ELSEVIER

Signal Processing: *Image Communication* 14 (1998) 71–94

SIGNAL PROCESSING: IMAGE COMMUNICATION

# 3D area matching with arbitrary multiview geometry[1]

Federico Pedersini*, Pasquale Pigazzini, Augusto Sarti, Stefano Tubaro

*Dipartimento di Elettronica e Informazione – Politecnico di Milano, Piazza L. Da Vinci, 32-20133 Milano, Italy*

**Abstract**

In this article we present a general and robust approach to the problem of close-range 3D reconstruction of objects from stereo correspondence of luminance patches. The method is largely independent on the camera geometry, and can employ an arbitrary number of CCD cameras. The robustness of the approach is due to the physicality of the matching process, which is performed in the 3D space. In fact, both 3D location and local orientation of the surface patches are estimated, so that the geometric distortion can be accounted for. The method takes into account the viewer-dependent radiometric distortion as well. The method has been implemented with a calibrated set of three standard TV-resolution CCD cameras. Experiments on a variety of real scenes have been conducted with satisfactory results. Quantitative and qualitative results are reported. 

*Keywords:* Stereopsis; 3D reconstruction; Area matching; Luminance transfer

## 1. Introduction

In the past decades we witnessed a proliferation of methods and algorithms for extracting information on the 3D structure of a scene from a sequence or an $n$-tuple of its views. The automatic estimation and storage of 3D information on objects and structures are, in fact, becoming more and more crucial in a broad range of applications. In the field of cultural heritage, for example, the automatic 3D reconstruction of works of art is becoming of paramount importance for purposes of restoration simulation and planning, analysis of the environmental impact through erosion's monitoring, creation of 3D catalogs, etc. The automatic creation of CAD models is gaining more and more importance for architectural applications or industrial problems of reverse engineering. Other industrial applications are obstacle avoidance for autonomous vehicle guidance in cluttered environments, monitoring and quality control for improving the efficiency of production plants. A wide variety of applications can also be found in the areas of telecommunications and entertainment, like 3D television, animation, etc.

* Corresponding author.

[1] Work supported in part by the ACTS Project 'PANORAMA', Proj. No. AC092, and in part by the CNR Strategic Project "*Conoscenza per immagini*; *un'applicazione ai beni culturali*" ("Knowledge acquisition from images: an application to cultural heritage").

0923-5965/98/$ – see front matter 
PII: S0923-5965(98)00029-0

The numerous approaches to 3D scene reconstruction can differ a great deal from each other, depending on the application that they are designed for and the type of scene to be reconstructed. For example, in real-time robotic applications great importance is attributed to computational efficiency, robustness and speed, while the environment is usually characterized by a certain geometrical regularity. In applications to cultural heritage, conversely, speed becomes less important but the scene can be much more complex. Furthermore, the reconstruction procedure is usually required to be non-invasive, i.e. to interact as least as possible with the object which is being analyzed.

Among the various available non-invasive approaches to the problem of automatic close-range measurement and reconstruction of object surfaces, some of the most popular ones are based on stereometric principles and make use of two or more cameras. All such methods share a common framework, according to which homologous primitives, i.e. stereo-corresponding object features, are detected, matched and back-projected onto the object space.

The image primitives that are used the most for 3D reconstruction are points (*fiducial marks*), edges and luminance patches. These types of features tend to provide information of a different nature. For example, image edges are particularly suitable for 3D reconstruction because of the intrinsic precision and reliability of their localization on the image [1]. Edge matching, however, can only generate sparse sets of 3D edges, as they are concentrated where the object surface is jagged or highly textured. Conversely, the matching/backprojection of the luminance profile of small image regions tends to provide us with much denser sets of 3D points but it is rather sensitive to the unavoidable viewer-dependent perspective/radiometric distortions, therefore this approach tends to be less stable and reliable.

In this paper we present a general and robust solution to the problem of 3D reconstruction from stereo correspondence of luminance patches. The method is largely independent on the camera geometry, and can employ an arbitrary number of standard TV-resolution CCD cameras. With three or more cameras, however, we have enough redundancy for removing possible matching ambiguities. The robustness of the approach can be mainly attributed to the physicality of the matching process, which is actually performed in the object space rather than on the image plane. In order to do so, besides the 3D location of the surface patches, it estimates their local orientation in 3D space as well, so that the geometric distortion of the luminance patch can be included in the model. Finally, the method adopts a non-Lambertian radiometric model in order to take the viewer-dependent radiometric distortion into account. In conclusion, the technique we propose in this article performs a first-order reconstruction in the sense that it provides information on the *tangent bundle* of the object surface, i.e. on both 3D coordinates and local orientation of the surface patches.

In Section 2, we summarize some facts from projective geometry that are used throughout the article. In particular, a projective model of the camera, a radiometric description of the object surface and some facts from the geometry of multiple views are briefly discussed. Section 3 is devoted to the characterization of the process of backprojection of a luminance profile onto the object surface and its re-projection onto another view. This whole operation is here described, under appropriate conditions, as a single mapping from view to view, and takes into account both geometric and radiometric model of the object surface. The 3D area matching problem is approached in Section 4. A definition for the correlation between an actual and a corresponding transferred image is here provided and used. In Section 5 we describe the implementative aspects of the 3D reconstruction system that we used for testing the 3D reconstruction strategy of Section 4. The results of the proposed 3D reconstruction method and of its accuracy are presented in Section 6 for a variety of real images. Section 7 contains conclusive considerations and proposals on future developments. A list of symbols adopted in this article is also included.

## 2. Preliminaries

The tools that are used in this article are basically all from projective geometry. In this section we will briefly summarize some fact and results that will prove useful in the following sections. The readers who are

already familiar with projective geometry, may skip this part, while those who would like to know more about it may refer, for example, to [1,7,13,15].

### 2.1. Projective spaces

The projective space $\mathscr{P}^n$ of dimension $n$ is defined as the quotient space of $\mathscr{R}^{n+1} - \mathbf{0}_{n+1}$, with respect to the equivalence relation

$$\boldsymbol{x} \sim \boldsymbol{x}' \Leftrightarrow \exists \gamma \neq 0 : \boldsymbol{x}' = \gamma \boldsymbol{x}.$$

$\mathscr{P}^n$ can thus be thought of as the set of all lines passing through the origin of $\mathscr{R}^{n+1}$.

The coordinates of an object point can be given either in $\mathscr{R}^3$ or in $\mathscr{P}^3$. The projective (or homogeneous) coordinates of a point of $\mathscr{P}^3$ are expressed as a quadruple of the form $\boldsymbol{x} = [x_1, x_2, x_3, x_4]^{\mathrm{T}}$ and, by definition, can be arbitrarily scaled. As far as image points are concerned, they can be defined either on $\mathscr{R}^2$ or $\mathscr{P}^2$.

In general, when dealing with camera geometry, it is often convenient to work on $\mathscr{P}^3$ and $\mathscr{P}^2$ rather than on $\mathscr{R}^3$ and $\mathscr{R}^2$, respectively. In fact, the camera models incorporates perspective projections, which are conveniently modeled as projective transformations, i.e. transformations action on projective spaces.

A projective transformation acting on $\mathscr{P}^n$ is called a *homography* when it is linear (in projective coordinates) and invertible. As a consequence, a homography is described by a non-singular matrix $\boldsymbol{H} \in \mathscr{R}^{(n+1)\times(n+1)}$, that maps the point $\boldsymbol{x}$ onto the point $\boldsymbol{x}' = \boldsymbol{H}\boldsymbol{x}$. Homographies map any projective subspace onto a projective subspace of the same dimension, and they form a transformation group $\mathscr{GL}_n$, called general linear group.

