereoscopic system to acquire reliable information. Moreover, the stereovision must help the acquisition of 3D information obtained from movement. Indeed, a stereoscopic sensor is unsuited for acquiring reliable information on contours whose orientation is close to epipolar lines, whereas the analysis of movement is problematic on contours whose orientation in the image is close to projection in the image from the direction of the translation movement. An optimization of these two systems is obtained by controlling the camera in such a way that the fixation point remains immobile along the image sequences, since the trajectory of movement is itself a constraint on a vertical plane. The apparent velocity field is obtained in two stages: the rotation velocities are obtained by measuring the rotation angle of the camera around the axis x; then, the translation components are obtained using this angle and the distance to the fixation point obtained by stereovision. This information and the components of normal velocity at contours allow a complete interpretation of the apparent velocity field, which leads to the determination of a depth chart. This chart is then fused with the depth chart obtained by stereovision.

The second approach suggested by Sandini and Tistarelli [SAN 90] uses an active monocular approach in the sense that here also, the camera is controlled in order to stabilize the fixation point along the image sequence (strategy close to that of Aloimonos [ALO 87] in the study of the problem of *"structure from motion"*). These constraints imposed on the movement of the scene in the image make it possible to easily reach the parameters of camera movement. The estimate of the 3D structure of the scene relies on the mapping of points, the analysis of the apparent movement of these points in image sequences and the estimate of the camera movement. It remains that the contribution of active

vision in these two approaches is no longer obvious. A quantitative (and even qualitative) study would have been interesting.

Huang and Aloimonos [HUA 91] are placed mainly within the context of purposive vision to solve the problem of estimating structure from motion. The studied problem deals with the localization of mobile objects with respect to one another. The proposed solution makes it possible to avoid the difficult stages of estimating the 3D movement and the estimation of the apparent motion field. It depends only on the use of the spatial and temporal derivatives of the function of light intensity and thus on the components perpendicular to apparent velocity contours. This estimate is indeed much easier to obtain. However, the depth chart thus determined will not be complete. The strategy of camera displacement is very simple since it consists of a translation movement along the optical axis. This approach gives good results which are relatively robust to noises except of course in the area close to the optical axis.

6.2.4.2. *Optimal 3D reconstruction of a primitive*

The purpose of the method described here is to determine adequate camera movements in order to obtain an optimal estimate of the characteristics of the primitive. This optimization problem is approached under two different aspects:

– suppression of discretization errors;

– minimization of the effects of measurement errors (measurement errors in the images and measurements of camera movement).

6.2.4.2.1. Suppression of discretization errors

The 3D reconstruction method that was previously presented relies on the apparent velocity measurement of the primitive in the image (i.e., $\dot{\mathbf{P}}$, the velocity of the parameters representing the projection of the primitive). However, this value $\dot{\mathbf{P}}$ is not directly measurable; the values measured in the sequence of images authorize only the measurement of $\Delta\mathbf{P}$, which represents the variation of parameters $\mathbf{P}$ during the interval of time Δt between two images. The use of $\Delta\mathbf{P}/\Delta t$ instead of $\dot{\mathbf{P}}$ in the estimate of 3D parameters of the primitive can lead to gross errors. A manner of solving this problem is to ensure the following equality:

$$\dot{\mathbf{P}} = \frac{\Delta\mathbf{P}}{\Delta t}, \quad \forall t. \tag{6.23}$$

Consequently, discretization will no longer have any effect on the quality of reconstruction. Such a condition will be satisfied only if:

$$\ddot{\mathbf{P}} = \cdots = \mathbf{P}^{[n]} = 0, \quad \forall t. \tag{6.24}$$

From (6.6), noted here by $\dot{\mathbf{P}} = f(\mathbf{P}, \mathbf{p_L}, \mathbf{T})$, it is possible to deduce:

$$\ddot{\mathbf{P}} = \mathbf{L}_{\mathbf{P}}^{T}\left(\mathbf{P}, \mathbf{p_L}\right) \dot{\mathbf{T}} + \frac{\partial f}{\partial \mathbf{P}} \dot{\mathbf{P}} + \frac{\partial f}{\partial \mathbf{p_L}} \dot{\mathbf{p_L}}. \tag{6.25}$$

A condition sufficient to satisfy (6.24) is to constrain the camera movements such that:

$$\dot{\mathbf{P}} = \dot{\mathbf{p_L}} = 0, \quad \forall t. \tag{6.26}$$

Indeed, in this case $\mathbf{T} \in \operatorname{Ker} \mathbf{L}_{\mathbf{P}}^{T}$, $\forall t$. Using (6.26), we can show that $\dot{\mathbf{T}} \in \operatorname{Ker} \mathbf{L}_{\mathbf{P}}^{T}$, $\forall t$, from which we can deduce that $\ddot{\mathbf{P}} = 0$, $\forall t$. Therefore, a recurrent method allows us to show that (6.24) is always verified.

In other words, a solution to remove discretization errors is that the limb surface of the primitive remains motionless in the pointer of the camera and that, moreover, 3D primitive projection appears in the same position in the image when the camera is moving. It is possible to show that except in the case of point and lines, the first condition $\dot{\mathbf{P}} = 0$ implies the second $\dot{\mathbf{p_L}} = 0$, which then reduces the problem to a *fixation* problem.

NOTE 6.1. The above suggested condition to remove discretization errors is only sufficient, but not compulsorily necessary. Indeed, there are camera movements such that $\ddot{\mathbf{P}} = 0$ with $\dot{\mathbf{P}} \neq 0$. For a point, for example, we can show [BOU 93] that $\ddot{\mathbf{P}} = 0$ when the camera moves according to a translation movement parallel to the image plane with constant velocity (i.e., $V_x = V_1$, $V_y = V_2$ and $V_z = \Omega_x = \Omega_y = \Omega_z = 0$). Generally, calculating all the solutions of the non-linear system (6.24) seems out of reach. Moreover, as these solutions depend on the knowledge of $\frac{\partial f}{\partial \mathbf{P}}$ and $\frac{\partial f}{\partial \mathbf{p}_L}$, they strongly depend on the nature of the primitive considered. On the other hand, condition (6.26) is valid irrespective of the nature of the primitive. Moreover, it has the advantage of maintaining the primitive in the vision field of the camera throughout the estimate process.

6.2.4.2.2. Minimizing the effects of the measurement errors

The configuration of a primitive in the image influences the quality of its estimate. It is thus shown in [BOU 93] that certain constraints on the primitive position in the image and constraints on the camera movement are necessary for a good parameter estimate of the primitive considered.

Let $\mathbf{p}$ be one of the parameters representing the primitive to be reconstructed; then, errors on the estimate of $\mathbf{p}$ are closely related to measurement errors on visual information $\mathbf{P}$ and $\dot{\mathbf{P}}$ as well as on camera movement $\mathbf{T}$. If the measurement errors on $\mathbf{P}$, $\dot{\mathbf{P}}$, and $\mathbf{T}$ are assumed to be decorrelated, the uncertainty $\sigma_{\mathbf{p}}$ on the estimation of $\mathbf{p}$ can then be expressed in the following manner:

$$\left(\sigma_p\right)^2 = \sum_{i=1}^{m} \left(\frac{\partial \mathbf{p}}{\partial P_i}\right)^2 \left(\sigma_{P_i}\right)^2 + \sum_{j=1}^{m} \left(\frac{\partial \mathbf{p}}{\partial \dot{P}_j}\right)^2 \left(\sigma_{\dot{P}_j}\right)^2 + \sum_{k=1}^{6} \left(\frac{\partial \mathbf{p}}{\partial T_k}\right)^2 \left(\sigma_{T_k}\right)^2. \tag{6.27}$$

where $P_{i(i=1...m)}$, $\dot{P}_{j(j=1...m)}$ and $T_{k(k=1...6)}$ are the different components of $\mathbf{P}$, $\dot{\mathbf{P}}$ and $\mathbf{T}$, respectively. Minimizing $\sigma_{\mathbf{p}}$ then amounts to minimizing each term $(\frac{\partial p}{\partial a})^2$ where a represents each of the variables of P_i, $\dot{P}_j$ and T_k. Then, there remains to determine the values of $\mathbf{P}$ such that:

$$\left(\frac{\partial \left(\frac{\partial \mathbf{p}}{\partial a}\right)^2}{\partial P_j}\right) = 0, \quad \forall a \text{ and } \forall j = 1 \text{ to } m. \tag{6.28}$$

Analytically determining the solution of such a system seems impossible. Yet, the derivation of interesting specific solutions remains possible. Thus, solutions were given for the principal geometric primitives (point, line, cylinder, sphere, circle) [BOU 93]. For example:

– In the case of a *point*, the effects of measurement errors on its depth z are minimized if the point constantly appears in the center of the image (i.e., $X = \dot{X} = Y = \dot{Y} = 0, \forall t$) and if in addition we have $V_z = \Omega_z = 0$. This means that the camera must move on a sphere whose center is the point to be reconstructed.

– In the case of a *line*, the effects of measurement errors are minimized if this appears centered and vertical (or horizontal) in the image and if, in addition, $V_y = V_z = \Omega_x = 0$ (or $V_x = V_z = \Omega_y = 0$) (see Figure 6.2a).

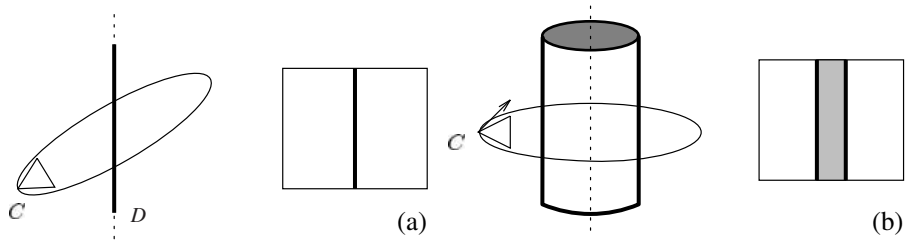


Figure 6.2. *Optimal movements of the camera for the (a) estimate of a line and (b) estimate of a cylinder. For each of these illustrations, the drawing on the left represents the object to be reconstructed as well as the trajectory of the camera (schematized by a triangle) and that on the right represents the image perceived by the latter*

– Lastly, in the case of a *cylinder*, the effects of measurement errors are minimized if it is projected in two parallel symmetric vertical line forms (or horizontal) in the image along with $V_y = 0$ (or) $V_x = 0$. The camera must then move on a circle whose center is crossed by the axis of the cylinder and is perpendicular to it (see Figure 6.2b).

The two tasks necessary for the process of optimal estimate of the 3D parameters of a primitive are thus a fixing task and a focusing task. These two tasks are similar to those specified by Sandini [SAN 90] or Aloimonos [ALO 87] in the solutions they proposed to the problem of structure from motion (let us recall that this task consisted of stabilizing fixation ("*gaze control*") point along the image sequence). The conclusions resulting from [BOU 93], however, go much further since in each case, additional constraints were imposed on the trajectory of the camera (suppression of movements in the direction and around the optical axis in the case of point ($V_z = \Omega_z = 0$), movement along the axis in the case of line ($V_y = V_z = \Omega_x = 0$) or cylinder ($V_y = 0$)). Indeed, these movements do not bring any information within the context of the 3D reconstruction process. Here, the principle of purposive vision [ALO 90] is applied, where useless movements (or actions) are removed.

6.2.5. *Generation of camera movements*

Visual servoing techniques [ESP 92, HUT 96] allow us to carry out various movements automatically satisfying the constraints necessary for an optimal estimate (fixation tasks and focusing tasks). Visual servoing consists of introducing information extracted directly from the image in a command loop directly and in closed loop. In our case, tasks are expressed as the

regulation at zero of a task function combining a primary task (such that the primitive appears in the image in its desired position) and a secondary task. This secondary task is constructed in order to allow camera movements while ensuring the realization of the primary task.

In section 6.2, we defined the interaction matrix $\mathbf{L}$ (see equation (6.6)). This matrix is the basis of the mechanism of visual servoing because it makes it possible to connect the movements of an object in the image to the movement of the camera.

Vision tasks can be expressed as the zero regulation of the following task function:

$$\mathbf{e_1} = \mathbf{C}\big(\mathbf{P} - \mathbf{P_d}\big) \tag{6.29}$$

where $\mathbf{P}$ is a vector representing the set of visual information selected to carry out the task (i.e., the current value of parameters representing the primitive considered, measured in the image in each iteration of command loop) and $\mathbf{P_d}$ represents its desired value. $\mathbf{C}$, called the combination matrix, can be defined by: $\mathbf{C} = \mathbf{W}\mathbf{L}_{\mathbf{P}}^{T^+}(\mathbf{P}, \widehat{\mathbf{p_L}})$, where the parameters $\mathbf{p_L}$ used in the interaction matrix are estimated online using the 3D estimate process, presented previously. $\mathbf{W}$ is defined as a full row matrix such as $\mathrm{Ker}\,\mathbf{W} = \mathrm{Ker}\,\mathbf{L}_{\mathbf{p}}^{\mathbf{T}}(\mathbf{p}, \widehat{\mathbf{P}})$. $\mathbf{L}^+$ indicates the pseudo-inverse of the matrix $\mathbf{L}$.

If the task of specified vision does not constrain n degrees of robot freedom, the redundancy available can be used to achieve a secondary task; then, we obtain the following task function:

$$\mathbf{e} = \mathbf{W}^+\mathbf{C}\big(\mathbf{P} - \mathbf{P_d}\big) + \big(\mathbf{I_n} - \mathbf{W}^+\mathbf{W}\big)\mathbf{g}_s^T \tag{6.30}$$

where:

– $\mathbf{W}^+$ and $\mathbf{I_n} - \mathbf{W}^+\mathbf{W}$ are two projection operators which ensure that the camera movement due to the secondary task is compatible with the regulation of $\mathbf{P}$ towards $\mathbf{P_d}$. In fact, as a result of choosing $\mathbf{W}$, $\mathbf{I_n} - \mathbf{W}^+\mathbf{W}$ theoretically belongs to kernel Ker $\mathbf{L}_{\mathbf{P}}^T$, which implies that the realization of the secondary task will not have any effect on the primary task ($\mathbf{L}_{\mathbf{P}}^T(\mathbf{I_n} - \mathbf{W}^+\mathbf{W})\mathbf{g}_s^T = 0$). However, if the parameters $\widehat{\mathbf{p}}_L$ are wrongly estimated, then the projection operator $\mathbf{I_n} - \mathbf{W}^+\mathbf{W}$ will not exactly belong to Ker $\mathbf{L}_{\mathbf{P}}^T$ and the secondary task will introduce a disturbance in achieving the visual task.

– $\mathbf{g_s}$ is a secondary task, which can be expressed as a gradient of a function h_s to be minimized ($\mathbf{g_s} = \frac{\partial h_s}{\partial \mathbf{r}}$). This cost function is thus minimized under the constraint that $\mathbf{e}_1$ is realized.

The control issue amounts to the regulation of the task function $\mathbf{e}$. The latter is perfectly achieved if, at each moment t: $\mathbf{e}(t) = 0$. A general method making it possible to regulate the task function $\mathbf{e}$ is presented in [SAM 91]. A simplified command law, which calculates the camera speed $\mathbf{T_c}$ and which ensures an exponential decrease of the task function $\mathbf{e}$ is given by [CHA 90, ESP 92]:

$$\mathbf{T_c} = -\lambda \mathbf{e} - \left(\mathbf{I_n} - \mathbf{W}^{+}\mathbf{W}\right) \frac{\partial g_s^T}{\partial t} \tag{6.31}$$

where λ is a gain regulating the exponential convergence velocity of $\mathbf{e}$.

We will now illustrate the control law used in the reconstruction process of 3D primitives for the particular case of the point.

The interaction matrix of a point is given by equation (6.12). By knowing $\mathbf{L_P}$, it is easy to calculate its pseudo-inverse $\mathbf{L_P^+}$. $\mathbf{L_P}$ being full row 2, we have $\mathbf{W} = \mathbf{L_P^T}$ and thus $\mathbf{W}^{+} = \mathbf{L_P^{T^+}}$.

The secondary task $\mathbf{g_s}$ is selected as equal to:

$$g_s = \begin{pmatrix} x(t) - x(0) - V_1 t \\ y(t) - y(0) - V_2 t \\ z(t) - z(0) + (X(t)V_1 + Y(t)V_2)t \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{6.32}$$

where $(x(t), y(t), z(t))$ is the present position of the camera and $(x(0), y(0), z(0))$ is its initial position. V_1 and V_2 are two constants that the user chooses in order to obtain a constant velocity movement in the $\vec{x}$ and $\vec{y}$ directions of the camera pointer. The choice of the third component g_s allows us to ensure the constraint $\dot{z} = 0$ canceling the discretization errors.

When the task is perfectly carried out (i.e., when $X = Y = 0$), we have:

$$\mathbf{Tc} = -\left(\mathbf{I} - \mathbf{W}^{+}\mathbf{W}\right) \frac{\partial g_s}{\partial t} = \frac{z}{1+z^2} \begin{pmatrix} -zV_1 \\ -zV_2 \\ 0 \\ V_2 \\ -V_1 \\ 0 \end{pmatrix} \tag{6.33}$$

which corresponds to a camera movement on a sphere centered on the point if the value estimated from z is correct.

6.3. Reconstruction of a complete scene

The problem which interests us now is that of the complete reconstruction of a scene containing several objects. Indeed, if the preceding method allows a very precise and reliable estimate of parameters of the primitives concerned, it makes it possible to rebuild only one primitive at a time. Thus, the purpose of this section is to describe perception strategies capable of providing a precise and complete 3D representation of the scene. In a way, the approach used consists of automatically selecting relevant image information and then successively focusing the camera on various objects of the scene in order to reconstruct them. Exploration phases are particularly necessary in order to ensure the completeness of the reconstruction.

6.3.1. *Automatic positioning of the camera for the observation of the scene*

The majority of works discussing observation or exploration of scenes assume as known a complete model of the scene [COW 88, TAR 95]. The problem is more complex if information on the scene is incomplete or zero (i.e., if the sensor moves in an unknown environment). Then, it is necessary to carry out an autonomous exploration task. Many articles [CON 85, WIX 94, MAV 93, WHA 94] deal with this problem under different angles. Connoly [CON 85] proposes using a laser sensor to determine a complete model of the scene from a set of viewpoints. He describes the algorithm known as *planetarium*, which uses decomposition in *octree* of the scene. The camera moves on the surface of a regularly sampled sphere circumscribed at the scene. The viewpoint making it possible to reveal the most important unseen zone is selected. Maver and Bajcsy [MAV 93] use occlusions to determine various viewpoints necessary to acquire 3D information of the hidden parts of the environment. The sensor is composed of a camera and a laser plane. In [WIX 94], Wixson describes strategies to find an object known in an encumbered zone. He explores various strategies for the exploration of a 2D world with the help of a 1D sensor. Two alternatives for the planetarium algorithm are proposed: the first based on the search for the point of view offering the maximum visibility of the still unknown zones and the second supporting the point of view that minimizes sensor displacement. Another strategy studied is based on the use of occluded edges. The principal

interest of this method rests on the cost and the benefits of the perception operations carried out.