A point $\boldsymbol{s} \in \mathscr{P}^n$ can also define a hyperplane in $\mathscr{P}^n$, which is given by all points $\boldsymbol{x} \in \mathscr{P}^n$ such that $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x} = 0$. Hyperplanes are subspaces of the projective space.

### 2.2. Projective cameras

The camera model adopted in this article is basically a pinhole to which a non-linear stretching of the image plane is applied in order to take the geometric distortion of the optics into account. This camera model can be completed with additional filtering, to account for the aperture of optics and CCD sensor. The model's bottleneck, however, is always the perspective projection, which will be the focus of this section.

A pinhole model performs a projection of the object point, through the optical center of the camera, onto the retinal plane. The relationship between image and object coordinates is linear projective and is specified by a rank-3 projection matrix $\boldsymbol{P} \in \mathscr{R}^{3\times 4}$,

$$\boldsymbol{u} = \boldsymbol{P}\boldsymbol{x}, \quad \boldsymbol{u} \in \mathscr{P}^2, \quad \boldsymbol{x} \in \mathscr{P}^3. \tag{1}$$

The projection matrix $\boldsymbol{P}$ can be easily derived from the geometry of the acquisition system. Let $\boldsymbol{x} \in \mathscr{P}^3$ and $\boldsymbol{x}' \in \mathscr{P}^3$ be, respectively, the projective world coordinates and the projective camera coordinates of a point in object space. Let $\boldsymbol{O} \in \mathscr{R}^3$ be the Euclidean world coordinates of the origin of the camera frame. The change of reference frame from world coordinates to camera coordinates can be immediately written as

$$\boldsymbol{x}' = \boldsymbol{T}_{\mathrm{wc}}\boldsymbol{x}, \quad \boldsymbol{T}_{\mathrm{wc}} = \begin{bmatrix} \boldsymbol{R} & -\boldsymbol{R}\boldsymbol{O} \\ 0\,0\,0 & 1 \end{bmatrix}, \tag{2}$$

$\boldsymbol{R} \in \mathrm{SO}(3)$ being the rotation matrix that describes the orientation of the camera frame in world coordinates. Eq. (2) corresponds to the more familiar relationship $\boldsymbol{X}' = \boldsymbol{R}(\boldsymbol{X} - \boldsymbol{O})$, where $\boldsymbol{X} \in \mathscr{R}^3$ and $\boldsymbol{X}' \in \mathscr{R}^3$ are the Euclidean coordinates of $\boldsymbol{x}$ and $\boldsymbol{x}'$, respectively.

Without loss of generality, the camera frame can be chosen in such a way that its origin $\boldsymbol{O}$ corresponds to the optical center and that the $x'_3$ axis is perpendicular to the image plane. With this choice of camera coordinates, the projection becomes purely perspective [15], therefore it can be performed by using the matrix

$$\boldsymbol{T}_{\mathrm{pr}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f & 0 \end{bmatrix},$$

$f$ being the focal length of the camera (distance between optical center and image plane).

In conclusion, the projection matrix $\boldsymbol{P}$ of Eq. (1) is given by

$$\boldsymbol{P} = \boldsymbol{T}_{\mathrm{pr}}\boldsymbol{T}_{\mathrm{wc}} = \begin{bmatrix} \boldsymbol{r}_1 & -\boldsymbol{r}_1\boldsymbol{O} \\ \boldsymbol{r}_2 & -\boldsymbol{r}_2\boldsymbol{O} \\ (1/f)\boldsymbol{r}_3 & -(1/f)\boldsymbol{r}_3\boldsymbol{O} \end{bmatrix}, \tag{3}$$

where $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$ are the rows of the rotation matrix $\boldsymbol{R}$.

The projective image coordinates $\boldsymbol{u} = [u_1, u_2, u_3]^{\mathrm{T}}$ can now be scaled in order to obtain the image coordinates $\boldsymbol{v} = [u_1/u_3, u_2/u_3, 1]^{\mathrm{T}}$, which are expressed, for example, in mm.

Notice that, in order to obtain *pixel* coordinates, we need to perform an extra 2D translation combined with a scale change. This transformation can be chained to that of Eq. (3) as follows:

$$\boldsymbol{P}_i = \boldsymbol{T}_i\boldsymbol{P}, \quad \boldsymbol{T}_i = \begin{bmatrix} 1/d_1 & 0 & -t_1 \\ 0 & 1/d_2 & -t_2 \\ 0 & 0 & 1 \end{bmatrix},$$

where the scale factors $d_1$ and $d_2$ represent the horizontal and vertical size of the pixel, respectively, while $t_1$ and $t_2$ represent the offset of the principal point, i.e. of the intersection between image plane and optical axis. The projective coordinates $\bar{\boldsymbol{v}}$, obtained through normalization of $\bar{\boldsymbol{u}} = \boldsymbol{P}_i\boldsymbol{x}$, are in fact expressed in pixel.

In the above description, nothing is said about the geometrical distortions introduced by the optical system. As a matter of fact, the predicted position of the projected point does not correspond to the actual one because of lens distortion [2], which can be seen as non-linear *stretching* of the image plane. Image points are, in fact, shifted from the perspective projection's location depending on the position on the image plane. Lens distortion is of crucial importance when dealing with applications of 3D measurement and reconstruction, and neglecting it may cause to serious reconstruction errors.

The shift introduced by lens distortion has a *radial* component and a *tangential* component with respect to the principal point, which is the intersection between optical axis and image plane. In most cases, however, the tangential component results as being negligible with respect to the radial one [12].

The magnitude of the shift due to radial distortion is quite a complex function of the position on the image plane. The radial shift we need to apply to the image coordinates in order to obtain the undistorted coordinates is most frequently expressed as a truncated power series of the form

$$r_{\mathrm{u}} = r_{\mathrm{d}}(1 + k_3 r_{\mathrm{d}}^2 + k_5 r_{\mathrm{d}}^4 + \cdots), \tag{4}$$

where $r_{\mathrm{d}}$ and $r_{\mathrm{u}}$ are the distances from the principal point of the *distorted* and *undistorted* image points, respectively. Usually, only the first one or two coefficients of Eq. (4) are considered, depending on the application [29].

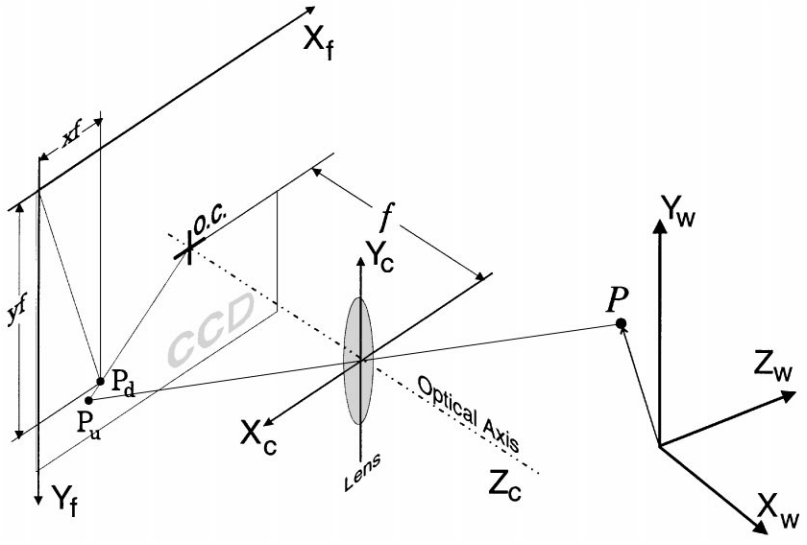


Fig. 1. Projective camera model.

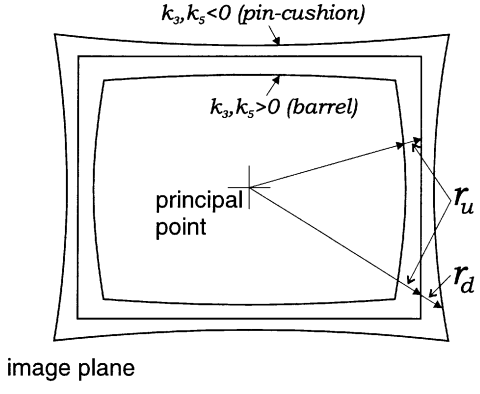


Fig. 2. Impact of the radial distortion on a square.

A global scheme of the adopted camera model is shown in Fig. 1.

Fig. 2 shows the impact of radial distortion on a square. Depending on the sign of the coefficients we obtain a *barrel* or *pin-cushion* type of distortion.

In the following sections we will assume that the image coordinates that we work with are all undistorted, which means that all image coordinates must be previously mapped onto undistorted ones through Eq. (4). Clearly, the inverse mapping will be required as well. However, as Eq. (4) cannot be easily written in explicit form with respect to $r_d$, this inversion needs to be performed iteratively. If only the first coefficient of the power series is considered, then Eq. (4) can be truncated to the third order and solved in closed-form by means of the *Cartan formula.*

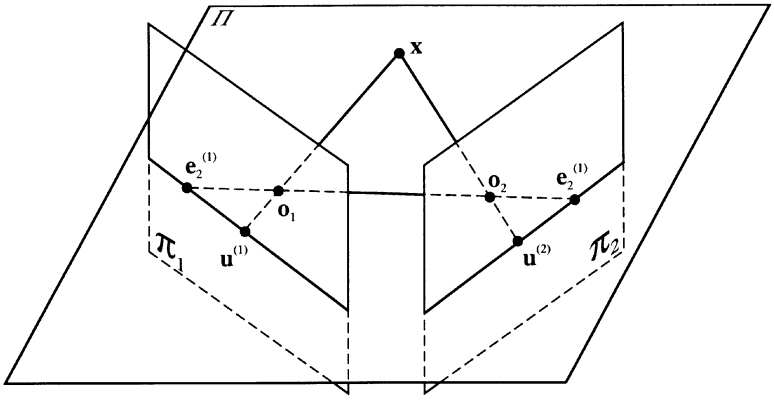


Fig. 3. Epipolar constraint: the optical rays are bound to be coplanar. $\Pi$ is the epipolar plane while $\pi_1$ and $\pi_2$ are the retinal planes.

### 2.3. Some facts from epipolar geometry

Two projective views of a point in object space, are bound to comply with the so-called "epipolar" (or "essential") constraint, according to which the two lines that connect the object point with the optical centers of the two projective cameras are coplanar (Fig. 3). The plane identified by the object point and the two optical centers is called *epipolar plane*. The intersection between the epipolar plane and an image plane is called *epipolar line*, and represents the projective view of the other optical ray. The intersection between an image plane and the line that connects the two optical centers is called *epipole*: all epipolar lines meet at that point, as all epipolar planes have the line that connects the optical centers in common.

Let $\boldsymbol{X}^{(1)} \in \mathscr{R}^3$ and $\boldsymbol{X}^{(2)} \in \mathscr{R}^3$ be the Cartesian coordinates of a point $\boldsymbol{X} \in \mathscr{R}^3$, referred to the coordinate frames of two different projective cameras. Let $\boldsymbol{R}_{21}$ and $\boldsymbol{T}_{21}$ be, respectively, the rotation matrix and the translation vector that describe the change of reference frame from camera 1 to camera 2. We thus have

$$\boldsymbol{X}^{(2)} = \boldsymbol{R}_{21}\boldsymbol{X}^{(1)} + \boldsymbol{T}_{21}.$$

The coplanarity of the two lines that connect the object point with the optical centers of the cameras can be expressed in terms of the orthogonality between $\boldsymbol{X}^{(2)}$ and the normal to the plane formed by $\boldsymbol{T}_{21}$ and $\boldsymbol{X}^{(1)}$. In matrix notation we have

$$(\boldsymbol{X}^{(2)})^{\mathrm{T}}\boldsymbol{E}_{21}\boldsymbol{X}^{(1)} = 0,$$

where $\boldsymbol{E}_{21} = \boldsymbol{T}_{21}\boldsymbol{R}_{21} = (\boldsymbol{t}_{21}\times)\boldsymbol{R}_{21} \in \mathscr{R}^{3\times 3}$ is called *essential matrix*.

Let $\boldsymbol{u}^{(1)} = \boldsymbol{P}^{(1)}\boldsymbol{x} \in \mathscr{P}^2$ and $\boldsymbol{u}^{(2)} = \boldsymbol{P}^{(2)}\boldsymbol{x} \in \mathscr{P}^2$ be the projective coordinates of a point $\boldsymbol{x} \in \mathscr{P}^3$, as seen by the two projective cameras, assuming that their projection matrices are $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$, respectively. As these projective coordinates are, up to a scale factor, equal to $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, respectively, the essential constraint can be rewritten in $\mathscr{P}^2$ as follows:

$$(\boldsymbol{u}^{(2)})^{\mathrm{T}}\boldsymbol{E}_{21}\boldsymbol{u}^{(1)} = 0. \tag{5}$$

The epipolar constraint (5) can be easily rewritten in image coordinates. For example, if the camera frames are chosen in such a way for the projections to result as perspective, then Eq. (5) becomes

$$(\boldsymbol{v}^{(2)})^{\mathrm{T}}\boldsymbol{E}_{21}\boldsymbol{v}^{(1)} = 0, \tag{6}$$

where

$$\boldsymbol{v}^{(1)} = \left[\frac{u_1^{(1)}}{u_3^{(1)}}, \frac{u_1^{(1)}}{u_3^{(1)}}, 1\right]^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{v}^{(2)} = \left[\frac{u_1^{(2)}}{u_3^{(2)}}, \frac{u_2^{(2)}}{u_3^{(2)}}, 1\right]^{\mathrm{T}}.$$

If the cameras are not perspective [15], then it is always possible to make them perspective through appropriate homographies $\boldsymbol{H}^{(1)} \in \mathscr{R}^{3\times 3}$ and $\boldsymbol{H}^{(2)} \in \mathscr{R}^{3\times 3}$,

$$(\boldsymbol{v}^{(2)})^{\mathrm{T}} \boldsymbol{Q}_{21} \boldsymbol{v}^{(1)} = 0, \qquad \boldsymbol{Q}_{21} = (\boldsymbol{H}^{(2)})^{\mathrm{T}} \boldsymbol{E}_{21} \boldsymbol{H}^{(1)}. \tag{7}$$

Notice that Eq. (7) provides us with a closed-form equation for the epipolar line on either one of the two cameras, associated to a point on the other camera. For example, by letting $\boldsymbol{w} = \boldsymbol{Q}_{21}\boldsymbol{v}^{(1)}$, Eq. (7) becomes

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{v}^{(2)} = 0, \tag{8}$$

which is the equation of the epipolar line on camera 2, corresponding to the image point $\boldsymbol{v}^{(1)}$.

### 2.4. Multi-ocular invariance

The epipolar constraint provides us with a tool for checking the correctness of a matching between homologous features in three or more views.