6.3.2. *Scene reconstruction: general principle*

In addition to the basic general assumptions on the maximum size of the scene and the nature of objects constituting it (in fact segments and cylinders), the only data that we have are those provided by 2D images acquired by the camera. The perception strategies presented thereafter are mainly based on the use of this 2D information. One of the fundamental stages of our algorithm is thus the creation of databases containing this information. The databases mainly used, noted by ω_{ϕ_t} (where ϕ_t represents the position of the camera), contain the list of segments corresponding to the projection of visible objects of the scene in the image from the ϕ_t position. Each segment is associated with information indicating its position in the image and whether or not it was processed.

The algorithms of image processing that we use during visual control phases and reconstruction authorize only a real-time sequence of a small number of segments [BOU 93]. Hence, in order to respect the real-time constraint, these databases cannot be created in each iteration of the estimate process, but only at the end of reconstructing a primitive (i.e., when a specified precision or a maximum number of iterations are reached).

In addition, a second database, noted by Ω_{Φ_t}, is also used. It regroups all databases ω_{ϕ_i}, $i = 1$ to t. More precisely, Ω_{Φ_t} contains the set of segments, which have not yet been subjected to a reconstruction at moment t and the position of the camera from which they are observed.

From a viewpoint ϕ_t and using the already collected 3D information, it is possible, by using an algorithm of launching rays, to calculate the zone observed $V(\phi_t)$. Let us note by $\mathcal{V}(\Phi)$ the space zone observed by the camera from the beginning of the reconstruction process (i.e., 3D primitives and known free space). We have:

$$\mathcal{V}(\Phi_t) = \bigcup_{i=1}^{t} V(\phi_i), \quad \text{with} \quad \Phi_t = \bigcup_{i=1}^{t} \phi_t \tag{6.34}$$

The reconstruction of the scene will be completed when:

$$\forall \phi_{t+1} \in \mathcal{E}, \quad \mathcal{V}(\Phi_t) \cup V(\phi_{t+1}) = \mathcal{V}(\Phi_t) \tag{6.35}$$

where $\mathcal{E}$ defines the set of viewpoints, which can be reached by the camera during the exploration process. This means that exploration is as complete as possible if, for all the pertaining viewpoints at $\mathcal{E}$, the camera observes an already known zone. By using the information observed (i.e., the database Ω_{Φ_t} and the partial chart of the environment), we can define strategies of positioning the camera, which will ensure a reconstruction of the scene as complete as possible. This exploration process, making it possible to calculate the position ϕ_{t+1}, is made up of two distinct levels:

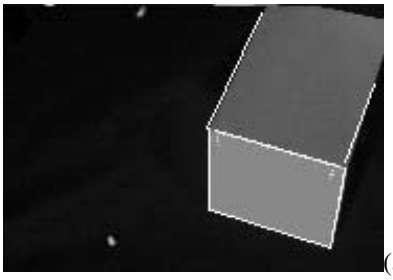
– a *local exploration* is carried out if a segment corresponding to a new primitive appears in the vision field of the camera or if such a segment was previously observed from another position of the camera. In this case, an explicit calculation from a viewpoint is not necessary;

– on the other hand, in other cases, when all the previously observed segments have been reconstructed, a more complex strategy must be implemented in a manner to focus the camera on the zones of the scene which have not been observed yet. Then, we will discuss *global exploration*.

6.3.3. *Local focusing strategy*

In order to minimize the displacement of camera, it is desirable to implement a *local* strategy using the current knowledge on the scene in an explicit way [MAR 99b]. This strategy is controlled by events detected in the image and by previously acquired 3D information.

The approach is based on the following assumptions: the scene consists of 3D objects connected by topological relations; the projection of 3D objects in the image can be represented by a graph where the nodes are the multiple junctions and the arcs, contours. Each arc of this graph corresponds to a 2D segment of the current database ω_{ϕ_t} (Figure 6.3b). These arcs are valued according to their position in the image and the possible knowledge that we have on the corresponding 3D primitive (i.e., whether they correspond or not to the projection in the image of an already reconstructed primitive). If several arcs of this graph prove to be the projection of non-reconstructed primitives, a choice is made to know which one to select. This choice paving way for minimizing the distance covered by the camera is realized using the graph and the current database. We search for an untreated segment related to the one that has just been reconstructed. If such a segment exists and is single, it is retained and reconstructed. If such a segment does not exist or is not single (case of a multiple junction), the segment whose position in the image is closest to an optimal position for reconstruction is then retained.


(a)

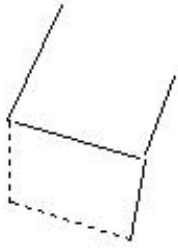

(b)

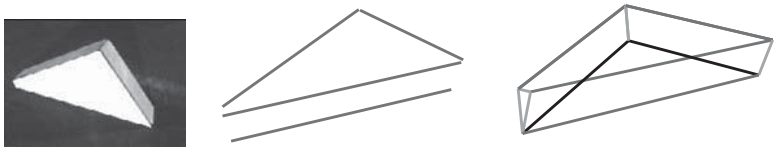
Figure 6.3. *(a) Acquired image (b) 2D databases (segments in dotted lines were already reconstructed)*

If all segments of ω_{ϕ_t} correspond to primitives that were already reconstructed, we search for an untreated segment Ω_{Φ_t} in the total database. The camera then moves to a position from which it had been observed (backward return phase) and a reconstruction of the primitive associated with this segment is carried out.

We finally obtain a 3D modeling of a part of the scene considered. However, this modeling remains a low level modeling and is sometimes incomplete. As regards the segments, it is desirable to go by a hierarchical representation in terms of 3D segments, 3D junctions, polygons and, as far as possible, sides, etc. The method developed for this purpose comes on top of the algorithm of incremental reconstruction and depends on prediction/*hypothesis verification* techniques. Due to uncertainties in measurements and observations, we used a probabilistic approach. The generation of an assumption is done using the present 3D model, and also by being based on a certain number of very general *deductions* on the characteristics of a polyhedric scene. Let us note that in no case are these *deductions* opposed to assumptions that were initially realized on the type of scenes that we decided to study (namely, any preliminary knowledge about the number and localization of the primitives and objects of the scene). The knowledge that we introduced is coded in Bayesian networks. Bayesian networks indeed lend themselves very well to reasoning and decision-making in the presence of uncertainty [PEA 88, DJI 96]. In our case, these networks make it possible to put forth assumptions on the existence and localization of new objects, then to propose the execution of an action leading to verify or confirm this assumption and, finally, according to the result of verification stage, to supplement the 3D model of the scene. The verification stage is simultaneously based on the observations already carried out on the scene,

and also on an acquisition of new information requiring a displacement of the sensor (displacement that is automatically produced by visual control). Lastly, modeling of the scene and creation of new objects (junctions, polygons, etc.) depend on 3D and 2D information coming from the set of reconstruction process already realized, and on the contribution of information introduced by validated assumptions [MAR 99b].

Following this approach enables us to have a modeling of the scene in terms of objects and no longer in terms of simple primitives like 3D segments (see Figure 6.4).



(a) (b) (c)

Figure 6.4. *"Polyhedral" scene (a) initial image of the scene (b) model acquired by making use of only the incremental reconstruction module (c) reconstructed model by making use of the prediction/hypothesis verification module*

The *local* strategy that we presented ensures an effective reconstruction of any primitive which was observed during the reconstruction process. It does not call for an explicit calculation from new points of view and locally minimizes the displacement of the camera. However, it does not ensure a complete reconstruction of the scene. To solve this problem, a *global* strategy must be implemented.

6.3.4. *Completeness of reconstruction: selection of viewpoints*

When all the primitives observed during local exploration phases are reconstructed (the database Ω_{Φ_t} is empty), we must determine the viewpoints of the camera allowing us to possibly discover new objects to be reconstructed and thus ensure the completeness of reconstruction. Such viewpoints are calculated using the current knowledge on the spatial geometry of the scene and a representation of the already observed spatial zones.

6.3.4.1. *Calculation of new viewpoints*

Viewpoint search is carried out by minimizing a cost function $\mathcal{F}(\phi)$, which represents the quality of a viewpoint ϕ. We have retained three criteria, which are integrated in this function:

– discovered volume gain brought by the new position: since the purpose of exploration is the acquisition of additional information on the scene, it is imperative to model this gain in the cost function;

– displacement cost towards the new position: such a criterion is justified by the fact that we wish to have a minimal length trajectory by avoiding camera movements of great amplitude;

– accessibility: the new point of view must obviously be accessible to camera.

Each of these criteria is associated with a value measurement in $[0, 1] \cup \infty$. A value of measurement close to 0 indicates that there is maximum satisfaction brought by this viewpoint with respect to the related criterion. On the other hand, a value close to 1 indicates that the position does not have any interest with respect to the related criterion. Lastly, an infinite value indicates that the position is to be automatically rejected.

The gain brought by a new position ϕ_{t+1} is defined by the volume of the unobserved zone, which appears in the cone of camera vision when it moves from ϕ_t to ϕ_{t+1}. The discovered zone from this position corresponds to the $\mathcal{G}(\phi_{t+1})$ zone defined by (see Figure 6.5a):

$$\mathcal{G}\big(\phi_{t+1}\big) = \mathcal{V}\big(\phi_{t+1}\big) - \mathcal{V}\big(\phi_{t+1}\big) \cap \mathcal{V}\big(\Phi_t\big) \tag{6.36}$$

The measurement of gain related to position ϕ_{t+1} is therefore given by:

$$g\big(\phi_{t+1}\big) = 1 - \frac{\text{volume}\left(\mathcal{G}\big(\phi_{t+1}\big)\right)}{\text{volume}\left(\mathcal{V}\big(\phi_{t+1}\big)\right)} \tag{6.37}$$

NOTE 6.2. Zone $\mathcal{G}(\phi_t)$ corresponds to a potentially discovered zone. In fact, if a new object appears in the vision field, occlusions due to this object cause the really observed zone $\mathcal{G}'(\phi_t)$ to be smaller than ($\mathcal{G}'(\phi_t) \subseteq \mathcal{G}(\phi_t)$).

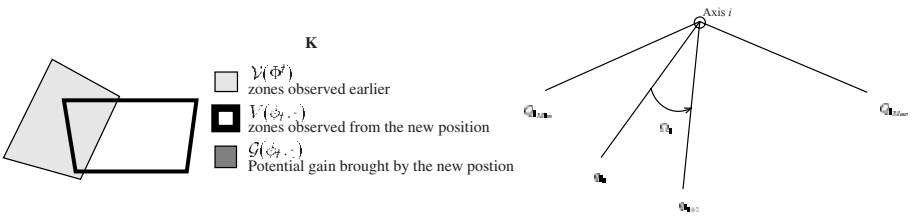


Figure 6.5. *Calculation of gain and cost function*

Cost measurement of displacement between two positions ϕ_t and ϕ_{t+1} is defined by calculating the distance between ϕ_t and ϕ_{t+1}. More precisely, this cost is given by (see Figure 6.5b):

$$\mathcal{C}(\phi_t, \phi_{t+1}) = \frac{1}{N_{ddl}} \sum_{i=1}^{N_{ddl}} \beta_i \frac{|q_{i_t} - q_{i_{t+1}}|}{|Q_{i_{\text{Max}}} - Q_{i_{\text{Min}}}|} \tag{6.38}$$

where:

– N_{ddl} represents the number of degrees of freedom of the robot;

– q_i represents the articular coordinate of the axis i ($\phi = (q_0, q_1, \ldots, q_{N_{ddl}})$);

– $|Q_{i_{\text{Max}}} - Q_{i_{\text{Min}}}|$ is the distance between two joint limits of the axis i;

– β_i coefficients are weights, which make it possible to support displacements on certain axes of the robot (rotation movements of the camera, for example, can be preferred to translation movements).

Moreover, an additional constraint is associated with certain positions of the robot. This constraint tends to ignore unattainable positions for the camera because of the joint limits of the robot or due to the presence of obstacles. An infinite penalty is given to a position if it is not accessible:

$$\mathcal{A}(\phi) = \begin{cases} 0 & \text{if } \phi \text{ is accessible} \\ \infty & \text{otherwise} \end{cases} \tag{6.39}$$

Cost function $\mathcal{F}(\phi_{t+1})$ is defined by the weighted sum of these different measurements:

$$\mathcal{F}(\phi_{t+1}) = \mathcal{A}(\phi) + \alpha_1 g(\phi_{t+1}) + \alpha_2 \mathcal{C}(\phi_t, \phi_{t+1}) \tag{6.40}$$

Determination of coefficients α_i in an optimization problem of this type is a non-trivial problem. We are contented to choose these coefficients in an empirical way. However, their value fixes the associated priority order to each criterion; accessibility of course has priority (the "binary" character of its result makes any weighting process pointless). Moreover, the discovery of new zones to be explored being our objective, we have chosen $\alpha_1 > \alpha_2$.

6.3.4.2. *Optimization*

Each position ϕ can *a priori* be a solution to this optimization problem. However, in order to overcome the problem, we authorize the camera to move on the surface of a circumscribed sphere at the scene. The position of the camera can then be described by a vector with five parameters $(\theta, \varphi, \alpha, \beta, \gamma)$, where θ and φ represent the latitude and the longitude of the camera on the sphere, respectively, and α, β and γ represent the orientation of the camera.

It is also possible to move in the interior of this sphere if the accessibility zones of the camera are restricted to already observed zones that do not contain possible obstacles [MAR 99a]. To minimize $\mathcal{F}(\phi)$, we chose to use a traditional deterministic method of gradients type combined with a diminishing step in various levels: first, we use huge increments in order to determine the space area of parameters where the optimum of the $\mathcal{F}(\phi)$ function is probably located. Then, we reiterate the process from this new position with a weaker increment. Contrary to stochastic methods of simulated annealing, we can only ensure that convergence is carried out towards the total minimum of the function. However, the gain in calculation time is very important and the experiments showed that a correct optimum is always reached in a lesser number of iterations. Moreover, the interest to find a total minimum of the cost function did not appear fundamental to us insofar as a position bringing important and additional information is found.

6.4. Results

The experiments presented in this chapter were carried out on the robotic vision cell of IRISA composed of a CCD camera mounted on the effector of a robot with 6 degrees of freedom.

The creation of databases ω_{ϕ_t}, including extraction of contours (Shen-Castan filter and thresholding) and a polygonal approximation of contours, was carried out in a SUN SPARCS Station 20 workstation in 1 second approximately. For 3D reconstruction phases by active vision, image processing was carried out on a specialized chart. The processing consists of following on the sequence of acquired images the 2D segment and determining the parameters (ρ, θ) describing its position in the image. The extraction and follow-up of the segment (in fact a list of contour points) were carried out in 80 ms. The method used is described in [BOU 93]. It is based on a local and reliable mapping of moving contour elements constituting the selected line. The estimate of 3D parameters of primitives as well as

the calculation of the command were also carried out in a SUN station at a rate of 10 Hz. The calibration process employed depends on the method suggested in [CHA 89]. Finally, calculation from a new viewpoint, resulting from the minimization of functional (6.40), was carried out in approximately 1 second with accuracy of a centimeter and in approximately 10 seconds with a precision in the order of a millimeter.

6.4.1. *Reconstruction of 3D primitive: case of the cylinder*

At first, we present results related to the reconstruction of a cylinder in order to show the interest of this reconstruction method by active vision. More complete results are described in [BOU 93]. In order to obtain a reliable and non-biased estimate of its parameters, the cylinder must appear centered and vertical or horizontal in the image (see Figure 6.6) during the movement of the camera, which undergoes a rotation in constant distance of the axis of the cylinder. The techniques of visual control were thus employed to carry out this task in real-time: an iteration of command loop and an estimate are simultaneously carried out in 100 ms. Figure 6.7 presents an error between the estimated value of the radius and its actual value (i.e., $r_i - r^*$) obtained using an estimate based on the two limbs of the cylinder.

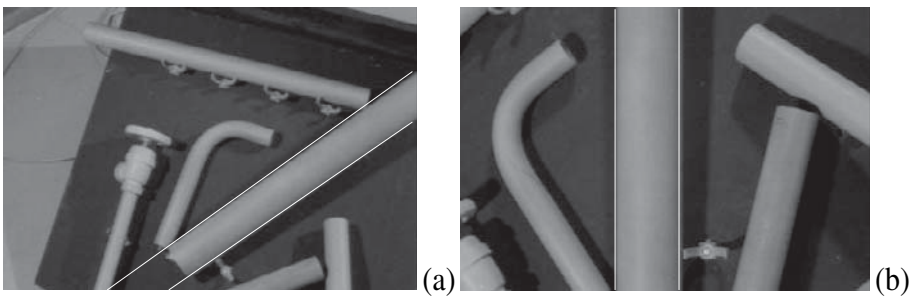


Figure 6.6. *Cylinder to be reconstructed before and after the focalization task*

Stability tests were also carried out. The parameters of the cylinder were estimated 50 times on the basis of different initial positions. The reported results in Figure 6.7b show, for each of the 50 estimates, the estimated radius $\widehat{r}$ as well as the standard deviation $\sigma_{\widehat{r}}$ on this estimate. For each estimate, the error between the estimated value and the actual value is less than 0.1 mm and the standard deviation on the set of averages is less than 0.02 mm. This shows that the reconstruction algorithm is reliable, stable and precise.