Given a point $\boldsymbol{x}$ in the object space, its image coordinates $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(n)}$ on the $n$ available views are bound to satisfy the epipolar constraint pairwise,

$$(\boldsymbol{v}^{(i)})^{\mathrm{T}} \boldsymbol{Q}_{ij} \boldsymbol{v}^{(j)} = 0, \qquad \boldsymbol{Q}_{ij} = (\boldsymbol{H}^{(i)})^{\mathrm{T}} \boldsymbol{E}_{ij} \boldsymbol{H}^{(j)}, \quad i,j = 1,\ldots,n,\ i > j. \tag{9}$$

If, conversely, we are considering the image coordinates of a point in each one of the available images, one can check for the correctness of the matching by checking whether the epipolar constraint is satisfied for all pairs of views. This operation corresponds to checking whether each image point lies on the intersection of the epipolar lines corresponding to the considered points on the other views, as shown in Fig. 4. Notice that three is the minimum number of views that allows us to check for this type of invariance. More than three views certainly provide us with more information but the problem becomes overconstrained (intersection of more than two epipolar lines), therefore we need to use least square techniques.

The above property can be seen as a form of point-wise multi-occular invariance. Other types of invariance properties for checking on the correctness of a match between image features can be found. For example, the equivalent of Eq. (9) for the case of lines is derived and described in [9].

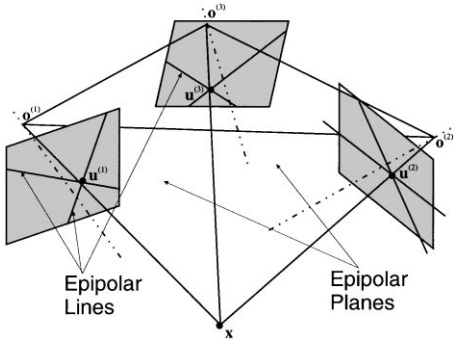


Fig. 4. Visualization of the point-wise multiocular invariance constraint in the case of three cameras: a point in each view must lie on the intersection of the epipolar lines associated to the homologous points in the other views.

### 2.5. Radiometric models

The luminance information associated to a camera view depends, in general, on the physical and geometrical properties of the object surface and of the illumination conditions. Several radiometric models are available for describing the viewing process of surfaces [4,16,28]. The most popular one is the Lambertian model, according to which surface reflectivity depends on the direction of illumination but not on the viewing direction. As a matter of fact, Lambertian surfaces are not very common, as specular reflections are very often present. Nonetheless matt surfaces are very well rendered through this type of models. Non-Lambertian reflectivity models are characterized by the fact that, besides a viewer-independent reflection lobe, they exhibit a viewer-dependent (specular) reflection lobe.

The radiometric model we refer to in this article is based on that of Torrance and Sparrow [28], with minor variations. In short, we assume that the environment's lighting is composed of a dominant illuminator at infinite distance and a certain amount of diffused light. Furthermore, surfaces are assumed as being only partially matt and never totally specular. These two assumptions amount to a reflectivity function $r$ of the following form:

$$r = a\cos\vartheta_i + k\frac{e^{-\alpha^2/2\sigma^2}}{\cos\vartheta_r} + d, \tag{10}$$

where $a$ is the local albedo map, which depends on the surface point and completely characterizes the surface texture,

$$\vartheta_i = \arccos\left(\frac{\boldsymbol{i}^{\mathrm{T}}\boldsymbol{n}}{\|\boldsymbol{i}\|\,\|\boldsymbol{n}\|}\right)$$

is the angle between surface normal $\boldsymbol{n}$ and the direction $\boldsymbol{i}$ of the dominant incident light,

$$\vartheta_r = \arccos\left(\frac{\boldsymbol{r}^{\mathrm{T}}\boldsymbol{n}}{\|\boldsymbol{r}\|\,\|\boldsymbol{n}\|}\right)$$

is the angle between surface normal $\boldsymbol{n}$ and the viewing direction $\boldsymbol{r}$ (third column of the rotation matrix that maps world coordinates onto camera coordinates), finally

$$\alpha = \arccos\left(\frac{(\boldsymbol{r}\times\boldsymbol{i})\times\boldsymbol{n}}{\|\boldsymbol{r}\times\boldsymbol{i}\|\,\|\boldsymbol{n}\|}\right)$$

is the angle between the surface normal $\boldsymbol{n}$ and the plane corresponding to the incident light $\boldsymbol{i}$ and viewing direction $\boldsymbol{r}$. The first term of Eq. (10) represents the viewer-independent (Lambertian) component, the second term is the specular (non-Lambertian) reflection lobe and the third one is a constant term that accounts for the diffused light. Size and shape of the specular reflection lobe associated to the surface material decided by $k$ and $\sigma$, respectively.

The above radiometric model is suitable for realistically describing a wide variety of surfaces and illumination conditions.

### 2.6. Surface patches

In this article, *a surface patch* $\mathscr{S}$ is intended as a smooth (or at least $C^2$) open and compact portion of a two-dimensional manifold $\mathscr{M}$ (object surface), embedded in $\mathscr{P}^3$. If we are viewing $\mathscr{M}$ through a projective camera that *sees* the whole surface patch $\mathscr{S}$, then we know that there is a one-to-one correspondence between

$\mathscr{S} \subset \mathscr{M}$ and its projection $\mathscr{S}' \subset \mathscr{P}^2$ on the image plane. In other words, the projection defines a map between $\mathscr{P}^2$ and $\mathscr{S}$

$$\begin{aligned} \boldsymbol{\mu}: \quad & \mathscr{P}^2 \to \mathscr{M} \\ & \boldsymbol{v} \in \mathscr{S}' \mapsto \boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{v}) \in \mathscr{S}, \end{aligned} \tag{11}$$

which is called *depth map*. Eq. (11) describes the surface patch $\mathscr{S}$ in parametric form and induces a smooth local parametrization on $\mathscr{M}$. If we have enough projective views to cover the whole object surface $\mathscr{M}$, then the set of all parametrizations can be used for constructing an atlas on $\mathscr{M}$.

The tangent plane $T\mathscr{M}$ of $\mathscr{M}$ at any point $\boldsymbol{x} \in \mathscr{S}$ is the vector space generated by the vectors

$$\boldsymbol{m}_1 = \partial\boldsymbol{\mu}/\partial v_1, \qquad \boldsymbol{m}_2 = \partial\boldsymbol{\mu}/\partial v_2,$$

therefore any tangent vector $\boldsymbol{m} \in T\mathscr{M}$ can be expressed as a linear combination of the form

$$\boldsymbol{m} = \alpha\boldsymbol{m}_1 + \beta\boldsymbol{m}_2,$$

while the normal to $T\mathscr{M}$, and therefore to $\mathscr{M}$, is given by

$$\boldsymbol{n} = \frac{\boldsymbol{m}_1 \times \boldsymbol{m}_2}{\|\boldsymbol{m}_1 \times \boldsymbol{m}_2\|}.$$

When two different projective cameras see the whole surface patch $\mathscr{S}$, the fact that $\mathscr{M}$ is a manifold guarantees the existence of a diffeomorphism (invertible map) between the views $\mathscr{S}^{(1)}$ and $\mathscr{S}^{(2)}$ of $\mathscr{S}$. The determination of this map is the focus of the next section.

## 3. Luminance transfer

Performing area matching requires comparing actual luminance profiles with those that we would obtain by *transferring* luminance profiles of other views, through a specific 3D surface model. How to perform this luminance transfer, given the object surface, will be discussed in this section.

Before doing so, let us consider the case in which no information is available about the surface patch $\mathscr{S}$, and two projective cameras view this patch under the conditions described in Section 2.6. The relationship between the Cartesian coordinates $\boldsymbol{X}^{(1)} \in \mathscr{R}^3$ and $\boldsymbol{X}^{(2)} \in \mathscr{R}^3$ of points on the two views can be immediately derived from the fact that

$$\begin{aligned} \boldsymbol{X}^{(1)} &= \boldsymbol{R}^{(1)}(\boldsymbol{X} - \boldsymbol{O}^{(1)}), \\ \boldsymbol{X}^{(2)} &= \boldsymbol{R}^{(2)}(\boldsymbol{X} - \boldsymbol{O}^{(2)}), \end{aligned} \tag{12}$$

where $\boldsymbol{X}$, $\boldsymbol{O}^{(1)}$ and $\boldsymbol{O}^{(2)}$ are the Cartesian world coordinates of the object point and the optical centers and $\boldsymbol{R}^{(2)}$ and $\boldsymbol{R}^{(2)}$ are the orientation of the camera frames in world coordinates. From Eq. (12) we derive the relationship between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ as

$$\begin{aligned} \boldsymbol{X}^{(2)} &= \boldsymbol{R}_{21}\boldsymbol{X}^{(1)} + \boldsymbol{D}_{21}, \\ \boldsymbol{X}^{(1)} &= \boldsymbol{R}_{12}\boldsymbol{X}^{(2)} + \boldsymbol{D}_{12}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} \boldsymbol{R}_{21} &= \boldsymbol{R}^{(2)}(\boldsymbol{R}^{(1)})^{\mathrm{T}} = \boldsymbol{R}_{12}^{T}, \\ \boldsymbol{D}_{21} &= \boldsymbol{R}^{(2)}(\boldsymbol{O}^{(1)} - \boldsymbol{O}^{(2)}), \\ \boldsymbol{D}_{12} &= \boldsymbol{R}^{(1)}(\boldsymbol{O}^{(2)} - \boldsymbol{O}^{(1)}). \end{aligned}$$

The transformation (13) acts on Cartesian coordinates and, because of the translational terms $\boldsymbol{D}_{21}$ and $\boldsymbol{D}_{12}$, is not scalable, therefore it cannot be seen as a transformation between points of $\mathscr{P}^2$. This fact should not surprise as, without any information on the object surface, the transfer between image coordinates cannot be determined. Conversely, when the 3D surface is known, then it is possible to determine the transfer between image coordinates in closed form.

### 3.1. Mapping between projective views

As already mentioned above, in order to characterize the transfer between points in different views, we need to know the geometry of the object surface. A general description of the object surface could be given through an implicit equation of the form $g(\boldsymbol{x}) = 0$. This type of description, however, does not easily allow us to derive a map between image coordinates in different views, unless linearity is assumed (planar surface). We will see later that this type of description can be sufficient for most of the cases of interest. In fact, when the surface is a $C^2$ manifold, being locally flat, it can be well-described by its *tangent bundle*.

Let us assume the patch $\mathscr{S}$ to be planar, i.e. lying on a plane of equation $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x} = 0$, where $\boldsymbol{s} = [s_1\ s_2\ s_3\ 1]^{\mathrm{T}}$, $\boldsymbol{n} = [s_1\ s_2\ s_3]^{\mathrm{T}}$ being the normal to the patch. Let us assume, without loss of generality, that $s_3 \neq 0$. In this case the equation of the plane can be rewritten in the form $x_3 = -(1/s_3)(s_1x_1 + s_2x_2 + 1)$, and the projective coordinates

$$\boldsymbol{u}^{(m)} = \boldsymbol{K}^{(m)}[\boldsymbol{R}^{(m)} \quad -\boldsymbol{R}^{(m)}\boldsymbol{O}^{(m)}]\boldsymbol{x}, \qquad \boldsymbol{K}^{(m)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/f^{(m)} \end{bmatrix}$$

of the $m$th view of the surface point $\boldsymbol{x}$ can be rewritten in such a way to eliminate $x_3$,

$$\boldsymbol{u}^{(m)} = \boldsymbol{M}^{(m)}(\boldsymbol{s}) \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix},$$

where

$$\boldsymbol{M}^{(m)}(\boldsymbol{s}) = \begin{bmatrix} r_{11}^{(m)} - \frac{1}{s_3}r_{13}^{(m)}s_1 & r_{12}^{(m)} - \frac{1}{s_3}r_{13}^{(m)}s_2 & -(\frac{1}{s_3}r_{13}^{(m)} + \boldsymbol{r}_1^{(m)}\boldsymbol{O}^{(m)}) \\ r_{21}^{(m)} - \frac{1}{s_3}r_{23}^{(m)}s_1 & r_{22}^{(m)} - \frac{1}{s_3}r_{23}^{(m)}s_2 & -(\frac{1}{s_3}r_{23}^{(m)} + \boldsymbol{r}_2^{(m)}\boldsymbol{O}^{(m)}) \\ \frac{1}{f^{(m)}}(r_{31}^{(m)} - \frac{1}{s_3}r_{33}^{(m)}s_1) & \frac{1}{f^{(m)}}(r_{32}^{(m)} - \frac{1}{s_3}r_{33}^{(m)}s_2) & -\frac{1}{f^{(m)}}(\frac{1}{s_3}r_{33}^{(m)} + \boldsymbol{r}_3^{(m)}\boldsymbol{O}^{(m)}) \end{bmatrix}, \tag{14}$$

which depends on the position and the orientation of the planar patch.

If we assume that all points of $\mathscr{S}$ are visible from the $i$th and $j$th projective cameras, i.e. that the planar patch lies on a plane that does not contain any one of their two optical centers, then both matrices $\boldsymbol{M}^{(i)}(\boldsymbol{s})$ and $\boldsymbol{M}^{(j)}(\boldsymbol{s})$ result as being invertible. As a consequence, the mapping between views can be expressed as

$$\boldsymbol{u}^{(j)} = \boldsymbol{M}_{ji}(s)\boldsymbol{u}^{(i)}, \qquad \boldsymbol{u}^{(i)} = \boldsymbol{M}_{ij}(s)\boldsymbol{u}^{(i)}, \tag{15}$$

where

$$\boldsymbol{M}_{ji}(s) = \boldsymbol{M}^{(j)}(\boldsymbol{s})(\boldsymbol{M}^{(i)}(\boldsymbol{s}))^{-1} = (\boldsymbol{M}_{ij}(s))^{-1}. \tag{16}$$

In conclusion, the exact knowledge of the planar surface allows us to specify the mapping between views as a homography.

Notice that in the case of a non-planar patch the relationship between projective coordinates on the two image planes is no longer linear.

### 3.2. Luminance transfer

Let $\mathscr{S}$ be a patch in object space, obtained by backprojecting a reference patch of any of the views on the surface $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x}=0$, and let $\mathscr{S}^{(i)}$ be its $i$th view. As already seen in Section 3.1, the transfer of projective coordinates from the $j$th view to the $i$th view through the plane $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x}=0$ is described by a homography of the form

$$\boldsymbol{u}^{(i)}=\boldsymbol{M}_{ij}(\boldsymbol{s})\boldsymbol{u}^{(j)},$$

where $\boldsymbol{M}_{ij}(\boldsymbol{s})$ is given by Eq. (16).