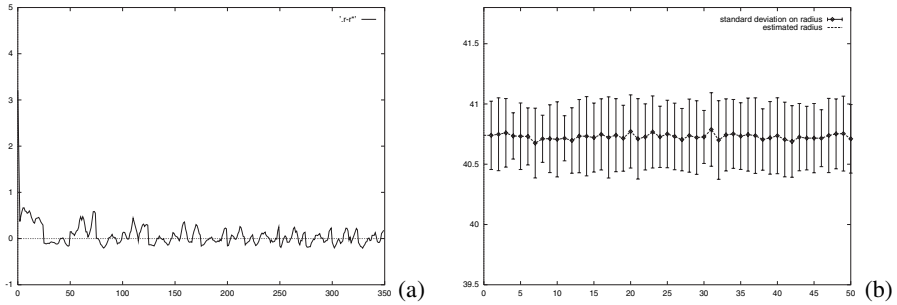


Figure 6.7. *(a) estimate of radius $R_i - R^*$ (in mm); (b) stability tests*

6.4.2. *Perception strategies*

The scene considered is made up of a cylinder and several polygons arranged on different planes. Figure 6.8 shows an exterior view of the scene and various objects that make it up.

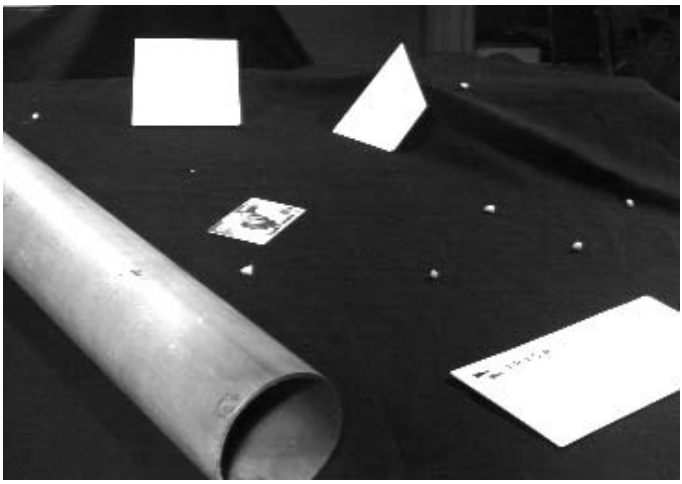
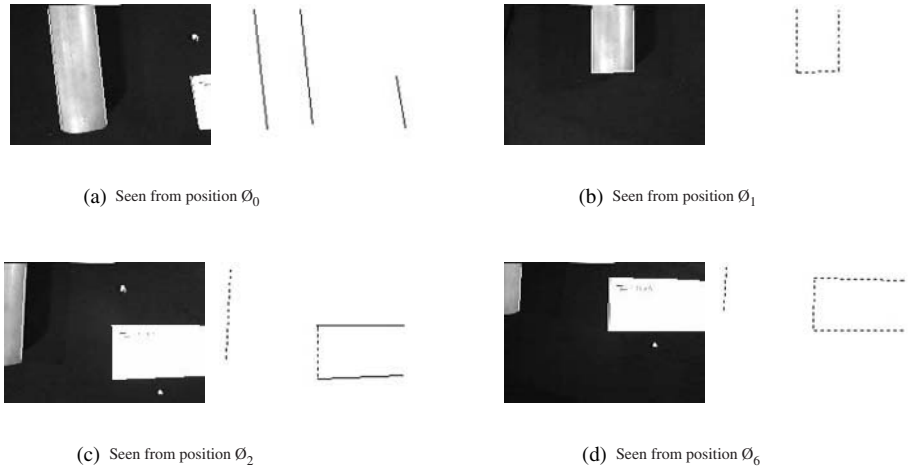


Figure 6.8. *Exterior view of the scene*

6.4.2.1. *Local exploration*

Figure 6.9 represents images acquired before each optimal reconstruction. Each of these images is associated with the corresponding 2D database. Continuous lines show database elements, which have not been treated yet.



(a) Seen from position $\emptyset_0$ (b) Seen from position $\emptyset_1$

(c) Seen from position $\emptyset_2$ (d) Seen from position $\emptyset_6$

Figure 6.9. *Local exploration of the scene*

Dotted lines represent segments corresponding to already reconstructed 3D primitives.

Figure 6.9a shows the image acquired from the position ϕ_0 of the camera. No further reconstruction was carried out and only three segments are visible from this position. We can note that the complete scene is not visible from this position of the camera. The segment extracted from database ω_{ϕ_0} is that corresponding to the right side limb of the cylinder. After the reconstruction phase of the cylinder, the camera is positioned in ϕ_1 (Figure 6.9b). All segments of the database ω_{ϕ_1} were treated. After consultation of the global database Ω_Φ, we note that a segment was observed from the position ϕ_0 and was not reconstructed yet. The camera thus moves in ϕ_0 and is focused on the segment selected. After the estimation of this primitive, the camera is positioned in ϕ_2 (Figure 6.9c). Two segments corresponding to non-reconstructed primitives appear in the database ω_{ϕ_2}. The segment nearest to the center of the image is selected and reconstructed. This process is renewed until all primitives observed during this phase of local exploration are estimated (Figure 6.9d corresponding to position ϕ_6 of the camera). Let us note that primitives that did not appear in the initial vision field of the camera were discovered and reconstructed. The estimated scene at this stage of the reconstruction process is presented in Figure 6.10.

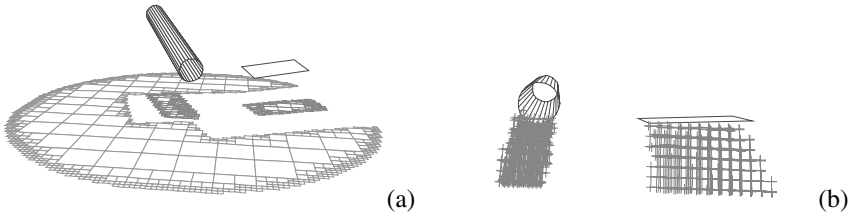


Figure 6.10. *Result of the local exploration of the scene considered: (a) reconstructed scene and projection on a virtual plane of the non-observed zone; (b) side view representing the reconstructed scene and the occluded zones*

6.4.2.2. *Total exploration*

Figure 6.11 shows the different stages necessary for a complete exploration of the scene. Each figure describes the reconstructed scene, the trajectory of camera up to its current position and a visualization of the non-observed zone. Figure 6.11a corresponds to position ϕ_6 of the camera obtained as

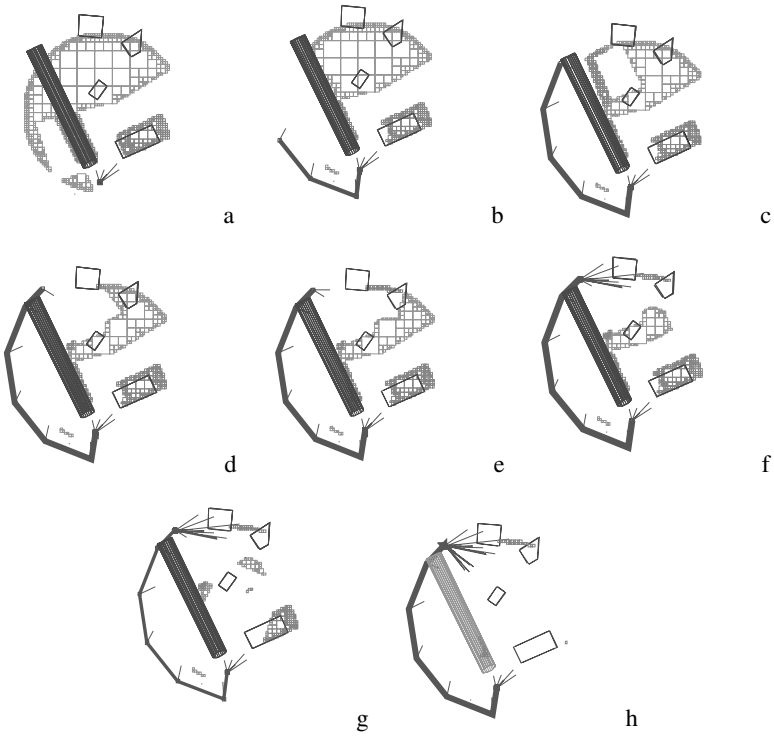


Figure 6.11. *Various stages of total exploration (trajectory of camera, reconstructed scene, and projection on the ground of the non-observed zone)*

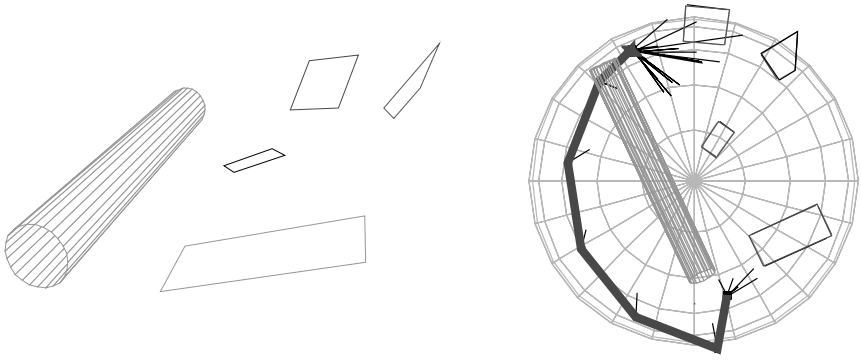
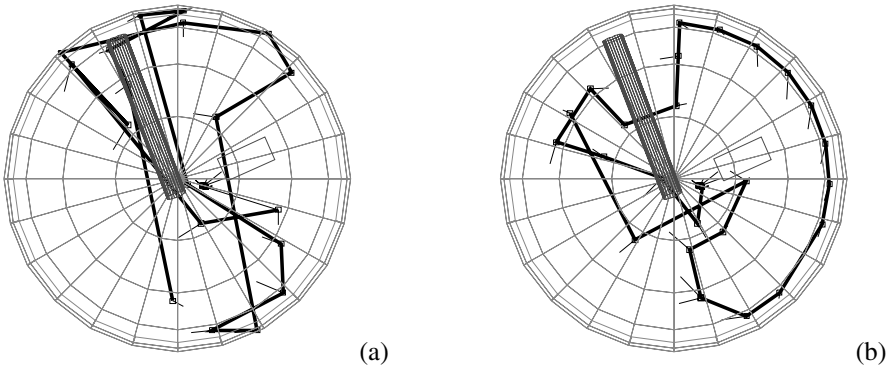


Figure 6.12. *Visualization of the reconstructed scene and polar view of the final trajectory of the camera during total exploration*

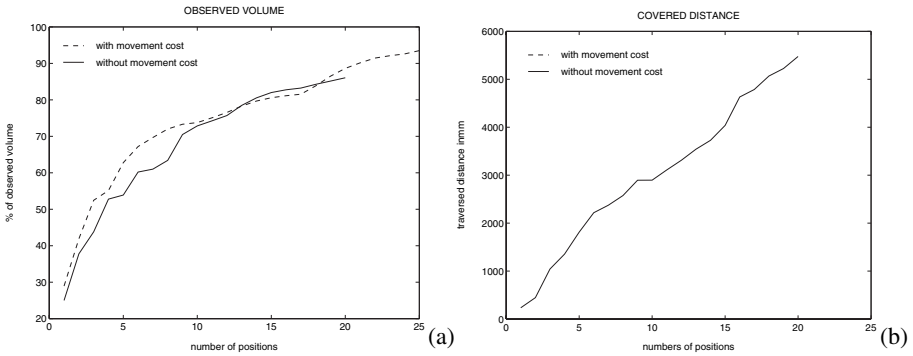
a result of local exploration described in the preceding section. The first displacements of the camera (see Figures 6.11b and 6.11c) make it possible to decrease the area of the non-explored scene. From position ϕ_{13} represented in Figure 6.11d, a new primitive is detected marking the beginning of a second phase of local exploration, which is completed in stage 24 (Figure 6.11f). The two polygons at the top of the scene are then reconstructed. A new total exploration brings the camera in ϕ_{25} (Figure 6.11g) where a segment pertaining to the last object (a telephone card) appears. After a last local exploration, allowing the reconstruction of 4 sides of the card, the camera is in position ϕ_{30} (Figure 6.11h). At this stage, 97% of space was observed thus ensuring a complete reconstruction of the scene. Figure 6.12 shows a 3D visualization of the scene such as it was reconstructed, as well as the trajectory of the camera during total exploration.

Presently, we analyze the influence of the coefficients α_i of equation (6.40) on the path of camera. In particular, we highlight the importance of considering the distance between two successive viewpoints in the energy function. The scene consists of a cylinder and a polygon, which were reconstructed during a phase of local exploration. The first strategy does not take into account the distance covered by the camera and thus is mainly based on a maximization of the zone discovered for each viewpoint (coefficient α_2 in equation (6.40) is zero). The second strategy takes into account this distance and thus tends to decrease the total distance covered by the camera. Figures 6.13b and 6.13c show the various trajectories carried out leading to a total exploration of the scene. Figure 6.14a shows the volume percentage of



(a) (b)

Figure 6.13. *Total exploration of the scene: trajectory of the robot without taking into account the displacement cost ($\alpha_2 = 0$) (a) and by taking into account the displacement cost ($\alpha_2 > 0$) (b)*



(a) (b)

Figure 6.14. *Percentage of observed scene (a) and distance covered by camera (b) according to the number of calculated viewpoints*

the scene observed after each new position of the camera. Figure 6.14b shows the cumulative distance covered by the camera for the two strategies. It will be noted that if the distance covered is not taken into account, the camera executes a trajectory in "bee flight" whereas such movements no longer occur if the distance is introduced in energy function.

Let us finally note that in this case the residual zones that were not observed (less than 3% of the total volume) correspond to occluded zones located under cylinder and rectangle (see Figure 6.15) or to zones located in the periphery of the scene. There are three solutions for this problem:

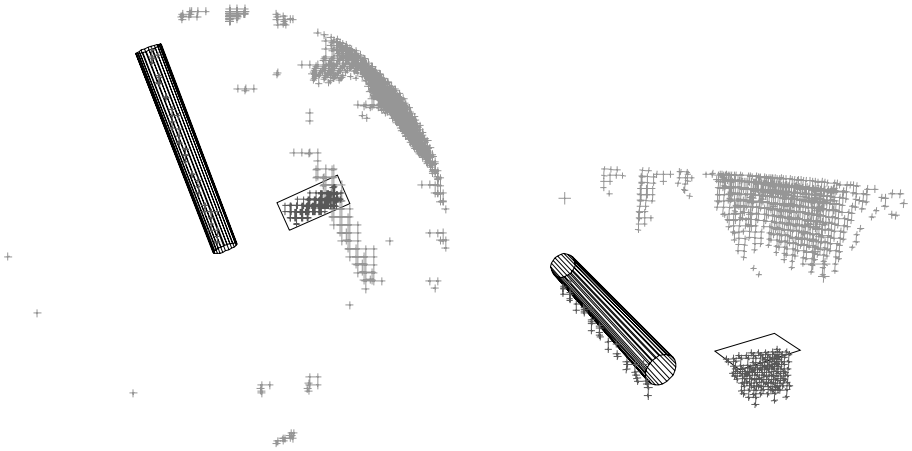


Figure 6.15. *Zones of the scene which remains to be observed (top view and front view)*

– initially, it is possible to no longer limit ourselves to displacements only on the surface of the sphere, but to be able to penetrate to its interior. Accordingly, it is advisable to verify that the viewpoint is accessible to the camera (i.e., whether it belongs to an already observed zone, which does not contain an object);

– we can then use the same algorithm of total exploration as earlier but by restricting ourselves to each small zone circumscribed in the residual non-observed zones. Then, the gain brought by a new viewpoint will no longer be negligible (see Figure 6.16).

– in addition, to solve the problems of occlusions, it also seems promising to be focused on the occluded segments and to then observe the zones occluded by these segments.

These different points are discussed (partly for the latter) in [MAR 96, MAR 99a].

6.5. Conclusion

In this chapter, we presented a method allowing the reconstruction of a 3D environment using a sequence of images acquired by a mobile camera. We described a reconstruction process allowing a precise and robust estimate of the parameters of a geometric primitive. Since this method is based on

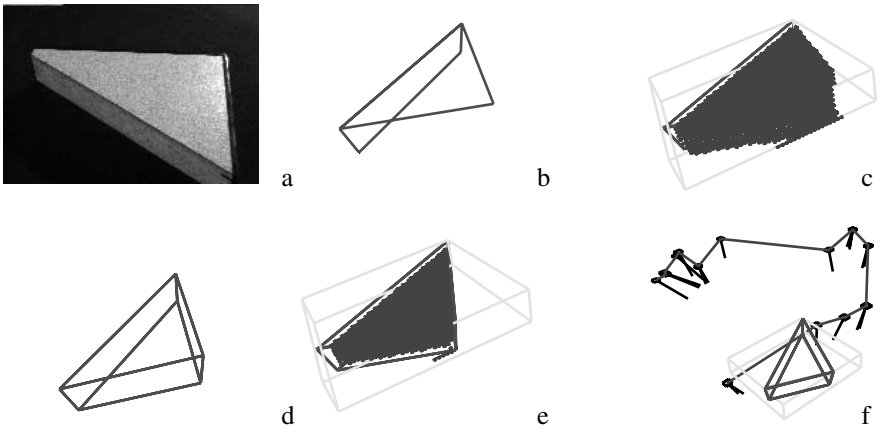


Figure 6.16. *Reconstruction of a polyhedron (a) image of the scene, (b) two polygons were reconstructed after a local exploration phase, (c) unobserved zones of the scene after this first phase resulting from occlusions, (d) Scene model at the end of reconstruction, (e) unobserved zones remaining at the end of total exploration (the unobserved zones are located inside the polygon), (f) trajectory of the camera*

particular movements of the camera, perceptive strategies making it possible to carry out a succession of reconstructions and ensuring the completeness of reconstruction were proposed. The experiments carried out on a robotic cell showed the validity of our approach and also its limitations: constraints on the movements of the camera imply a strong sequencing of reconstruction tasks. Hence, it will be useful to determine the optimal movements of the camera necessary for the simultaneous reconstruction of several primitives, and this without notable degradation in the reconstruction quality. The reconstruction of more complex objects is also considered. Then, it will no longer be a question of globally regarding the object as a parameterable geometric primitive, but of carrying out a local reconstruction of the object surface using yet another active approach there. Perception strategies will have to be implemented to ensure a complete reconstruction and to take into account the problems linked to occlusions or topological changes of the object.

6.6. Appendix: calculation of the interaction matrix

Here, we present the calculation of the interaction matrix $\mathbf{L}$ in the case of a point. The traditional equations of projected movement are then found; see [ESP 92] for the calculation of matrices associated with other geometric primitives.