Let $I^{(i)}$ and $I^{(j)}$ be the luminance profiles of the views $i$ and $j$, respectively. With reference to Eq. (10), the luminance $I^{(i)}(\boldsymbol{u}^{(i)})$ associated to the projective image coordinates $\boldsymbol{u}^{(i)}$, can be written as

$$I^{(i)}(\boldsymbol{u}^{(i)})=a(\boldsymbol{u}^{(i)})g(\boldsymbol{s},\boldsymbol{i})+h(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma)+d, \tag{17}$$

whose first term is the Lambertian (viewer-independent) component, in which $a(\boldsymbol{u}^{(i)})$ is the albedo map,

$$g(\boldsymbol{s},\boldsymbol{i})=\frac{\boldsymbol{i}^{\mathrm{T}}\boldsymbol{n}}{\|\boldsymbol{i}\|\,\|\boldsymbol{n}\|},\quad \boldsymbol{n}=\begin{bmatrix}s_1\\ s_2\\ s_3\end{bmatrix}, \tag{18}$$

and $\boldsymbol{i}$ is the direction of illumination. The term $h(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma)$ represents the specular (viewer-dependent) component (see Eq. (10)) and $d$ is the diffuse lobe.

The luminance transfer from image $j$ to image $i$, through the surface $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x}=0$, can be derived from Eq. (17), by taking into account that $a(\boldsymbol{v}^{(i)})=a(\boldsymbol{v}^{(j)})$ (texture transfer). We obtain

$$I^{(i)}_{j}(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma)=I^{(j)}(\boldsymbol{M}_{ji}(\boldsymbol{s})\boldsymbol{u}^{(i)})+\Delta^{(i)}_{j}(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma), \tag{19}$$

where

$$\Delta^{(i)}_{j}(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma)=h(\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma)-h(\boldsymbol{M}_{ji}(\boldsymbol{s})\boldsymbol{u}^{(i)},\boldsymbol{s},\boldsymbol{i},k,\sigma) \tag{20}$$

is the corrective term that compensates for the variations of the viewer-dependent portion of the luminance (migration of reflexes with viewpoint changes). Notice that the term (18) does not appear in Eq. (19) because it is not viewer dependent.

## 4. 3D area matching principles

In general, we can look at correlation-based 3D reconstruction methods as those that determine a 3D surface and its reflectivity properties whose projective views result as close as possible to the actual views. More precisely, this inverse problem can be thought of as that of determining the parameters of a geometric and radiometric model of the surface, which maximize a similarity measure (such as the *correlation*) between actual views and *transferred* versions of the other views. A *global* optimization approach, however, is definitely not feasible because of size of the search space and the fact that such a non-linear search would almost certainly converge to a local minimum rather than the global minimum.

In order to limit the search space and reduce the risk of converging to a relative minimum, it is customary to proceed with a *local* approach, under specific stereometric constraints. Acting *locally* means describing the

whole surface as a patchwork of smaller surfaces, each one of which determined though a matching of luminance profiles of *homologous* image regions in different views. Two image regions in different views are considered as *homologous* when they are projective views of the same 3D surface patch. When the object surface is unknown, verifying whether two image regions are homologous can be a rather difficult task, which requires to take the geometry of the projective cameras into account, and to cope with possible matching ambiguities through proper invariance constraints.

In order to be able to find homologous regions on the views of a multi-camera system with arbitrary geometry, we need to take the perspective distortion of the image region into account. In order to do so, we can perform area matching in object space rather than on the images.

Let us consider a patch in object space, which lies on a parametric surface. This patch is a good approximation of the object surface when there is a match between the *back-projection* of all the corresponding image regions onto the 3D patch. Notice that the match needs to be found through texture comparison, therefore back-projecting an image region onto a 3D patch must be intended as *painting* the albedo map on the surface patch. In conclusion, area matching consists of looking for the parameters of the parametric surface where the patch lies and those of the radiometric model, which maximize the correlation between the back-projected corresponding image regions.

The correlation can be computed indifferently on the planar surface in the 3D space, or on either one of the retinal planes. In this last case we need to characterize the luminance transfer from view to view. As already seen in Section 3.1, the transfer between points in different views, can be easily modeled only when the patch is planar. On the other hand, as the surface is assumed as being $C^2$, it can be well-described by its *tangent bundle*. As a consequence, if the surface patch is small enough, we can choose the parametric surface that it lies upon to be planar and characterize the luminance transfer as done in Section 3.2. We will thus look simultaneously for position and orientation of a locally planar 3D patch that originated in the corresponding image areas.

Let us assume that the portion of the object surface $\mathscr{M}$ that we want to reconstruct is being imaged by a set of projective cameras that satisfy the hypotheses of Section 2.6. In order to determine the tangent bundle of the imaged portion of $\mathscr{M}$, we need to find a way of scanning its surface. Under the hypothesis of Section 2.6, such an operation can be easily performed with reference to any of the available views. In fact, scanning the image with an image patch of pre-determined shape and size corresponds to scanning the visible portion of the manifold $\mathscr{M}$.

Let $\mathscr{S}$ be a patch in object space, obtained by backprojecting a reference patch of any of the views on the surface $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x} = 0$, and let $\mathscr{S}^{(i)}$ be its $i$th view. As already seen in Section 3, the luminance transfer from the $j$th view to the $i$th view through the plane $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x} = 0$ is given by

$$I_j^{(i)}(\boldsymbol{u}^{(i)}) = I^{(j)}(\boldsymbol{M}_{ji}(\boldsymbol{s})\boldsymbol{u}^{(i)}) + \Delta_j^{(i)}(\boldsymbol{u}^{(i)}, \boldsymbol{s}, \boldsymbol{i}, k, \sigma), \tag{21}$$

where the radiometric correction

$$\Delta_j^{(i)}(\boldsymbol{u}^{(i)}, \boldsymbol{s}, \boldsymbol{i}, k, \sigma) = h(\boldsymbol{u}^{(i)}, \boldsymbol{s}, \boldsymbol{i}, k, \sigma) - h(\boldsymbol{M}_{ji}(\boldsymbol{s})\boldsymbol{u}^{(i)}, \boldsymbol{s}, \boldsymbol{i}, k, \sigma)$$

is the luminance offset that accounts for reflex migration with viewpoint. Eq. (21) allows us to define a similarity function between original and transferred luminance profiles, to be maximized for 3D reconstruction.

In alternative, we can define an MSE-like cost function of the form

$$C_j^{(i)}(\boldsymbol{s}, \boldsymbol{i}, k, \sigma) = \int_{\mathscr{S}^{(i)}} |I^{(i)}(\boldsymbol{u}^{(i)}) - I_j^{(i)}(\boldsymbol{u}^{(i)})|^2 \, \mathrm{d}\boldsymbol{u}^{(i)}, \tag{22}$$

to be minimized with respect to $\boldsymbol{s}, \boldsymbol{i}, k$ and $\sigma$, in order to determine geometric and radiometric properties of an image patch. Notice that neither the diffused light component $d$ nor the albedo map $a$ appear in the cost function (22).

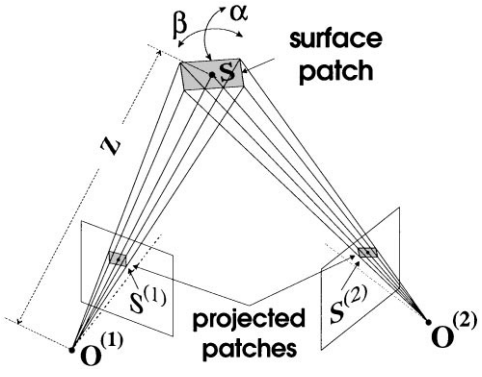


Fig. 5. Search for homologous image patches through minimization of a binocular cost function. Notice that, due to the perspective projection, the luminance transfer results in a distortion in the local texture and in the patch shape.

The above-described binocular area matching process is summarized in Fig. 5 where the distortion of local texture and patch shape due to perspective projection are emphasized.