Let $\mathbf{m}(x, y, z)$ be the position of a point in the reference pointer camera. Let $\mathbf{T} = (\mathbf{V}, \boldsymbol{\Omega}) = (V_x, V_y, V_z, \Omega_x, \Omega_y, \Omega_z)$ be the camera speed. The relation that links point movement $\dot{\mathbf{m}}$ to the camera speed $\mathbf{T}$ is given by:

$$\dot{\mathbf{m}} = -\mathbf{V} - \boldsymbol{\Omega} \times \mathbf{m} \Longleftrightarrow \begin{cases} \dot{x} = -V_x - \Omega_y z + \Omega_z y \\ \dot{y} = -V_y - \Omega_z x + \Omega_x z \\ \dot{z} = -V_z - \Omega_x y + \Omega_y x \end{cases} \tag{6.41}$$

From perspective equations $X = x/z$ and $Y = y/z$, by deriving in relation to time, we can write:

$$\begin{cases} \dfrac{dX}{dt} = \dfrac{\partial X}{\partial x}\dfrac{dx}{dt} + \dfrac{\partial X}{\partial z}\dfrac{dz}{dt} \\ \dfrac{dY}{dt} = \dfrac{\partial Y}{\partial y}\dfrac{dy}{dt} + \dfrac{\partial Y}{\partial z}\dfrac{dz}{dt} \end{cases} \Longleftrightarrow \begin{cases} \dot{X} = \dfrac{\dot{x}}{z} - \dfrac{x}{z^2}\dot{z} \\ \dot{Y} = \dfrac{\dot{y}}{z} - \dfrac{y}{z^2}\dot{z} \end{cases} \tag{6.42}$$

By replacing in (6.42) the values of $(\dot{x}, \dot{y}, \dot{z})$ calculated by equation (6.41) and by simplifying, we finally obtain:

$$\begin{cases} \dot{X} = -\dfrac{1}{z}V_x + \dfrac{X}{z}V_z + XY\Omega_x - \big(1 + X^2\big)\Omega_y + Y\Omega_z \\ \dot{Y} = -\dfrac{1}{z}V_y + \dfrac{Y}{z}V_z + \big(1 + Y^2\big)\Omega_x - XY\Omega_y - X\Omega_z \end{cases}$$

We find the relation $\dot{\mathbf{M}} = \mathbf{L}\mathbf{T}_\mathbf{c}$ with $\mathbf{L}$ defined by equation (6.12).

6.7. Bibliography

[ADI 85] ADIV G., "Determining Three-Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, p. 384–401, 1985.

[ADI 89] ADIV G., "Inherent ambiguities in recovering 3D motion and structure from a noisy flow field", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, p. 477–489, 1989.

[AGG 87] AGGARWAL J. and WANG Y., "Analysis of a sequence of images using point and line correspondences", *IEEE Int. Conf. on Robotics and Automation*, vol. 2, Raleigh, North Carolina, p. 1275–1280, March 1987.

[AGG 88] AGGARWAL J. and NANDHAKUMAR N., "On the computation of motion from sequences of images: a review", *Proceedings of the IEEE*, vol. 76, no. 8, p. 917–935, 1988.

[ALO 87] ALOIMONOS Y., WEISS I. and BANDOPADHAY A., "Active Vision", *Int. Journal of Computer Vision*, vol. 1, no. 4, p. 333–356, 1987.

[ALO 90] ALOIMONOS Y., "Purposive and qualitative active vision", *IAPR Int. Conf. on Pattern Recognition, ICPR'90*, vol. 1, Atlantic City, New Jersey, p. 346–360, June 1990.

[ARB 91] ARBOGAST E., Modélisation automatique d'objets non polyédriques par observation monoculaire, PhD Thesis, Institut National Polytechnique de Grenoble, July 1991.

[BAJ 88] BAJCSY R., "Active Perception", *Proceedings of the IEEE*, vol. 76, no. 8, p. 996–1005, 1988.

[BAL 91] BALLARD D., "Animate vision", *Artificial Intelligence*, vol. 48, no. 1, p. 57–86, February 1991.

[BLA 90] BLAKE A. and CIPOLLA R., "Robust Estimation of Surface Curvature from Deformation of Apparent Contours", *1st European Conf. on Computer Vision, ECCV'90*, Antibes, France, p. 465–474, April 1990.

[BOU 89] BOUKARRI B., Reconstruction 3D récursive de scènes structurées au moyen d'une caméra mobile. Application à la robotique, PhD Thesis, University of Orsay, France, October 1989.

[BOU 93] BOUKIR S., Reconstruction 3D d'un environnement statique par vision active, PhD Thesis, University of Rennes 1, IRISA, October 1993.

[CHA 89] CHAUMETTE F. and RIVES P., "Modélisation et calibration d'une caméra", *7ème congrès AFCET Reconnaissance des formes et Intelligence artificielle, RFIA'89*, vol. 1, Paris, p. 527–536, December 1989.

[CHA 90] CHAUMETTE F., La relation vision-commande: théorie et application à des tâches robotiques, PhD Thesis, University of Rennes I, IRISA, July 1990.

[CHA 96] CHAUMETTE F., BOUKIR S., BOUTHEMY P. and JUVIN D., "Structure from controlled motion", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 5, p. 492–504, May 1996.

[CHI 89] CHIEN C. and AGGARWAL J., "Model Construction and Shape Recognition from Occluding Contour", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 4, p. 372–389, 1989.

[CON 85] CONNOLLY C., "The Determination of Next Best Views", *IEEE Int. Conf. on Robotics and Automation*, St Louis, Missouri, p. 432–435, March 1985.

[COW 88] COWAN C. and KOVESI P., "Automatic Sensor Placement from Vision task Requirements", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, p. 407–416, 1988.

[CRO 92] CROWLEY J., STELMASZYK P. and PUGET P., "Measurement and integration of 3D structures by tracking edge lines", *Int. Journal of Computer Vision*, vol. 8, no. 1, p. 29–52, 1992.

[DJI 96] DJIAN D., PROBERT P. and RIVES P., "Reconnaissance de modèles géométriques simples à l'aide de réseaux bayésiens", *10ème Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, RFIA'96*, vol. 1, Rennes, France, p. 396–404, January 1996.

[ESP 87] ESPIAU B. and RIVES P., "Closed-Loop Recursive Estimation of 3D Features for a Mobile Vision System", *IEEE Int. Conf. on Robotics and Automation*, vol. 3, Raleigh, North Carolina, p. 1436–1443, April 1987.

[ESP 92] ESPIAU B., CHAUMETTE F. and RIVES P., "A new approach to visual servoing in robotics", *IEEE Trans. on Robotics and Automation*, vol. 8, no. 3, p. 313–326, 1992.

[HUA 91] HUANG L. and ALOIMONOS J., "Relative depth from motion using normal flow: an active and purposive solution", *IEEE Workshop on Visual Motion*, Princeton, New Jersey, p. 196–204, October 1991.

[HUT 96] HUTCHINSON S., HAGER G. and CORKE P., "A tutorial on Visual Servo Control", *IEEE Trans. on Robotics and Automation*, vol. 12, no. 5, p. 651–670, 1996.

[MAR 96] MARCHAND E., Stratégies de perception par vision active pour la reconstruction et l'exploration de scènes statiques, PhD Thesis, University of Rennes 1, IRISA, No. 1589, June 1996.

[MAR 99a] MARCHAND E. and CHAUMETTE F., "Active vision for complete scene reconstruction and exploration", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, p. 65–72, 1999.

[MAR 99b] MARCHAND E. and CHAUMETTE F., "An autonomous active vision system for complete and accurate 3D scene reconstruction", *International Journal of Computer Vision*, vol. 32, no. 3, p. 171–194, 1999.

[MAT 89] MATTHIES L., KANADE T. and SZELISKI R., "Kalman filter-based algorithms for estimating depth from image sequences", *Int. Journal of Computer Vision*, vol. 3, no. 3, p. 209–236, 1989.

[MAV 93] MAVER J. and BAJCSY R., "Occlusions as a guide for planning the next view", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 5, p. 417–433, 1993.

[NEG 87] NEGAHDARIPOUR S. and HORN B., "Direct passive navigation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, p. 168–176, 1987.

[PEA 88] PEARL J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publisher Inc., San Mateo, California, 1988.

[RIV 87] RIVES P. and ESPIAU B., Estimation récursive de primitives 3D au moyen d'une caméra mobile, Report no. 652, INRIA-IRISA, March 1987.

[SAM 91] SAMSON C., LE BORGNE M. and ESPIAU B., *Robot Control: The Task Function Approach*, Clarendon Press, Oxford, 1991.

[SAN 86] SANDINI G. and TISTARELLI M., "Recovery of depth information: camera motion as an integration to stereo", *IEEE Workshop on Motion*, Niawah, Iceland, p. 39–43, May 1986.

[SAN 90] SANDINI G. and TISTARELLI M., "Active Tracking Strategy for Monocular Depth Inference over Multiple Frames", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, p. 13–27, 1990.

[SUB 87] SUBBARAO M., "Solution and uniqueness of image flow equations for rigid curved surfaces in motion", *IEEE Int. Conf. on Computer Vision, ICCV'87*, London, p. 687–692, June 1987.

[SWA 93] SWAIN M. and STRICKER M., "Promising Direction in Active Vision", *Int. Journal of Computer Vision*, vol. 11, no. 2, p. 109–127, 1993.

[TAR 95] TARABANIS K., TSAI R. and ALLEN P., "The MVP Sensor Planning System for Robotic Vision Tasks", *IEEE Trans. on Robotics and Automation*, vol. 11, no. 1, p. 72–85, 1995.

[VER 90] VERNON D. and TISTARELLI M., "Using camera motion to estimate range for robotic part manipulation", *IEEE Trans. on Robotics and Automation*, vol. 6, no. 5, p. 509–521, 1990.

[VIA 92] VIALA M., Contribution à la reconstruction de scènes constituées d'objets cylindriques et polyédriques à partir d'une séquence d'images acquises par une caméra en mouvement, PhD Thesis, University of Orsay, November 1992.

[WAX 87] WAXMAN A., PARSI B. and SUBBARAO M., "Closed-form Solutions to Image Flow Equations for 3D Structure and Motion", *Int. Journal of Computer Vision*, vol. 1, no. 3, p. 239–258, 1987.

[WEL 89] WELLS W., "Visual estimation of 3D lines segments from motion. A mobile robot vision system", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 5, no. 6, p. 820–825, 1989.

[WEN 92] WENG J., HUANG T. and AHUJA N., "Motion and structure from line correspondences closed-form solution, uniqueness, and optimization", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, p. 318–336, 1992.

[WHA 94] WHAITE P. and FERRIE F., "Autonomous exploration: Driven by uncertainty", *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, CVPR'94, Seattle, Washington, p. 339–346, June 1994.

[WIX 94] WIXSON L., "Viewpoint Selection for Visual Search", *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'94*, Seattle, Washington, p. 800–805, June 1994.

[XIE 89a] XIE M., Contribution à la vision dynamique: reconstruction d'objets 3D polyédriques par une caméra mobile, PhD Thesis, University of Rennes 1, IRISA, June 1989.

[XIE 89b] XIE M. and RIVES P., "Toward dynamic vision", *IEEE Workshop on Interpretation of 3D Scenes*, Austin, Texas, p. 91–99, November 1989.

[ZHA 94] ZHAO C. and MOHR R., "Relative 3D regularized B-Splines surface reconstruction through image sequences", *3rd European Conf. on Computer Vision, ECCV'94*, vol. 2, Stockholm, Sweden, p. 417–426, May 1994.

[ZHA 95] ZHANG Z., "Estimating Motion and Structure from Correspondences of Line Segments between Two Perspective Images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, p. 1129–1139, 1995.

Part 4

This page intentionally left blank

Chapter 7

Shape Recognition in Images

7.1. Introduction

We can classify recognition systems into two major classes. The first class is a group of general recognition systems and the second class is specialized systems. These two classes differ from each other in the definition of similarity between images.

The general systems are dedicated to non-specific search operations: we do not seek a precise image, but a category of images, which in general is not defined formally: are there city images or sunsets in the base? Requests such as these are common in search systems on the web and in systems open to the general public. This type of system frequently uses global descriptors of images, based on the form, color, or texture, which make it possible to define concepts of approximate similarity.

Vision or robotics applications often search for a specific object, which requires a highly precise definition of sought similarity. We must then choose specialized systems. An example is the recognition of objects. It is a classical problem of vision, which is most of the time carried out using an exposure or reconstruction algorithm; we try such a calculation with all the basic models, and the one that gives acceptable results indicates the recognized object.

Chapter written by Patrick GROS and Cordelia SCHMID.

Such a process is clearly complex in computing time; it obliges to consider all the models successively and proceeds with a complex calculation for each one of them. It is important to be able to uncouple recognition from its application and to aim at a sub-linearity for the number of models stored in the base. This is what we propose in the following sections.

After state of the art recognition systems, we present two methods whose purpose is to recognize images, which represent the same object or, more precisely, the same aspect of an object, but under different conditions. These differences can relate to viewpoint, light conditions, composition of image, or the position of various elements of the image. The two methods presented are based on a common principle: use of quasi-invariants associated with local descriptors. The description of this principle and its advantages is followed by the presentation of two methods. They were tested with bases that did not require the storage of descriptors on disk. In the last section, we consider the problems arising from the usage of external disks like memory to store the data.

7.2. State of the art

In this section, we present methods that find the 3D image or the object corresponding to the requested image. This refers to the same image or object seen under different conditions. We classify the existing methods in other works into image searching methods, i.e., 2D objects and methods of 3D object recognition. In the case of 3D objects, the additional difficulty is that there are no invariants [BUR 90].

7.2.1. *Searching images based on photometric data*

The first group of image searching methods uses information on the brightness of an object, i.e., its aspect, like signature. In this context, Swain [SWA 91] used color. He showed that histograms of color can be used to map objects. The largest defect of this approach is its lack of robustness with respect to changes in luminosity. Several authors improved the performances of this technique by introducing fairly robust measurements to changes in luminosity. Thus, Funt [FUN 95] proposed to use distribution of color ratios and showed that these ratios provide a color constant for an object. Slater [SLA 96] showed that moments of color distribution are invariant to a change of luminosity in the assumption of a linear model of luminous reflection. Nayar [NAY 93] and Nagao [NAG 95] used photometric invariants based on

reflection ratios. It is also possible to use histograms of local filters, which were used by Schiele [SCH 96a].

Another approach of image search from photometric information is that by Turk [TUR 91]. His method uses a large collection of images, which is broken up into principal components. He used this method to find faces. Components corresponding to the greatest eigenvalues represent fuzzy shapes of faces. This approach was applied by Murase [MUR 95] to recognize images of unspecified objects. The advantages of this method are its speed, its generality and its robustness to small occlusions. On the other hand, it requires centering of images and it is not robust to image transformations or significant occlusions.

Lades [LAD 93], Rao [RAO 95], Wu [WU 95] and Schmid [SCH 96d] used local measurements based on the image of gray levels. The signal is locally characterized by adjustable filters in case of Rao, by Gabor transforms in case of Wu and Lades, and by differential invariants in the case of Schmid. For Lades, Rao and Wu, these filters are calculated on a grid, which is centered on the object by a simple calculation of the center of the object. Rao uses a circular grid while Wu and Lades use a rectangular grid. It is difficult to position the grid as soon as the object is presented in front of a complex background. In addition these methods do not allow us to recognize an object from images of a portion of this object. This is due to the fact that the grid cannot be positioned if only part of the object is given. On the other hand, the use of a grid implies that some of its points represent little of the object. Rao [RAO 95] thus proposes to use characteristic vectors comprising up to 45 components. With this purpose, this characterization is calculated in a multi-scale context. To recognize an object in the presence of scaling, it is necessary to relocate characteristic vectors in order to find the subsets of components, which correspond to each other. Mapping is then done on a part of the characterization and the results degrade significantly. To overcome these disadvantages, Schmid calculated her descriptors in characteristic points by detecting interest points [SCH 00]. In addition, Schmid [SCH 97] showed that the use of semi-local constraints increases the performances of recognition. These constraints locally represent the geometric feature of key points.

7.2.2. *Search for images based on geometric data*

A second group of methods that looks for images uses geometric data such as segments, junctions and ellipses. Such data are extracted beforehand from

images and research is carried out using only these data. These methods thus depend on the given symbolic data even if they use these data to calculate numerical sizes. A certain number of approaches are based on the following paradigm: calculation of assumption and verification. During the first phase, characteristics are extracted from the sought image and then they are associated with the characteristics of 2D model candidates. The exhaustive search for all the possible candidates generates a polynomial calculation cost. The major contribution of the various search systems was to control and decrease the complexity of the phase of matching. For example, Ayache and Faugeras [AYA 86] use a recursive evaluation of assumptions. Lamdan [LAM 88] proposed to use methods of indexing and hashing to obtain significant acceleration. In case of indexing, the association of characteristics and the search for a basic model are replaced by a *look-up table* mechanism. In a similar context, Rothwell *et al*. [ROT 93] used projective invariants as indexing elements. In the case of 2D objects, such invariants can be calculated for any object.

Other methods of image search are based on Hough transform. They choose a model by searching for a point of accumulation in the transformations space (see, for example, Ballard [BAL 81]). Grimson [GRI 90], however, showed that such an approach is not very robust to the noise of an image. Indeed, in the presence of noise, it is not possible to distinguish between two different models. To solve this problem, Gros [GRO 95] uses invariants with similarities and only votes in Hough space if these invariants match. This reduces the number of votes in Hough space and thus makes the distinction of various candidates possible.

7.2.3. *Recognition using a 3D geometric model*

Geometric models of a 3D object are based on geometric characteristics such as edges, junctions, ellipses, surfaces and volumes. A 3D model of the object to be recognized is established with the help of these characteristics. These models are often based on CAD[1] models. Among the existing CAD models, iron-wire modeling consists of a list of junctions and connections between these junctions. As for constructive solid geometry, it models an object by collective operations from voluminous primitives. Spatial occupation representation describes the volume occupied by a 3D object.

1. CAD: computer-aided design. This design allows the automation of the design process and manufacturing.

Representation by surface envelope (B-Rep.) models an object by parts of a surface. Besl *et al.* [BES 85] and Chin *et al.* [CHI 86] present a state of the art of geometric models and object recognition systems based on such modeling. These systems map a 2D image and a 3D geometric model. They can be divided into methods based on a mechanism of prediction/verification, on Hough transform or the use of an interpretation tree. Systems based on a mechanism of prediction/verification relate some characteristics of the model with some characteristics of the image. This allows an initial calculation of the model-image transformation. This transformation is used to project other characteristics of the model on the image and then to verify the correspondence with the characteristics of the image. In case of polyhedric objects, Huttenlocher *et al.* [HUT 90] and Lowe *et al.* [LOW 87] developed such an approach. Bolles *et al.* [BOL 86] and Faugeras *et al.* [FAU 86] developed similar approaches in the case of depth images. However, the exhaustive search for all the models existing in the base generates an exponential calculation cost. The major contribution of various recognition systems was to control and decrease the complexity of the pairing phase. For example, Bolles *et al.* [BOL 86] use a search tree.

Kriegman *et al.* [KRI 90] presented an approach for curved models. An implicit representation of these curves in the image is parametrized by the position and orientation of the object. The calculation of these parameters is reduced to the adjustment problem between the theoretical contour and contour points in the image. Verification is carried out by comparing the adjustment errors for various models.

Other works, such as Mundy *et al.* [MUN 90], calculate the transformations between the detected primitive images and the primitives of CAD data. Then, they use Hough transform in the parameter space of these transformations to find a point of accumulation. This point of accumulation simultaneously gives the corresponding model and the transformation between the image and the model.

The trees of interpretation contain all the possible combinations between primitives detected in the image and primitives of the model. These combinations are organized in a tree, for example, the first level of the tree contains combinations between an extracted primitive and the primitives of the model. This tree creates an enormous space for research. It is thus essential to introduce additional constraints, which ignore the exhaustive course of the tree. Brooks [BRO 83], for example, developed such an approach. Besides, his approach makes it possible to use constraints with a

confidence interval in the case of generic objects. Grimson *et al.* [GRI 87, GRI 89] used interpretation trees in the case of depth images. In the case of gray level images, their approach is limited to 2D objects.

Lastly, for specific classes of 3D objects, it is possible to characterize an object by way of a single view because there are invariants. As Zisserman [ZIS 95] showed, there are projective invariants for certain classes of 3D object, such as polyhedrons, revolving surfaces, tubes and symmetric objects. These invariants can thus be extracted from a view of such an object and it is then possible to recognize the object from any point of view.

7.2.4. *Recognition using a set of images*

The idea of approaches using a set of images is to no longer use the theoretical model located far away from images, but to use characteristic images to represent an object. These images in the continuation of this chapter are called model-images. To compare the image of a 3D object with these model-images, we can apply one of the techniques of image search (2D object) presented earlier.

Nayar *et al.* [NAY 93] and Schmid *et al.* [SCH 96c] modeled a 3D object from 2D gray level images. As model-images, they used a set of images regularly spaced on a circle or a sphere. From these images, search methods cited previously make it possible to recognize the object and also its position and posture.

Gros [GRO 95] determined the necessary model-images to represent a 3D object using a clustering algorithm. In the same way, Gdalyahu [GDA 96] used contours of model-images to represent an object. This approach underlines the need to choose "good" views to represent the object. Such views are intrinsically more stable and more representative of the object.

7.3. Principle of local quasi-invariants

Making use of local information provides natural robustness to partial occlusions and translations in the image. This information is either obtained by cutting images *a priori* (cutting into 9 sub-images, regular sampling of image, etc.), or using segmentation tools. This second solution is finer, in the sense that it can better take into account all small translations, but it is clearly dependent on the quality of the extractor used. The principal quality of such

a detector is repeatability, i.e., the aptitude to extract the same elements in images taken under different conditions. Indeed, if we start from elements that are so precise, but different, the continuation of the recognition process cannot succeed.

The concept of quasi-invariant was introduced in vision in a theoretical manner by Binford [BIN 93], and in an independent and pragmatic way by Ben-Arie [BEN 90b, BEN 90a]. The quasi-invariants theory makes it possible to define first order invariants in the case of perspective transformations. To locate these quasi-invariants, let us commence by presenting the invariants [GRO 93].

Let us take a quantity f, which we can calculate from an unspecified point $\mathbf{x}$ of an image. It is said that f is invariant for a set of transformations T if:

$$\forall t \in T \quad f(\mathbf{x}) = f(t(\mathbf{x})) \tag{7.1}$$

In general, the set of transformations is a Lie group, i.e., a group parametrized by coefficients (a_i). In addition, using the coordinates of $\mathbf{x}$, we can note:

$$t(\mathbf{x}) = \Phi\big(x_1, \ldots, x_n; a_1, \ldots, a_m\big) = \begin{pmatrix} \phi_1\big(x_1, \ldots, x_n; a_1, \ldots, a_m\big) \\ \vdots \\ \phi_n\big(x_1, \ldots, x_n; a_1, \ldots, a_m\big) \end{pmatrix} \tag{7.2}$$

With this notation, the fact that f is an invariant is translated by:

$$\begin{gathered} \forall a_1 \ldots a_m, a'_1 \ldots a'_m \\ f\big(\Phi\big(x_1 \ldots x_n; a_1 \ldots a_m\big)\big) = f\big(\Phi\big(x_1 \ldots x_n; a'_1 \ldots a'_m\big)\big) \end{gathered} \tag{7.3}$$

In practice, this invariance concept often appears too strong and constraining. In fact, we insist that quantity $f(t(\mathbf{x}))$ be the same for all elements of T, including those which correspond to degenerated transformations: with nearly zero ratio similarities plane view "on the section". Considering these transformations makes it necessary to manipulate complex quantities (biratios, for example), which are sensitive to noise, whereas the corresponding configurations cannot generate calculations because according to the discretization of images, the points that we must make use of are generally indistinguishable or known with almost zero relative precisions.

The constraint imposed on quasi-invariants is weaker: it is necessary that there is a transformation t of parameters a such that the size measured in image $f(t(\mathbf{x}))$ is equal to the size measured in the scene, and that the variation of $f(t(\mathbf{x}))$ as compared to t is infinitely small in first order, or another formulation, the derivative of $f(t(\mathbf{x}))$ as compared to parameters of t must be zero:

$$\begin{aligned}&\forall j \in \{1 \dots m\}\\ &\frac{\partial f \circ \Phi}{\partial a_j}(x_1 \dots x_n; a_1 \dots a_m)\\ &= \sum_{i=1}^{n} \left[\partial_i f\big(\Phi(x_1 \dots x_n; a_1 \dots a_m)\big) \frac{\partial \phi_i}{\partial a_j}(x_1 \dots x_n; a_1 \dots a_m)\right] = 0.\end{aligned} \tag{7.4}$$

If the second derivatives are also zero, then we talk about strong quasi-invariant. As we see it, this constraint is only local: instead of imposing nullity of all non-constant terms of Taylor's development, we only impose that of the term of order 1, or that of terms of order 1 and 2. For example, the angle between two segments is a quasi-invariant and the ratio of three aligned points is a strong quasi-invariant.

7.4. Photometric approach

In the following section, we present a method based on local photometric information [SCH 96b]. An image is characterized by a set of local invariants of gray levels calculated at the key points. After a presentation of the detector of key points and invariants, we explain how to compare these invariants. Then, we show that using an algorithm of vote and semi-local constraints makes recognition robust and that a mechanism of indexing allows an effective recognition of a base containing more than 1,000 images. The experimental results presented show the possibility of recognizing in cases of partial visibility, image transformations of the similar type, changes of luminosity, additional characteristics and light perspective deformations. Lastly, the extensions of this recognition method are briefly presented.

7.4.1. *Key points*

Calculating image descriptors for each pixel of the image induces too great a quantity of information of which a majority is not very significant. Using key points allows the selection of points with significant informative contents: in

these points, the signal changes in two directions at a time. Moreover, key points are local characteristics and are therefore stable in partial visibility (occlusions).

A large variety of key point detectors exist in other works. In the context of recognition, the extracted points must be repeatable and must present significant informative contents. A comparison of several detectors under varied and changing conditions [SCH 00] showed that the detector satisfying the preceding criteria most was the Harris detector [HAR 88].

The essential idea of this detector is to use the autocorrelation function to determine the positions where the signal changes in two directions at a time. To avoid discrete calculations, the matrix related to the autocorrelation function is calculated from the primary derivatives of the signal integrated on a neighborhood. The eigenvalues of this matrix correspond to the principal curves of the autocorrelation function. Two significant eigenvalues indicate the presence of a key point.

Figure 7.1 presents the key points detected for the same scene in the case of a rotation. The repeatability rate is 92%, i.e., 92% of the points detected in the first image are detected in the second. Experiments with images taken under different conditions showed that the average rate of repeatability is about 90%. Moreover, 50% repeatability rate is sufficient if we use a robust recognition method.

7.4.2. *Differential invariants of gray levels*

Local characterization is based on differential invariants of gray levels, invariants in image rotations and translations. Using a multi-scale approach

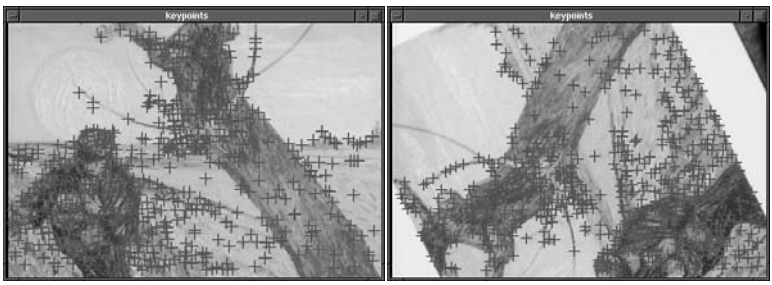


Figure 7.1. *Example of key points detected for the same scene in the case of a rotation. Rotation between the left image and the image on the right is of 155 degrees. Repeatability rate is 92%*

helps to make this characterization robust to changes in scale, and therefore robust to matching. These invariants are thus quasi-invariants for a perspective transformation [BIN 93]. Lastly, the characterization can be made robust to illumination changes.

An image in the neighborhood of a point can be described by the set of its derivatives. These derivatives can be calculated in a stable way using convolutions with Gaussian derivatives.

For an image I and a scale factor σ, the "local jet" [KoE 87] of order N in a point $\vec{x} = (x_1, x_2)$ is defined by:

$$J^N[I](\vec{x}, \sigma) = \left\{ L_{i_1 \ldots i_n}(\vec{x}, \sigma) \mid (\vec{x}, \sigma) \in I \times \mathbb{R}^+;\ n = 0, \ldots, N \right\}$$

with $L_{i_1 \ldots i_n}(\vec{x}, \sigma)$ convolution of image I with Gaussian derivatives $G_{i_1 \ldots i_n}(\sigma)$ where $i_k \in \{x_1, x_2\}$.

The coefficient σ of the Gaussian function determines the quantity of smoothing. This coefficient also corresponds to the definition of scale space and it is significant in a multi-scale approach. Hereafter, the *size* of the Gaussian will refer to σ.

The calculation of differential invariants from "local jet" [KoE 87, HAA 94] makes it possible to obtain an invariance in image rotations. In this chapter, a complete set of invariants up to order three has been used. Equation (7.5) gives the formula in tensorial notation – this notation uses Einstein's convention of addition. We can notice that the first element represents the average illumination, the second represents the square of the magnitude of the gradient and the fourth represents the Laplacian

$$\mathcal{D}(\vec{x}, \sigma) = \begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij} \left(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l \right) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{bmatrix} \tag{7.5}$$

with $L_{i_1 \ldots i_n}$ calculated elements of "local jet" at position $\vec{x}$ in scale σ. ε_{ij} is the Epsilon anti-symmetric 2D tensor defined by $\varepsilon_{12} = -\varepsilon_{21} = 1$ and $\varepsilon_{11} = \varepsilon_{22} = 0$.

To obtain robustness in scaling, a multi-scale approach must be used [LIN 94b, WIT 83] because there are no invariants in scale in a discrete representation. If there is change in scale α between two images I_1 and I_2, the derivatives between these two images are related by:

$$I_1(\vec{x}) \star G_{i_1 \ldots i_n}(\sigma) = \alpha^n I_2(\vec{u}) \star G_{i_1 \ldots i_n}(\sigma\alpha) \tag{7.6}$$

where $\vec{x}$ and $\vec{u}$ are the positions in two images ($\alpha\vec{x} = \vec{u}$) and $G_{i_1 \ldots i_n}$ the Gaussian derivatives of order n.

Equation (7.6) shows that the size of the Gaussian must be adjusted to take into account scaling. This implies that the reference size of calculation varies and induces a multi-scale approach. Since it is impossible to calculate the invariants in all the scales, it is necessary to discretize the scale space. Experiments showed that a calculation of mapping using differential invariants is tolerant to a scale variation of 20%. To be robust in scaling up to a scale factor of 2, it is sufficient to use σ between 0.48 and 2.07. For example, good results were obtained with the following discretization: 0.48, 0.58, 0.69, 0.83, 1.00, 1.20, 1.44, 1.73, 2.07.

Characterization must also be robust in illumination change. There are several possibilities to model a variation in luminosity, for example, by a translation, an affine transformation, or a monotonous transformation. Here, we have retained modeling by an affine transformation. Any quotient of two derivatives is invariant in such a transformation. Thus, there are various ways of making the descriptor $\vec{\mathcal{D}}$ invariant in an affine transformation. We chose to divide by the adequate power of the gradient magnitude. Since it is also necessary to eliminate illumination, our descriptor is of dimension 7.

7.4.3. *Comparison of descriptors with Mahalanobis distance*

It is traditional to model the uncertainty of descriptors (invariant vectors) by a Gaussian distribution and to use Mahalanobis distance d_M to compare them. This distance takes into account the differences in the magnitude nature of various components as well as covariance $\mathbf{\Lambda}$ of these components. Given two descriptors $\vec{a}$ and $\vec{b}$, Mahalanobis distance is defined by: $d_M(\vec{b}, \vec{a}) = \sqrt{(\vec{b} - \vec{a})^T \mathbf{\Lambda}^{-1} (\vec{b} - \vec{a})}$.

The square of Mahalanobis distance follows a distribution law of χ^2. Since the square root is a bijection of $\mathbb{R}^+$ in $\mathbb{R}^+$, it is possible to use the fractile table of this distribution for thresholding the distance. Indeed, the fractile determines the threshold t_k for which $P(d_M^2 \leq t_k) = k\%$.

In order to obtain accurate results using this distance, it is important to use a representative covariance matrix, i.e., one which takes into account the noise of the signal, illumination variations and inaccuracy in the position of key points. It is theoretically impossible to calculate this matrix by taking into account realistic assumptions. To obtain an estimate, a set of key points is followed in sequences of images. The descriptors (invariant vectors) calculated for each series of points make it possible to obtain a covariance matrix. The average of these matrices is then used for comparison.

Usage of Mahalanobis distance is difficult when we wish to carry out a fast indexing method. However, it is possible to carry out a change of reference mark in order to use the Euclidean distance d_E. The covariance matrix is a symmetric real and semi-defined positive matrix. Thus, it can be divided in the following way: $\mathbf{\Lambda}^{-1} = \mathbf{P}^T\mathbf{D}\mathbf{P}$ with $\mathbf{P}$ an orthogonal matrix and $\mathbf{D}$ a diagonal matrix. Then we get $d_M(\vec{a}, \vec{b}) = d_E(\sqrt{\mathbf{D}}\mathbf{P}\vec{a}, \sqrt{\mathbf{D}}\mathbf{P}\vec{b})$.

7.4.4. *Voting algorithm*

The voting algorithm enables us to select the image of the base which is most similar to the request image. The idea of the algorithm is to compare the descriptors of the request image with basic descriptors and to add the number of times that an image of the base is selected.

An image base contains a set $\{M_k\}$ of model-images. Each model-image M_k is defined by descriptors calculated in key points of this model-image. During the training process – or storage in the base – each descriptor is added to the base with a link to the model k from which it was calculated. Formally, the simplest image base consists of a table of couples $(\mathcal{D}_j, k)$.

The process of recognition consists of finding the model-image $M_{\hat{k}}$ which corresponds to request image I, i.e., the most similar image. For the request image, a set of descriptors $\{\mathcal{D}_l\}$ is calculated at key points extracted on this image. These descriptors are then compared with basic vectors $\{\mathcal{D}_j\}$ by calculating Mahalanobis distance $d_M(\mathcal{D}_l, \mathcal{D}_j)$. If this distance is below threshold t according to the distribution of χ^2, the corresponding model acquires a vote.

The sum obtained is stored in a vector $T(k)$. The model with the greatest number of votes is regarded as the best matching one; the request image corresponds to model $M_{\hat{k}}$ for which $\hat{k} = \arg\max_k T(k)$.

Figure 7.2a shows an example of a $T(k)$ vector in the form of a histogram. Model-image 0 is correctly recognized. However, other images have got scores – or number of votes – of the same order.

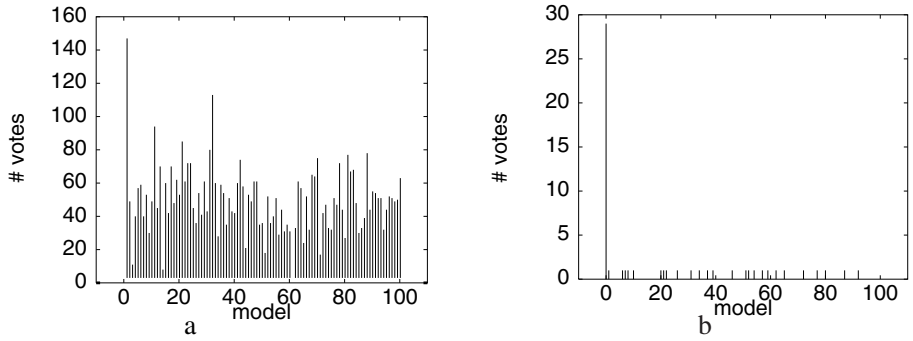


Figure 7.2. *(a) Result of the voting algorithm: the number of votes is posted for each model image of the base. Image 0 is correctly recognized. (b) Result obtained by adding semi-local constraints. The semi-local constraints decrease the probability of false votes. Image 0 is recognized in a much more discriminating manner than in the case of (a)*

7.4.5. *Semi-local constraints*

A descriptor of the request image can vote for several model-images. A great number of model-images or several very similar model-images increase the probability that a descriptor votes for several models. To solve this problem, we used semi-local constraints. These constraints impose for each couple of paired points that the descriptors of neighboring points must match and that their geometric configuration is compatible (see Figure 7.3). Similar constraints were used by [VAN 91, ZHA 95].