In general, in order to make sure that the information extracted through the above optimization approach is correct, two images are not enough, as they do not allow us to apply the multi-ocular invariance property (see Section 2.4). The invariance property can be included into the optimization procedure by defining a cost function that incorporates the cost (22) computed on all pairs of views

$$C(\boldsymbol{s},\boldsymbol{i},k,\sigma)=\sum_{i}\sum_{j>i}C_j^{(i)}(\boldsymbol{s},\boldsymbol{i},k,\sigma). \tag{23}$$

For example, with three views we have

$$C(\boldsymbol{s},\boldsymbol{i},k,\sigma)=C_2^{(1)}(\boldsymbol{s},\boldsymbol{i},k,\sigma)+C_3^{(1)}(\boldsymbol{s},\boldsymbol{i},k,\sigma)+C_3^{(2)}(\boldsymbol{s},\boldsymbol{i},k,\sigma), \tag{24}$$

which can be maximized with respect to $\boldsymbol{s},\boldsymbol{i},k$ and $\sigma$.

The implementation of an area matching method based on the minimization of the cost function (24), with respect to position and orientation of the 3D surface patch and of the radiometric parameters, depends on the available a priori information on the shape of the objects to be reconstructed. For example, when there is a fair amount of information on the object, which consists of a rough approximation of its surface, it is possible to adopt an iterative optimization process such as the *steepest descent* algorithm [27], or the *downhill simplex* method [17], or the Levemberg–Marquardt method [27].

When the a priori information is poor, the optimization process will have to be combined with some method for sampling the parameter space in order to make the search simpler and safer. This case will be discussed in the next section.

## 5. Implementation

An assumption that is often made for area matching is that the distance between cameras is much smaller than the distance of the object from the acquisition system. Furthermore, the cameras are often assumed to have parallel or almost parallel optical axes. These two hypotheses allow us to ignore the geometric distortion that the texture within the image patch undergoes during the transfer from image to image. An immediate consequence of this fact is that the image patch preserves its shape when mapped onto the other

image, therefore the matching procedure can be implemented like a block-based correlation algorithm [30].

The above approach can be extended to the case in which the optical axes are not parallel, though the distance from the object is still assumed as being much greater than the distance between cameras. In this case it is possible to apply a homography to the projective image coordinates, which returns the image coordinates that we would have if we used parallel retinal planes. This image warping process is called *image rectification* [6,11].

Notice that, in both the above solutions, the assumption that the distance between cameras is much smaller than the distance of the object from the acquisition system is responsible for the fact that the orientation of the surface patch in object space cannot be retrieved from the images, and all that can be estimated is depth. In practice, the traditional area matching solutions can be thought of as *order-zero* 3D reconstruction methods, in the sense that only 3D coordinates are retrieved. What we propose, on the other hand, can be seen as an *order-one* approach as it provides us with the tangent bundle of the object surface. A first consequence of this fact is that a sparser set of matches can be retrieved without giving up reconstruction accuracy. In principle, it could be possible to work on *order-k* methods, with $k > 1$. In order to do so, however, we would have to deal with a nonlinear transfer between projective image coordinates, which would enormously complicate the problem.

The area matching approach presented in Section 4 is valid for an arbitrary camera geometry and for a generic number of cameras. In this section we will illustrate the specific solutions we adopted for implementing and testing the method.

### 5.1. Calibrated acquisition system

The acquisition system we adopted for our experiments is a calibrated trinocular camera system. We used a set of three standard TV-resolution CCD cameras (2/3″ CCD sensor, 16 mm lenses) mounted on a rigid frame in a triangular configuration (non-collinear optical centers), as shown in Fig. 6. In all our experiments, the distance from the object is kept comparable with the sides of the triangle (close-range reconstruction) and, as a consequence, the optical axes are strongly convergent in order to have the cameras pointing toward the object.

The frame-grabber used for digitalizing the analog signal generated by the cameras is synchronized with the camera pixel-clock in order to make sure that the pixel size of the images corresponds to the actual pixel

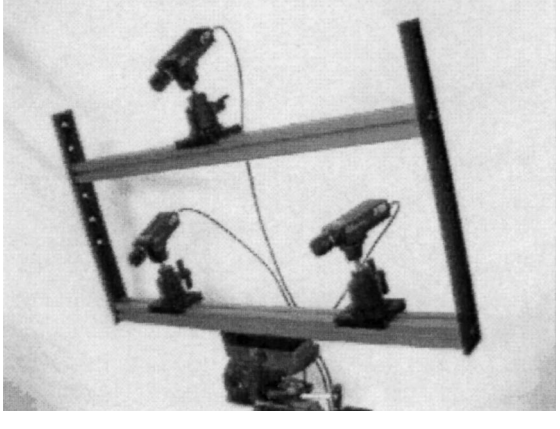

Fig. 6. Trinocular camera system used for the experiments.

size on the CCD sensor. Before beginning the acquisition process, camera calibration is performed by using an improved version of Tsai's calibration method [18,19,25,26], which includes fifth-order radial distortion and principal point estimation.

### 5.2. Implementation of the matching strategy

In order to speed up the computation of the optimization strategy described in Section 4, instead of minimizing the MSE-like cost function (24), we minimize a cost function of the form

$$D(\boldsymbol{s}, k_g, \Delta_j^{(i)}) = D_2^{(1)}(\boldsymbol{s}, k_g, \Delta_j^{(i)}) + D_3^{(1)}(\boldsymbol{s}, k_g, \Delta_j^{(i)}) + D_3^{(2)}(\boldsymbol{s}, k_g, \Delta_j^{(i)}),$$

where

$$D_j^{(i)}(\boldsymbol{s}, k_g, \Delta_j^{(i)}) = \int_{\mathscr{S}^{(i)}} |I^{(i)}(\boldsymbol{u}^{(i)}) - I_j^{(i)}(\boldsymbol{u}^{(i)})| \, \mathrm{d}\boldsymbol{u}^{(i)}$$

and

$$I_j^{(i)}(\boldsymbol{u}^{(i)}) = k_g I^{(j)}(\boldsymbol{M}_{ji}(\boldsymbol{s})\boldsymbol{u}^{(i)}) + \Delta_j^{(i)}, \tag{25}$$

$\Delta_j^{(i)}$ being the radiometric correction of Eq. (20), which is here assumed constant in the whole patch. An additional gain factor $k_g$ is included to account for inhomogeneity in the CCD sensor's sensitivity.

By choosing a luminance transfer function of the form (25), we give up on the radiometric parameters $\boldsymbol{i}, k$ and $\sigma$, in order to retrieve only the shape of the object surface.

As a general rule, we need to make sure that the maximum size of the patch is small enough to guarantee a limited error on the texture distortion. This choice, however, depends on the degree of smoothness of the surface to be reconstructed.

#### 5.2.1. Relative minima

The above area matching process is based on the minimization of a highly non-linear cost function. Therefore we can expect the process to be quite sensitive to the presence of relative minima. In order to avoid such a problem, several strategies can be adopted, the choice of which depends on the type of object surface to be reconstructed.