For each basic key point and request image, the closest points p in the image, in the context of Euclidean distance, are selected. Imposing that all closest neighbors p are mapped correctly is equivalent to impose that there is no error in detecting the key point (and that there is a 100% repeatability rate). It is in fact sufficient to impose that at least 50% of the neighbors match. An additional verification is possible by adding geometric constraints. For example, angles defined by neighboring mapping points must match. Indeed,

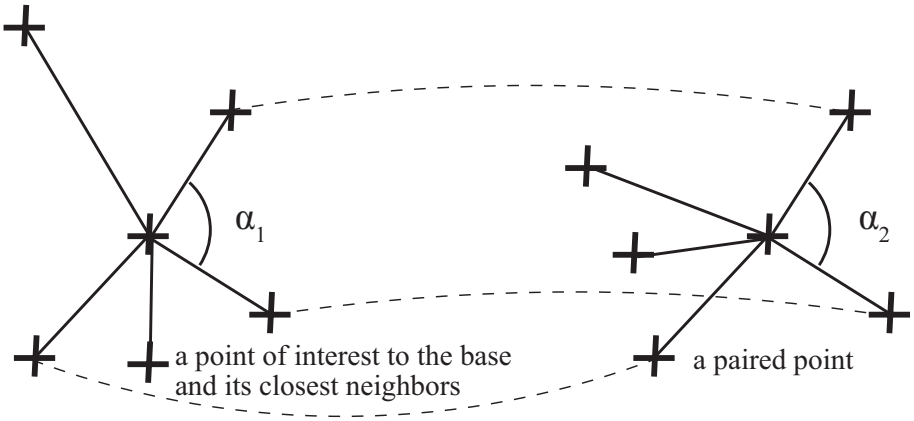


Figure 7.3. *Example of semi-local constraints: neighbors of points as well as angles which they define must correspond exactly*

if it is assumed that the transformation between two images can locally be approximated by a similarity – which is exact in the first order – the angles must be locally consistent. Figure 7.3 illustrates this constraint with angles α_1 and α_2.

Figure 7.2 presents an example on the use of semi-local constraints. The image on the right-hand side shows reduction in the number of votes when constraints are applied to the example on the left. The score of the correctly recognized object (model 0) is definitely more discriminating.

7.4.6. *Multi-dimensional indexing*

Without indexing, the complexity of the voting algorithm is in the order of $l \times N$, where l is the number of descriptors in the request image and N is the total number of descriptors in the image base. Since N is large (about 150,000 in our tests), it is necessary to use an adequate structuring of the stored data.

The search structures were largely studied. A state of the art of all tree structures, which allow a fast and/or compact access to data is presented in [SAM 84]. The structure presented hereafter can be seen like an alternative of a kD tree.

In this structure, each dimension of space is sequentially considered. The access to data in a dimension is realized by one-dimensional samples of fixed

size. Matching samples and their neighbors can be reached directly. An access to neighbors is necessary to take into account uncertainties in measurements. A sample is extended in a second dimension if the number of descriptors that it contains is superior to the threshold given. Thus, the data structure can be seen as a tree whose depth is at most equal to the number of dimensions of the stored descriptors. The complexity of indexing is in the order of l (number of descriptors of the request image).

This indexing technique leads to a very effective recognition. Our base of test images contains 1,020 images, which are represented by 154,030 descriptors. The average time of recognition is less than 5 seconds on a Sparc 10. These performances can be further improved by paralleling since each descriptor is treated separately.

7.4.7. *Experimental results*

Experiments were conducted with an image base containing 1,020 images. They exhibited the robustness of the method to image rotations, scaling, illumination changes, limited variations of viewpoint, partial visibility and the emergence of additional characteristics. The rate of recognition obtained is above 99% for a variety of test images taken under different conditions.

The base used includes various types of images such as 200 masterpieces, 100 aerial images and 720 images of 3D objects (see Figure 7.4). The 3D objects include the image base of Columbia. These images represent a large variety. However, certain images of masterpiece and certain aerial images are very similar. This induces ambiguities, which the method of recognition is able to manage.

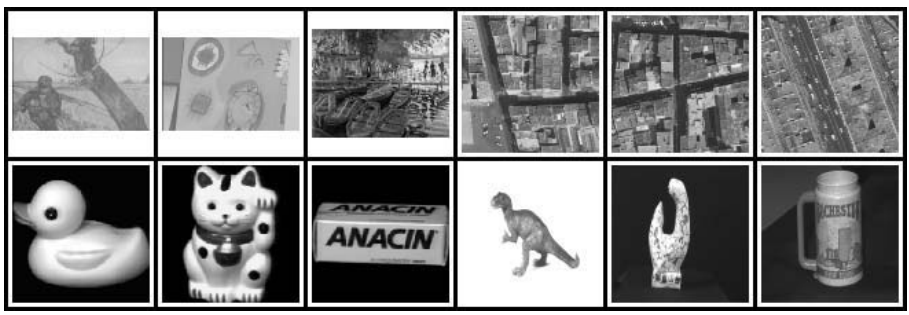


Figure 7.4. *Some images of the base. The complete base contains 1,020 images*

In the case of 2D objects, an object is represented by an image in the base. It is the same for the quasi-planar objects such as aerial images. A 3D object must be represented by several images taken from different viewpoints. The images are stored in the base with a change in the view angle of 20 degrees.

In what follows, three examples are presented for each type of image. For each one of them, the right side image is stored in the base. This image is recognized from any image on the left. Figure 7.5 presents the recognition of a masterpiece in the case of an image rotation and scaling. This figure also shows that recognition is possible if only a part of the image is found.

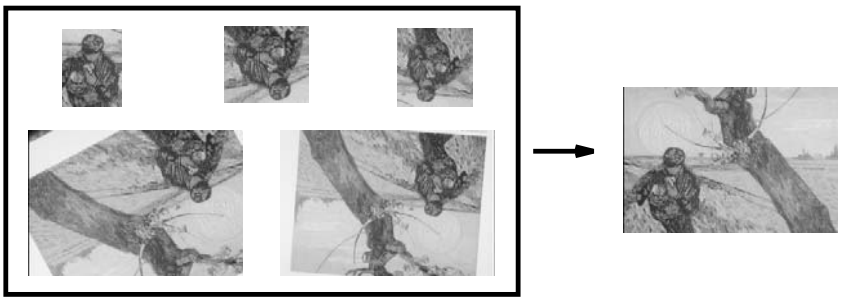


Figure 7.5. *The image on the right is correctly recognizable from the images on the left. These images underwent a rotation, a scaling even when only a part of the image is sought*

Figure 7.6 presents an example of aerial image. It shows a correct recognition in the case of image rotation and when only a part of the image is sought for. In the case of aerial images, it is necessary to be robust to a changing viewpoint and additional data (data present only on one of the images). Buildings appear different because of the change in view angle. In the same way, cars move, some disappear and others appear.

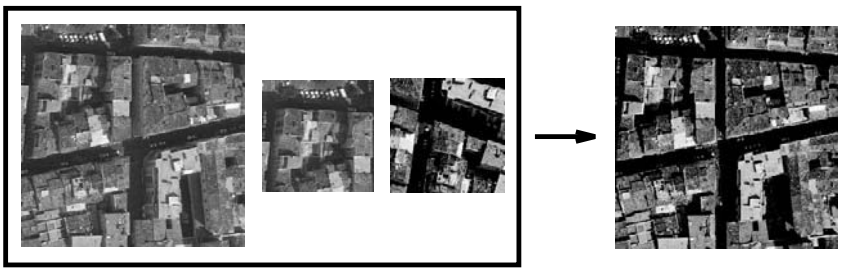


Figure 7.6. *The image on the right is correctly recognizable from the images on the left. These images are taken from different view angles (Istar property)*

Figure 7.7 presents the recognition of a 3D object. This object was correctly recognized in the presence of a image rotation, a scaling, a change in background of the image, and a partial visibility. Moreover, there is a difference of 10 degrees of view angle between observations. It should be noted that not only was the object recognized, but it is the most similar stored image.

Figure 7.7. *The image on the right is correctly recognizable from the images on the left. The 3D object can be in front of a complex or partially visible background*

A systematic evaluation of the method for a great number of test images is presented in [SCH 96b]. These test images are taken under different conditions: image rotation, scaling, change of luminosity, variation in viewpoint and partial visibility. This evaluation can be summarized as follows: perfect results (no error) are obtained in the case of rotations. Using a multi-scale approach, the recognition rate reaches 100% up to a scaling of factor 2. Such a factor seems to be a limit of this method. However, this limit is not due to the characterization with invariants but to the stability of the detector of key points. Indeed, the repeatability decreases rapidly when the scale factor is greater than 1.6. In the case of changes in the illumination intensity and the position of the source of light, the recognition rate also reaches 100%.

Variation of viewpoint. Test images were taken from viewpoints different from those of the images stored in the base. Each aerial image was taken from 4 different viewpoints. Only the first image is stored in the base. For these images taken from different viewpoints, the recognition rate is 99%. The only image that is not recognized correctly corresponds to an aerial image of the port and contains only water on which there is no reliable key point.

In the case of 3D objects, the method is robust to the change of viewpoint which is equal to 10 degrees at least. In the case of the base of Columbia,

i.e. a benchmark for object recognition, the recognition rate obtained is 100%. In this base, images are spaced in 20 degrees and the test images are spaced in 5 degrees.

Partial visibility. Pieces of different sizes were randomly extracted from masterpiece images. The relative size of the extracted images varied between 10% and 100%. For the pieces whose relative size is higher or equal to 30%, the recognition rate is 100%. For a relative size of 20%, the rate obtained is 95%. For a relative size of 10%, the rate obtained is 90%. These high recognition rates can be explained by the fact that the extracted points are very discriminating and thus only few points are sufficient to recognize the image correctly. Thus, it is possible to recognize an image even if only a part of this image is used for searching.

7.4.8. *Extensions*

In the following section, we briefly present extensions of the approach described earlier. In the case of 3D objects, a geometric verification based on the trifocal tensor makes it possible to improve results [SCH 96c]. The request image is mapped with the two most similar images of the base. The mappings between these three images are used to estimate the trifocal tensor in a robust manner. This enables the rejection of false pairings and thus helps to make a robust vote. This tensor makes it possible, in addition, to project additional information on the image used for research. For this purpose, we add additional information to images of the base which model a 3D object. Such an addition is illustrated in Figure 7.8a. This information is then projected to label the request image. Figure 7.8b presents an example where labeling is correctly projected on the request image.

The addition of a probabilistic model makes it possible to improve the results as compared to of a simple algorithm of vote [SCH 99]. This model allows us to take into account the variability of descriptors, their relevance, as well as the variability of spatial configurations. The method is structured for clearly separating the probability of correspondences based on descriptors of spatial coherence of a correspondence set. For each descriptor of the request image, several correspondences exist in the base. Each of these correspondences is weighted by its variability and discriminance. The search for coherent sets of correspondences reinforces each one of them. The recognized model is that which obtains the greatest probability from these sets. The experimental results presented for this method show a significant gain.

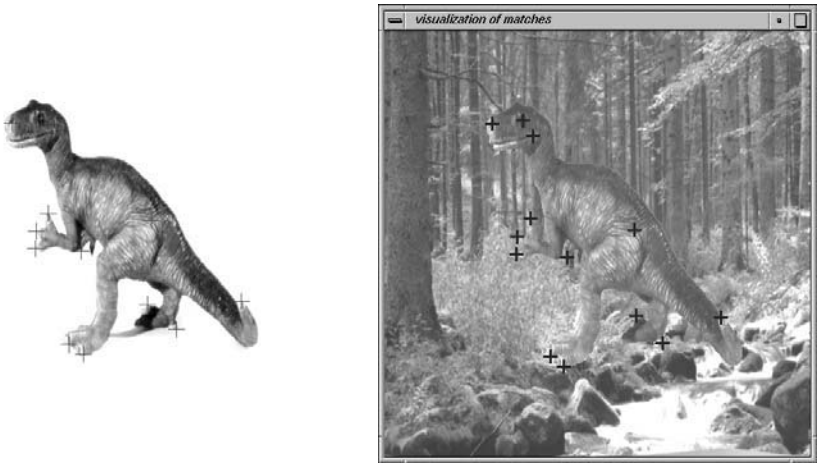


Figure 7.8. *(a) Image of the base with added additional information. (b) Example of addition of information on the request image. Additional information is correctly added*

The multi-scale approach that was previously presented is limited to a scale factor of 2. This is due to the quality of the extracted key points. For scale factors superior to 2, these points are no longer repeatable [SCH 00], i.e., they are no longer localized at the same place. The key to success for such a problem is the adequate representation of characterization in the scale space [DUF 00]. It is shown how to extract key point in different scales and how to compare two images of different resolution. The results show that the method can be used for changes of scale up to a scale factor of 6. Figure 7.9 shows an example of mapping in a significant scaling. The 16 pairings are correct.

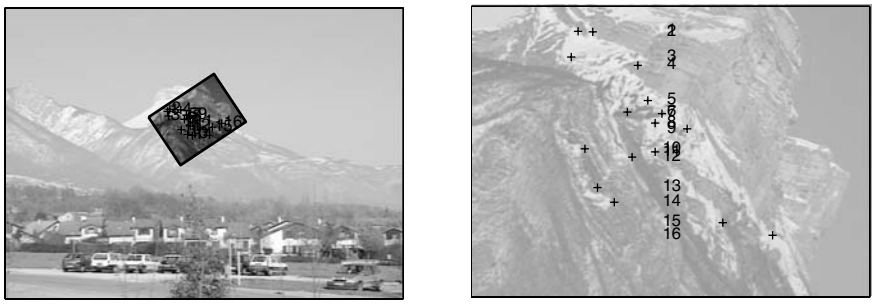


Figure 7.9. *Example of mapping in the case of a significant scaling. The 16 pairings are correct. The high resolution image (image on the right side) is projected on the low resolution image with homography calculated from 16 pairings. The rotation angle is estimated at 34 degrees and the scaling at 5*

In the case of color images, the transposition of descriptors which have just been presented is possible, except for some specific difficulties to be solved [GRO 00]. These difficulties arise owing to the fact that there is only one possible coding of color information, and that the influence of lighting conditions on the value of pixels is a complex process, which we cannot completely model. The selected solution consists of using the non-linear RGB components, which are among the most common. A change of luminosity is modeled by a diagonal matrix and a translation vector.

Once a transformation model of the pixels is chosen, the same diagram as the previous one is followed: the detection of points, the calculation of descriptors by formulae similar to those presented, which gives invariant vectors whose dimension lies between 18 and 29. The methods of comparison and vote are the same. The use of color makes it possible to improve the results to an extent. The problem is that the point detector does not take into account the photometric type of transformation, which is considered for invariance. To circumvent this problem, it is necessary to pretreat the image using the local standardization technique and to detect the points only in the second time. Then, there is a greater improvement in the results.

7.5. Geometric approach

In this section, we are specifically interested in structured images, i.e., those where we can make use of segments to approach the contours of the objects. They are thus generally images of scenes where manufactured objects predominate which frequently occur, for example, in robotics manufacturing applications. We present the basic algorithm and some extensions.

7.5.1. *Basic algorithm*

Initially, the algorithm seeks to determine whether two images represent the same object or have a common part. For this, the following stages are followed:

– We start by extracting the contours of images. Then, we approach these contours by segments, which we connect between them by their ends; according to the type of descriptors used, it can also be necessary to extract the key point of the image.

– We enumerate the configurations on which we want to calculate descriptors: it can be a group of 2 or 3 connected segments, or of configurations including 1 or 2 segments and 1 or 2 key points.

– We then calculate quasi-invariants on these configurations; these quantities are based on calculations of angles or distance ratios to obtain a simple quasi-invariance, which is equivalent to the invariance in the plane similarities of the image, or the ratios of aligned points if we want an invariance with affine transformations of the image.

– If only two images are considered, the descriptors are paired: we thus find that the two images have close values. Since these descriptors are not very discriminating, it is necessary to use a total constraint to remove ambiguities: each time two descriptors are paired, we calculate the transformation between the two corresponding configurations. Therefore, we retain the pairings that have coherent transformations from a geometric viewpoint [GRO 98].

– If we confront an image in a set of images, we carry out paralleling: all descriptors of the images of the base are gathered in an indexing structure (tree structure in main memory or on disk). Each descriptor of the image to be recognized is confronted with those in the base and, for each possible pairing, we give a vote to the image of the corresponding base. To remove ambiguities, it is possible to use transformations between configurations in the same way and consider only the geometrically coherent votes [LAM 96].

Once the first pairing is made between two images, it is possible to easily refine the first pairing carried out: we can calculate the epipolar geometry by a robust method, or an approximation of the apparent movement by homography and then supplement pairing by making use of this new information.

It is in addition possible to pair n images at the same time. In this case, we pair the images two by two and then we consider the graph of pairings thus realized that we prune to keep only the coherent pairings. Such a technique enables us, by the use of intermediate images, to propagate pairings between two images, which cannot be paired directly.

7.5.2. *Some results*

7.5.2.1. *Pairing results*

Let us schematically illustrate the method presented using some examples. Figure 7.10 shows, on top, the first pairing carried out between two images and, at the bottom, the one obtained after correction and improvement. The paired points are those that carry the same number. We can notice that even the non-polyhedric objects were paired. On the other hand, it is clear that if

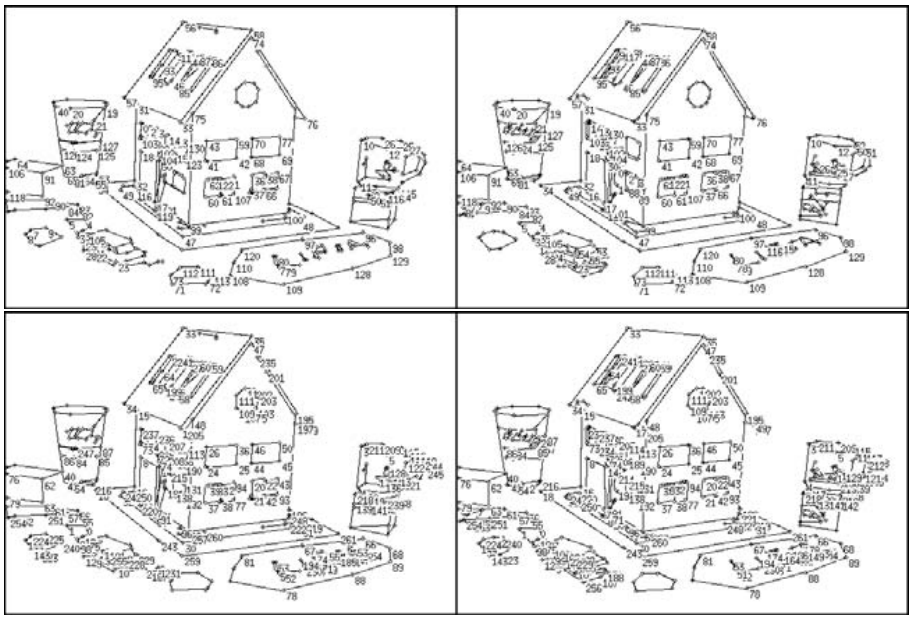


Figure 7.10. *Initial pairings on top, those improved at the bottom*

some "obvious" points were not paired, there are few errors of pairing. In this case, we use angles and distance ratios calculated on the connected segment pairs.

Figure 7.11 shows two images on the left for which the use of mixed configurations points-segments is essential in obtaining correct results. In the images on the right-hand side the paired elements are in red. Even without classification, it is clear that there are many errors. This is a particularly difficult case having made many techniques fail.

7.5.2.2. *Results of indexing and recognition*

Figure 7.12 illustrates the recognition process in the case where this process is used to capture data in robotics; the image to recognize is pretreated to extract primitives and the desired configurations from it, and the vote takes place and three images of the base having the most (geometrically coherent) votes are shown. Once the pairing is done, it is possible to calculate the exposure of the object and then to seize it using tongs through visual control.

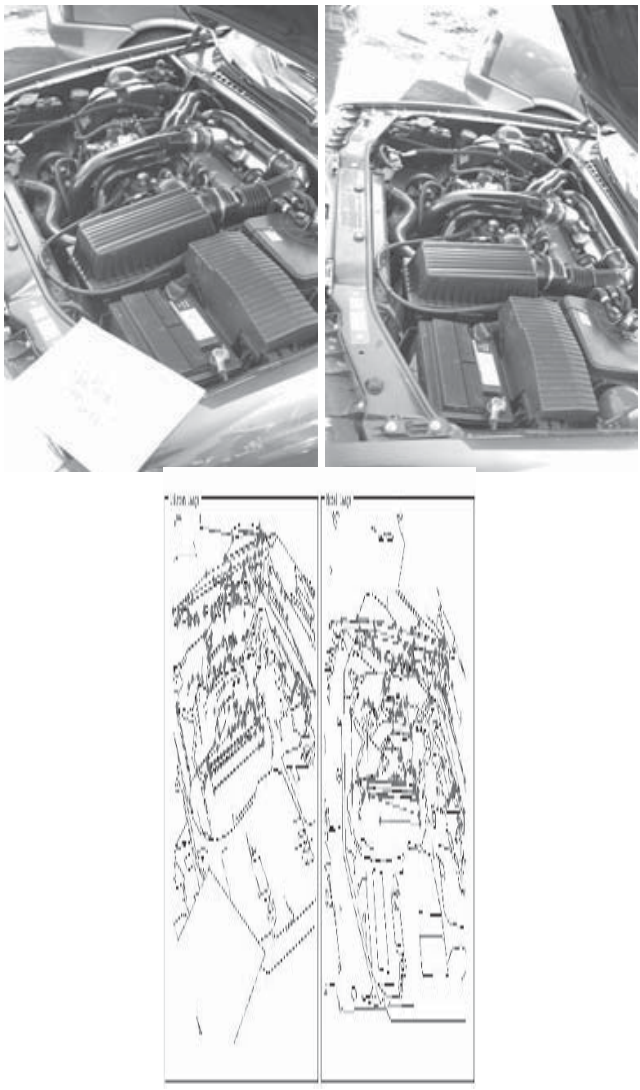


Figure 7.11. *Two images of engines (top) and their pairing (bottom). The paired elements are marked by diamonds*

Figure 7.13 shows typical recognition results. Images on the left are the requests (which are not present in the base), images on the right are the answers of the systems. The base contains images of engines and some other images.

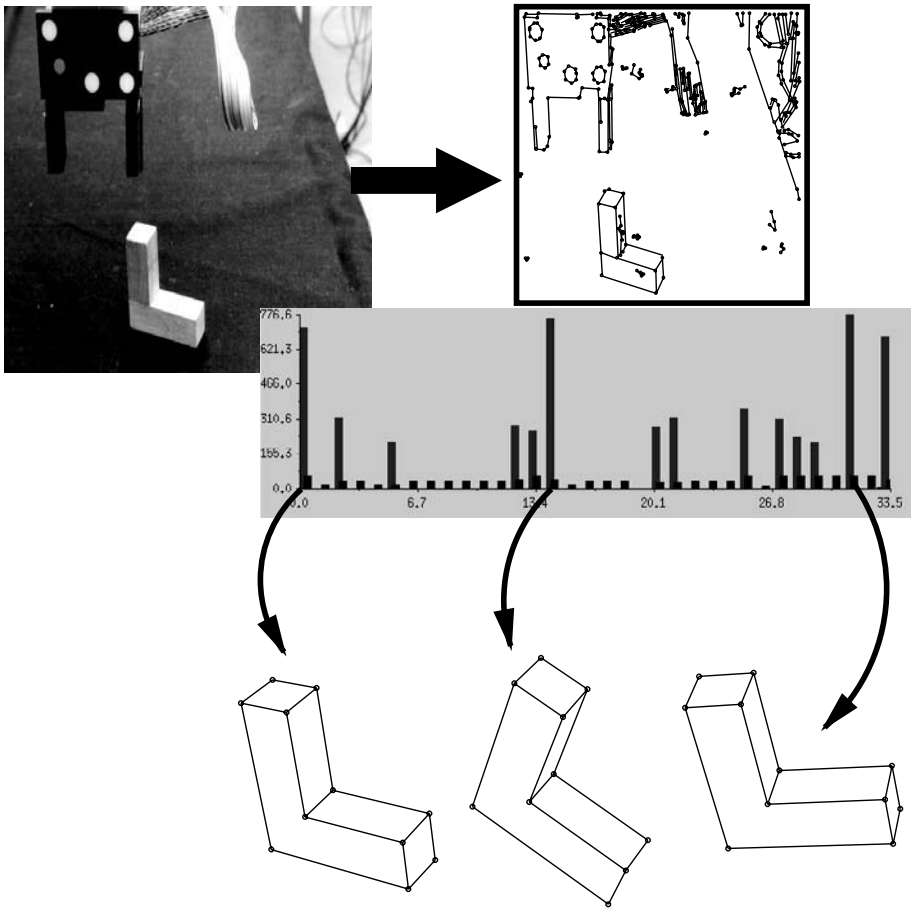


Figure 7.12. *Recognition of an object to capture data*

7.6. Indexing of images

Since we work with several images, the difficulty of managing this package of images arises, and using a DBMS and storing in a disk appears a natural solution. Database techniques are then used to increase the search speed and to minimize costs related to input-outputs. The databases indeed know how to specifically organize the data on secondary memory, offer structuring and indexing techniques accelerating research, and also offer guarantees of reliability and coherence of competitive accesses, i.e. desirable guarantees within a system of real interrogation.

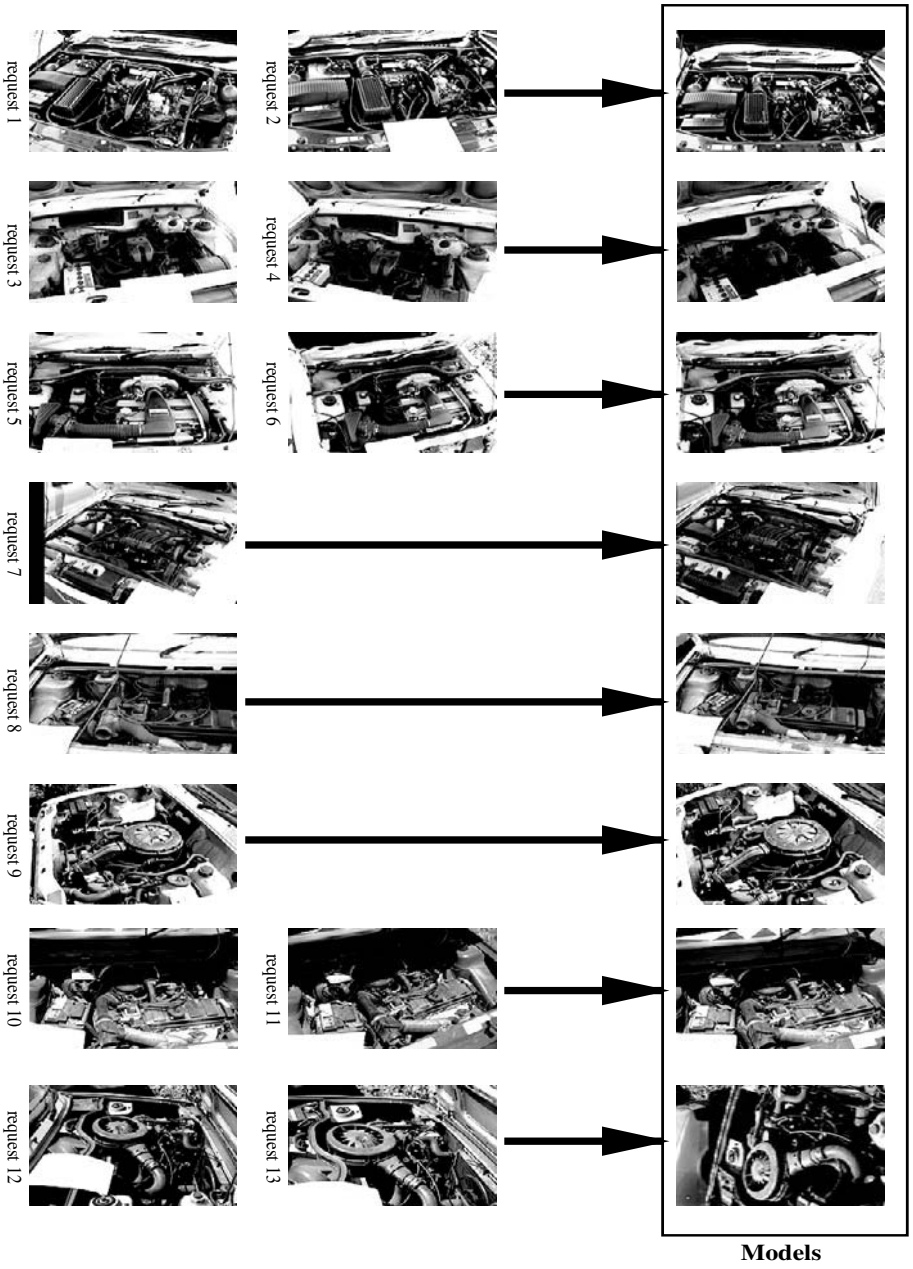


Figure 7.13. *Recognition of engines*

Unfortunately, databases have not been designed for the type of data that we use and the storage of coded descriptors in the form of real unprecise vectors presents some difficulties:

– The descriptors are real vectors whose dimension can be large. The difficulty is then twofold. First, the comparative distance of the descriptors uses all dimensions at the same time. Thus, we find ourselves looking within large real spaces, without being able to separate dimensions and treat them one after the other. In addition, the variability of descriptors demands that we must *explore* space and not merely search for one particular point. Very often, this type of exploration has an exponential complexity with the dimension of space, which is a problem known as *dimensionality curse* [WER 96, LAM 97, BER 98].

– When local descriptors are used, generally each image has between 50 to 600 descriptors. This imposes a strong constraint on the performances of the system, because the response time felt by the user is that of 50 to 600 successive interrogations and not that due to only one. Moreover, the response time must integrate the decision-making process.

– Different indexing systems of other works are tested using randomly generated data sets, normally according to a uniform law, which is not the case with real data. Thus, there is a risk of obtaining inferior performances as compared to those reported by the designers.

This section therefore begins by describing specific techniques developed for these data, by granting more details about two of them that appear more promising to us and then presents some evaluation results of these techniques, before concluding with some perspectives.

7.6.1. *Traditional approaches*

Multimedia indexing algorithms are classified into two main categories: those which gather data according to their proximity *(data-partitioning index methods)* and those which partition the multi-dimensional space *a priori* and then store the data according to this partitioning *(space-partitioning index methods)*. All these approaches populate the space with image descriptors or approximations of these descriptors.

All algorithms of the first category are derived from R-Tree [GUT 84], which was originally known to index 2D data. The central concept is that of a bounding box: the leaves of the tree refer to each object through its

bounding box and the internal levels of the tree store the lower level bounding boxes. Several criteria can be employed for deciding to include or preserve two distinct lower level boxes [BER 96]. The principle of R-Tree was then generalized to be applied to multi-dimensional data and there are alternatives: SS-Tree [WHI 96] uses spheres instead of bounding boxes and SR-Tree [KAT 97] uses the intersections of spheres and rectangles.

Lin, Jagadish and Faloutsos [LIN 94a] presented a technique with TV-Tree where we distinguished dimensions which the system must systematically use, those that it is unaware of and those that it can use to refine a search. The major defect of this technique is that it requires *a priori* knowledge of the precise distribution of data according to each dimension.

All these approaches use a filling factor equal to 50% at the time of bursting of the nodes of the tree, which guarantees obtaining balanced trees and also makes it possible to maximize the usage rate of pages on disk. However, [BER 98] shows that this factor generally induces a probability of access to each page of the index close to 100% during each research. Thus, there is a high chance that the totality of the index will be covered, and this by means of many random accesses to the disk.

Techniques in the second category divide multi-dimensional space according to a complex and regular grid, as the grid-files do [NIE 84], K-D-B-Tree [ROB 81], LSD^h-Tree [HEN 98]. Then, the data are stored in suitable boxes. These techniques soon become ineffective because the number of boxes grows exponentially with the dimension of the descriptors and finally exceeds that of the descriptors. Moreover, when the required point is close to the borders of space division, research then has to explore the neighboring boxes in large numbers (generally to realize that there is nothing in each). The cost of a search is then prohibitive. The most modern techniques like [AGR 98] or [HIN 99] are evaluated as effective for a low number of dimensions and for less significant noise (ε).

7.6.2. *VA-File and the Pyramid-Tree*

[WEB 98] shows that above 10 dimensions, a sequential search becomes more efficient than any navigation in an index, which invalidates the approaches presented above. Recently, two methods were proposed to overcome this obstacle: the VA-File and the Pyramid-Tree.

With the VA-File [WEB 98], Weber, Schek and Blott developed a method, which aims to improve the performances of the sequential research. This

technique manages two sets of data, the first containing descriptors and the second containing an approximation of the latter. The latter file must save them in the main memory to guarantee good performances.

During the creation of the index, to calculate approximations, the VA-File divides each dimension d into 2^{b_i} boxes, so that the boxes approximately contain the same number of points. Thus, there are 2^b boxes, where $b = \sum_i b_i$, which are numbered from 0 to $2^b - 1$. The descriptors of the base are then read one after the other and the approximation of each descriptor is then equal to the number of boxes between the limits from which it emerges. The approximation file is thus made up of pairs (identifying descriptor, box number). Only information related to boxes containing at least one descriptor is stored, thus avoiding the management of a large number of empty boxes.

At the time of research, the request descriptor undergoes the same process of approximation. The algorithm compares the approximations between them and then sequentially reads the contents of the boxes that will most likely provide the closest neighbors, on the disk. The first stage thus acts as a filter, which eliminates the boxes from the search (and thus the descriptors) which do not have a chance of belonging to the answer, thus limiting the number of comparisons to be made by sequential research.

Berchtold, Böhm and Kriegel, with Pyramid-Tree [BER 98], proposed a method, which divides space $[0, 1]^d$ into $2 \times d$ pyramids. Each pyramid has its point placed at the center of the space (0.5, 0.5, etc.), has a base with a surface of $d-1$ dimensions and is numbered. Each pyramid is then divided into sections, parallel to its base. The sections close to the top are thus smaller than those that are closer to the base. This space division has the property to create a number of boxes growing linearly, and not exponentially, with dimension.

This double division allows any point of the multi-dimensional space to be expressed as a pair (pyramid number, height in the pyramid). This contraction of the multi-dimensional space allows the usage of a particularly effective index for this type of data and interval type requests: B^+-Tree. A page of B^+-Tree matches a specific section of a particular pyramid.

7.6.3. *Some results*

Results in terms of performance in recognition were provided in the preceding sections. Here, we are interested in the speedy performances according to three scenarios.

7.6.3.1. *Context of experiments*

The experiments given below were carried out using two bases: the first includes 413,412 invariant descriptors of dimension 24, which are descriptors resulting from 1,826 images having 50 seconds of a film and varied fixed images. The second base consists of 413,412 vectors of dimension 24 drawn in a uniform and pseudo-random way. We tested various search scenarios using 3 algorithms: VA-Files, Pyramid Trees and sequential search. A typical research consists of taking a request made up of 150 descriptors, which are not in the base and to search for the 10 closest descriptors of the base. This gives 1,500 descriptors.

7.6.3.2. *First experiment*

The first scenario aims to study the influence of the dimension of the descriptors over search time. For this, we constituted the bases of various dimensions, by truncating real data, or by generating vectors of adequate size.

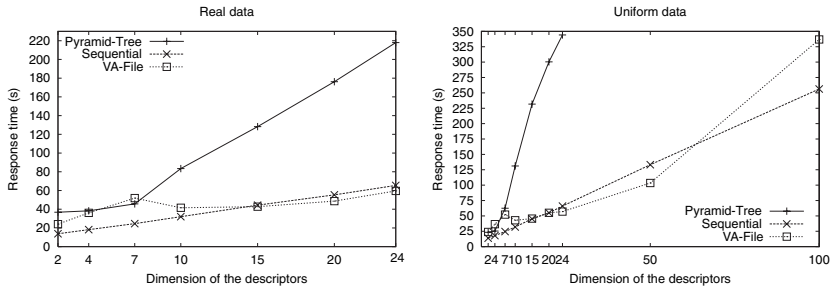


Figure 7.14. *Base of 413,412 descriptors, 150 request descriptors, real and uniform distributions, increasing dimension of the descriptors*

Mainly, we can notice that VA-Files have performances very close to those of sequential research, except for few dimensions, but that is not the difficult case for which this structure was developed.

On the contrary, Pyramid Trees have rapidly decreasing performances and we did not carry out an experiment beyond 24 dimensions because the computing time becomes crippling.

7.6.3.3. *Second experiment*

In the second experiment, we observed the influence of base size. There too, VA-Files and sequential research are at par. We can notice that in spite of

the reduced size of the base (1,816 images, which is little as compared to real needs), we obtain a long computing time for online use: the users of search sites on the Internet are accustomed to a search time of a few seconds, not of a minute or more.

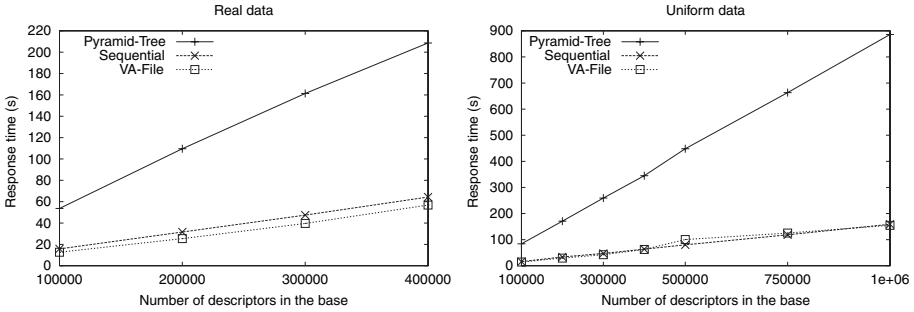


Figure 7.15. *Influence of base size, with a request of 150 descriptors and descriptors of dimension 24*

7.6.3.4. *Third experiment*

In the third experiment, the influence of the number of descriptors in the request is studied. We observe a linear behavior which, once again, shows that the limits of the current systems are not far.

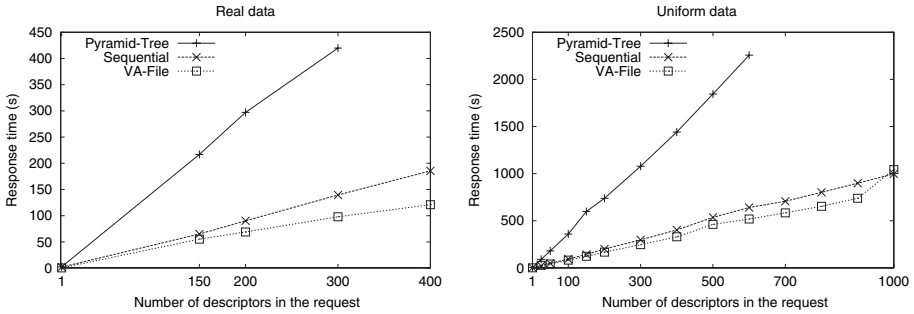


Figure 7.16. *Influence of the request size, the base contains 413,412 descriptors of dimension 24*

7.6.4. *Some prospects*

The present study remains compartmental, but it very well shows the way that remains to be crossed to obtain a system that can manage many images (from 100,000 to a million) while having response times like those obtained

for textual research by the presently available engines on the Web. If this objective is not achieved yet, it is important to move away from search paths to arrive there.

Here are some proposals inspired from the use of VA-Files:

– First, we can try to visit only one box per request descriptor. To minimize the problems involved in borders, we can duplicate the base with a double system of boxes in staggered rows distributed on two machines. For a descriptor, we then use the system, which keeps this descriptor farther away from the borders between boxes.

– Then, we can try to use redundancy connected to the use of local descriptors. Indeed, a request is not limited to searching for a correspondence between a descriptor resulting from the request image and that coming from the base, but proceeds mainly by accumulation of evidence. Thus, we can try to not conduct interrogations until their end, but to stop the process as soon as there is a strong probability of obtaining a good result. Such an algorithm requires to be validated from the speed point of view as well as from the recognition point of view.

– The distribution of descriptors is not uniform. Moreover, if some are found in many images, others are very specific. The quantity of information brought by each descriptor of the request image is thus not the same. We can use a Bayesian formalism to estimate it. This can make it possible to choose the order in which request descriptors are used and to prematurely stop the research as indicated in the preceding point.

– A better management of hidden memory can limit the number of pages loaded in the memory. For this, we can gather the descriptors of the request image that are close and relate research on these descriptors.

– Since the dimension of descriptors is one of the sources of complexity of the search, we can divide these descriptors to carry out, for example, 4 searches on descriptors of dimension 6 rather than only one on a descriptor of dimension 24. These four searches can be made on different machines to save time. On the other hand, these results should be synthesized, but the total complexity should be decreased.

The tracks are thus numerous, but they remain to be explored and evaluated.

7.7. Conclusion

In this chapter, we presented two methods whose purpose is to enable recognition of images that represent the same object, or more precisely the

same aspect of an object, but under different conditions. These differences can relate to viewpoint, illumination conditions, composition of image or the position of the various elements of the image. The two methods that are presented are based on a common principle: using quasi-invariants associated with local descriptors. The principle and advantages of this approach were presented. The two methods were then tested with respect to image databases of great dimension but not requiring the storage of descriptors on disk. In the last section, we dealt with specific problems related to the use of external disks as a medium of storing data.

7.8. Bibliography

[AGR 98] AGRAWAL R., GEHRKE J., GUNOPULOS D. and RAGHAVAN P., "Automatic subspace clustering of high dimensional data for data mining applications", *Proceedings of the* ACM SIGMOD *International Conferenceon Management of Data*, Seattle, Washington, USA, p. 94–105, 1998.

[AYA 86] AYACHE N. and FAUGERAS O., "HYPER: a New Approach for the Recognition and Positioning of 2D Objects", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, p. 44–54, 1986.

[BAL 81] BALLARD D., "Generalizing the Hough Transform to Detect Arbitrary Shapes", *Pattern Recognition*, vol. 13, no. 2, p. 111–122, 1981.

[BEN 90a] BEN-ARIE J., "Probabilistic Models of Observed Features and Aspect Graphs", *Pattern Recognition Letters*, June 1990.

[BEN 90b] BEN-ARIE J., "The Probabilistic Peaking Effect of Viewed Angles and Distances with Application to 3D Object Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 8, p. 760–774, 1990.

[BER 96] BERCHTOLD S., KEIM D. and KRIEGEL H., "The X-tree: An Index Structure for High-Dimensional Data", *Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India, the Very Large Database Endowment*, Fri Oct 10 1997, p. 28–39, 1996.

[BER 98] BERCHTOLD S., H-P. C.B. and KRIEGEL, "The Pyramid-Technique: Towards Breaking the Curse of Dimensionality", *ACM SIGMOD*, p. 142–153, 1998.

[BES 85] BESL P. and JAIN R., "Three-Dimensional Object Recognition", *ACM Computing Surveys*, vol. 17, no. 1, 1985.

[BIN 93] BINFORD T. and LEVITT T., "Quasi-Invariants: Theory and Exploitation", *Proceedings of DARPA Image Understanding Workshop*, p. 819–829, 1993.

[BOL 86] BOLLES R. and HORAUD R., "3DPO: A Three-Dimensional Part Orientation system", *The International Journal of Robotics Research*, vol. 5, no. 3, p. 3–26, 1986.

[BRO 83] BROOKS R., "Model-Based Three-Dimensional Interpretations of Two-Dimensional Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, p. 140–150, 1983.

[BUR 90] BURNS J., WEISS R. and RISEMAN E., "View Variation of Point Set and Line Segment Features", *Proceedings of DARPA Image Understanding Workshop*, Pittsburgh, Pennsylvania, USA, p. 650–659, 1990.

[CHI 86] CHIN R. and DYER C., "Model-based Recognition in Robot Vision", *ACM Computing Surveys*, vol. 18, no. 1, p. 67–108, 1986.

[DUF 00] DUFOURNAUD Y., SCHMID C. and HORAUD R., "Matching Images with Different Resolutions", *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, p. 612–618, June 2000.

[FAU 86] FAUGERAS O. and HEBERT M., "The Representation, Recognition, and Locating of 3D Objects", *The International Journal of Robotics Research*, vol. 5, p. 27–52, 1986.

[FUN 95] FUNT B. and FINLAYSON G., "Color Constant Color Indexing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, p. 522–529, 1995.

[GDA 96] GDALYAHU Y. and WEINSHALL D., "Measures for Silhouettes Resemblance and Representative Silhouettes of Curved Objects", *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, p. 363–375, 1996.

[GRI 87] GRIMSON W. and LOZANO-PEREZ T., "Localizing Overlapping Parts by Searching the Interpretation Tree", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, p. 469–482, 1987.

[GRI 89] GRIMSON W., "On the Recognition of Curved Objects", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, p. 632–643, 1989.

[GRI 90] GRIMSON W. and HUTTENLOCHER D., "On the Sensitivity of the Hough Transform for Object Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, p. 225–274, 1990.

[GRO 93] GROS P., Outils géométriques pour la modélisation et la reconnaissance d'objets polyédriques, PhD Thesis, Institut National Polytechnique de Grenoble, July 1993.

[GRO 95] GROS P., "Matching and clustering: two steps towards object modelling in computer vision", *The International Journal of Robotics Research*, vol. 14, no. 6, p. 633–642, 1995.

[GRO 98] GROS P., BOURNEZ O. and BOYER E., "Using Local planar geometric invariants to match and model images of line segments", *Computer Vision and Image Understanding*, vol. 69, no. 2, p. 135–155, 1998.

[GRO 00] GROS P., "Experimental evaluation of color illumination models for image matching and indexing", *Proceedings of the RIAO'2000 Conference Content-Based Multimedia Information Access*, p. 567–574, 2000.

[GUT 84] GUTTMAN A., "R-Trees: a dynamic index structure for spatial searching", *ACM SIGMOD International Conference on Management of Data*, Boston, Massachussets, USA, p. 47–57, June 1984.

[HAA 94] TER HAAR ROMENY B., FLORACK L., SALDEN A. and VIERGEVER M., "Higher Order Differential Structure of Images", *Image and Vision Computing*, vol. 12, no. 6, p. 317–325, 1994.

[HAR 88] HARRIS C. and STEPHENS M., "A combined corner and edge detector", *Alvey Vision Conference*, p. 147–151, 1988.

[HEN 98] HENRICH A., "The LSD^h-tree: an access structure for feature vectors", *Proceedings of the Fourteenth ICDE International Conference on Data Engineering*, Orlando, Florida, USA, p. 362–369, 1998.

[HIN 99] HINNEBURG A. and KEIM D., "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering", *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, p. 506–517, 1999.

[HUT 90] HUTTENLOCHER D. and ULLMAN S., "Recognizing Solid Objects by Algnment with an Image", *International Journal of Computer Vision*, vol. 5, no. 2, p. 195–212, 1990.

[KAT 97] KATAYAMA N. and SATOH S., "SR-tree: an index structure for high-dimensional nearest neighbor queries", *Proceeding of ACM SIGMOD Conference*, Tucson, Arizona, p. 369–380, 1997.

[KoE 87] KOENDERINK J. and VAN DOORN A., "Representation of local geometry in the visual system", *Biological Cybernetics*, vol. 55, p. 367–375, 1987.

[KRI 90] KRIEGMAN D. and PONCE J., "On recognizing and positioning curved 3D objets from image contours", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, p. 1127–1137, 1990.

[LAD 93] LADES M., VORBRÜGGEN J., BUHMANN J., LANGE J., MALSBURG C., WÜRTZ R. and KONEN W., "Distortion Invariant Object Recognition in the Dynamic Link Architecture", *IEEE Transactions on Computers*, vol. 42, no. 3, p. 300–311, 1993.

[LAM 88] LAMDAN Y., SCHWARTZ J. and WOLFSON H., "Object recognition by affine invariant matching", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Diego, California, USA, p. 335–344, 1988.

[LAM 96] LAMIROY B. and GROS P., "Rapid object indexing and recognition using enhanced geometric hashing", *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, vol. 1, p. 59–70, April 1996.

[LAM 97] LAMIROY B. and GROS P., "Object indexing is a complex matter", *Proceedings of the 10th Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, vol. I, p. 277–283, June 1997.

[LIN 94a] LIN K.-I., JAGADISH H. and FALOUTSOS C., "The TV-tree: an index structure for high-dimensional data", *VLDB Journal*, p. 517–542, 1994.

[LIN 94b] LINDEBERG T., *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, 1994.

[LOW 87] LOWE D., "Three-dimensional object recognition from single two-dimensional images", *Artificial Intelligence*, vol. 31, no. 3, p. 355–395, 1987.

[MUN 90] MUNDY J. and HELLER A., "The Evolution and Testing of a Model-Based Object Recognition System", *Proceedings of the 3rd International Conference on Computer Vision*, Osaka, Japan, Mon Aug 5 1996, p. 268–282, 1990.

[MUR 95] MURASE H. and NAYAR S., "Visual learning and recognition of 3D objects from appearance", *International Journal of Computer Vision*, vol. 14, p. 5–24, 1995.

[NAG 95] NAGAO K., "Recognizing 3D objects using photometric invariant", *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, Massachusetts, USA, p. 480–487, 1995.

[NAY 93] NAYAR S. and BOLLE R., "Computing reflectance ratios from an image", *Pattern Recognition*, vol. 26, no. 10, p. 1529–1542, 1993.

[NIE 84] NIEVERGELT J. and HINTERBERGER H., "The grid file: an adaptable, symmetric multikey file structure", *ACM Transactions on Database Systems*, vol. 9, no. 1, p. 38–71, 1984.

[RAO 95] RAO R. and BALLARD D., "Object indexing using an iconic sparse distributed memory", *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, Massachusetts, USA, p. 24–31, 1995.

[ROB 81] ROBINSON J., "The K-D-B-tree: a search structure for large multi-dimensional dynamic indexes", *SIGMOD '81*, Ann Arbor, MI, Association for Computing Machinery, p. 10–18, 1981.

[ROT 93] ROTHWELL C., "Hierarchical object descriptions using invariants", *Proceeding of the DARPA-ESPRIT Workshop on Applications of Invariants in Computer Vision*, Azores, Portugal, p. 287–303, October 1993.

[SAM 84] SAMET A., "The quadtree and related hierarchical data structures", *ACM Computing Surveys*, vol. 16, no. 2, p. 189–259, 1984.

[SCH 96a] SCHIELE B. and CROWLEY J., "Object recognition using multi-dimensional receptive field histograms", *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, p. 610–619, 1996.

[SCH 96b] SCHMID C., Appariement d'images par invariants locaux de niveaux de gris, PhD Thesis, Institut National Polytechnique de Grenoble, GRAVIR – IMAG – INRIA Rhône–Alpes, July 1996.

[SCH 96c] SCHMID C., BOBET P., LAMIROY B. and MOHR R., "An image oriented CAD approach", PONCE J., ZISSERMAN A. and HÉBERT M. (Ed.), *Object Representation in Computer Vision II*, no. 1144 Lecture Notes in Computer Science, Springer-Verlag, p. 221–245, April 1996, *ECCV'96 International Workshop*, Cambridge, UK, Proceedings.

[SCH 96d] SCHMID C. and MOHR R., "Combining Grayvalue Invariants with Local Constraints for Object Recognition", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Francisco, California, USA, June 1996, ftp://ftp.imag.fr/pub/labo-GRAVIR/MOVI/publications/Schmid_cvpr96.ps.gz.

[SCH 97] SCHMID C. and MOHR R., "Local Grayvalue Invariants for Image Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, p. 530–534, 1997.

[SCH 99] SCHMID C., "A Structured Probabilistic Model for Recognition", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, USA, vol. II, p. 485–490, 1999.

[SCH 00] SCHMID C., MOHR R. and BAUCKHAGE C., "Evaluation of interest point detectors", *International Journal of Computer Vision*, vol. 37, no. 2, p. 151–172, 2000.

[SLA 96] SLATER D. and HEALEY G., "The illumination-invariant recognition of 3D objects using color invariants", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, p. 206–210, 1996.

[SWA 91] SWAIN M. and BALLARD D., "Color indexing", *International Journal of Computer Vision*, vol. 7, no. 1, p. 11–32, 1991.

[TUR 91] TURK M. and PENTLAND A., "Face recognition using eigenfaces", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, USA, p. 586–591, 1991.

[VAN 91] VAN GOOL L., KEMPENAERS P. and OOSTERLINCK A., "Recognition and semi-differential invariants", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, USA, p. 454–460, June 1991.

[WEB 98] WEBER R. and ZEZULA P., "A quantitative analysis of performance study for similarity-search methods in high-dimensional spaces", *Proceedings of the 24th VLDB Conference*, 1998.

[WER 96] WERMAN M. and WEINSHALL D., "Complexity of Indexing: Efficient and Learnable Large Database Indexing", *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, vol. 1, p. 660–670, 1996.

[WHI 96] WHITE D. and JAIN R., "Similarity Indexing with the SS-tree", *12th International Conference on Data Engineering*, New Orleans, LA, IEEE, p. 516–523, February 1996.

[WIT 83] WITKIN A., "Scale-Space Filtering", *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, Germany, p. 1019–1023, 1983.

[WU 95] WU X. and BHANU B., "Gabor wavelets for 3D object recognition", *Proceedings of the 5th International Conference on Computer Vision*, Cambridge, Massachusetts, USA, p. 537–542, 1995.

[ZHA 95] ZHANG Z., DERICHE R., FAUGERAS O. and LUONG Q., "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry", *Artificial Intelligence*, vol. 78, p. 87–119, 1995.

[ZIS 95] ZISSERMAN A., FORSYTH D., MUNDY J., ROTHWELL C., LIU J. and PILLOW N., "3D object recogntion using invariance", *Artificial Intelligence*, vol. 78, no. 1-2, p. 239–288, 1995.

List of Authors

François CHAUMETTE
IRISA-INRIA
Rennes
France

Rachis DERICHE
INRIA
Sophia Antipolis
France

Michel DHOME
LASMEA-CNRS
Clermont-Ferrand
France

François GASPARD
DRT
CEA
Gif-sur-Yvette
France

Patrick GROS
IRISA-INRIA
Rennes
France

Jean-Thierry LAPRESTÉ
LASMEA
Blaise Pascal University
Clermont-Ferrand
France

Jean-Marc LAVEST
LASMEA
Blaise Pascal University
Clermont-Ferrand
France

Diane LINGRAND
I3S Laboratory
University of Nice
Sophia Antipolis
France

Éric MARCHAND
IRISA-INRIA
Rennes
France

Luce MORIN
IRISA
University of Rennes I
France

Lionel OISEL
Corporate Research
Thomson
Rennes
France

Long QUAN
Computer Science Department
HKUST
Hong Kong

Gérard RIVES
LASMEA
Blaise Pascal University
Clermont-Ferrand
France

Cordelia SCHMID
GRAVIR
INRIA
Grenoble
France

Thierry VIÉVILLE
INRIA
Sophia Antipolis
France

This page intentionally left blank

Index