1. *Initialization through rough surface estimation*: the fastest and safest solution to the problem of relative minima consists of using an initial guess of the surface shape, which helps the minimization process converge to a global minimum and dramatically speeds up the matching process thanks to a dramatic reduction of the size of the search space. In principle, any method can be used for obtaining a first guess of the surface shape. For example, we could use an edge-based approach [5,21–23], whose reliability is guaranteed by the accuracy of the camera model and the calibration procedure. In this case the result is usually a sparse, though accurate, set of 3D points which, in order to obtain a first guess of the surface to be reconstructed, needs to be interpolated. Other initial surfaces could be extracted, for example, through the analysis of rims at occlusion boundaries [24,31], or through shape-from-shading methods [10,20] or, when more views are available, through volumetric intersection [8].
2. *Progressive-scan initialization*: when no initial information on the 3D structure of the surface is available at all, we can adopt a blind strategy whose robustness is paid by an increase of computational efficiency. The object space is *scanned* by a sequence of $n$ parallel planes whose distance from each other is $\Delta$. The area matching is first performed by taking as an initial surface the plane which is the closest to the camera system. If a 3D patch is estimated and its distance from the plane is not greater than a certain threshold

(which depends on $\Delta$), then the match is used for modifying the shape of the next plane. Roughly speaking, the result of the first estimate will be used for *bulging* the next surface. The area matching process is then iterated, each time using a progressively more bulged version of the plane as an initial guess of the object surface. We can also proceed in a more *parallel* fashion by repeating the area matching process (with narrow thresholds) each time with a different one of the parallel planes, each time estimating a different set of 3D points/normals. At the end we can merge all the estimates and perform surface interpolation only once.

3. *Multi-scale refinement*: in some cases the surface geometry is such that a multi-resolution approach can be adopted for 3D reconstruction without any initial information on the object surface. In these cases, we can perform an initial area matching with relatively large surface patches. After locating the surface patches in object space, we can perform surface interpolation and obtain a first rough approximation of the object surface. At this point the area matching process can start over with a smaller patch size and a reduced search space.

#### 5.2.2. Surface interpolation

As already mentioned earlier, the interpolation of estimated 3D points in object space is an important step in the reconstruction process [3]. This operation, in principle, could be done through standard interpolation of sparse data with non-uniform density, by taking any of the camera frame as a reference frame. This type of representation, however, may cause some problems of inconsistency, particularly with non-parallel optical axes. An alternative solution which overcomes such problems consists of directly generating a 3D surface which passes through all estimated 3D points.

The object surface may exhibit depth discontinuity about its boundaries while its normal may exhibit discontinuity about surface ridges. The depth function can thus be assumed as being smooth almost everywhere. In order to be able to interpolate surfaces with such characteristics, we need an interpolation algorithm that generates surfaces that are smooth everywhere except for some special locations that correspond to object ridges and boundaries (which can often be recognized through edges analysis). This type of interpolator, introduced by Mallet [14], it is known as Discrete Smooth Interpolation (DSI), and is based on a modification of the thin-plate spline algorithm. Its capability of preserving discontinuities is obtained through the specification of both local and global surface roughness parameters, which account for the presence of discontinuities in the neighborhood. An optimized version of this interpolator has been implemented and employed for surface interpolation.

## 6. Examples of application

Some experiments of 3D scene reconstruction have been made on a variety of test scenes. The first test concerned a measurement of the accuracy of the area matching method proposed in this article. The object to be reconstructed was a newspaper's page glued to a flat glass surface and viewed from tilted viewpoints. Fig. 7 shows the original views that we started with. The object was at about 1 m of distance from the camera system, and the camera baseline was about 0.5 m. Area matching was performed in one pass using image patches of $8 \times 8$ pixel. The reconstructed surface resulted as being flat within 0.15 mm of standard deviation (see Fig. 8). Considering patches of larger size it is possible to further increase this precision.

Another reconstruction experiment was performed on a large stone of the ruins of the Roman Amphitheater of Aosta, Italy. The acquisition was done in rather difficult conditions (camera frame mounted on a scaffold) and in non-controlled illumination conditions. In Fig. 9 one of the original views of the stone is shown. The object was at 1 m of distance from the camera system, whose baseline was 0.5 m. No initial guess for the surface was available, and a multi-scale refinement was performed for surface reconstruction. The retrieved points and the 3D reconstruction after texture mapping are shown in Figs. 10 and 11.

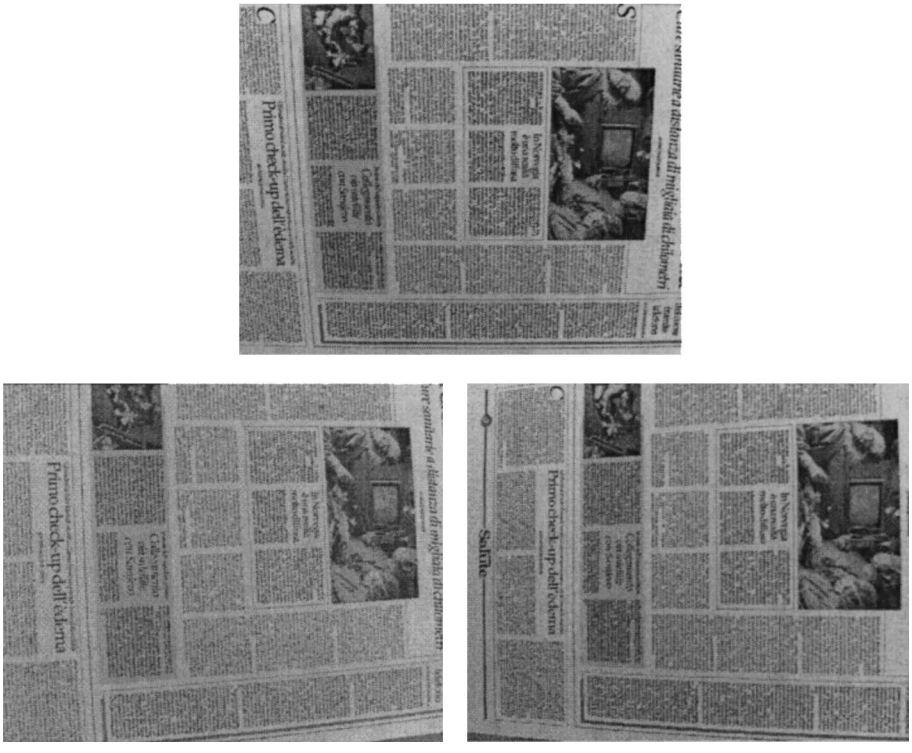


Fig. 7. Triplet of views (top, left and right) of a flat newspaper's page.

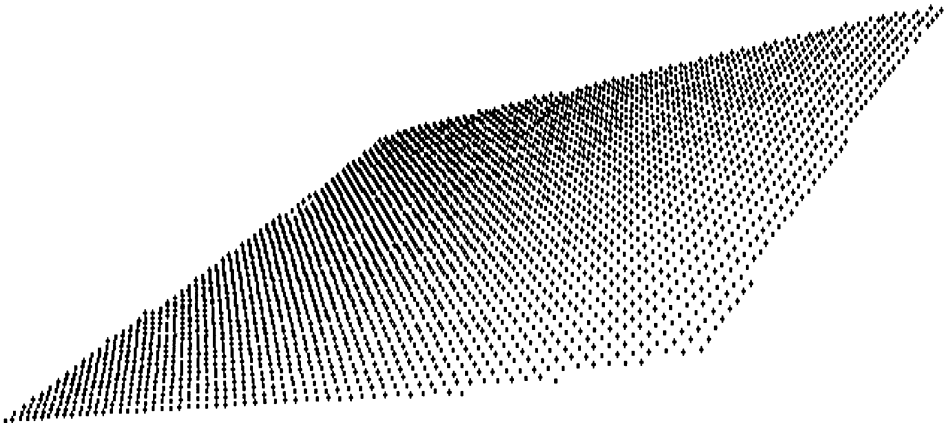

Fig. 8. Oblique view of a set of 3D points of the newspaper's page, retrieved through area matching. The standard deviation from planarity is 0.1 mm.

A quantitative evaluation of the quality of the results has been possible through photogrammetric techniques. The markers that are visible in Fig. 9 have in fact been placed for photogrammetric measurements. We found that our reconstruction results agreed with the measurements taken with classical photogrammetric methods with a maximum error of 1 mm. Notice, however, that part of the difference between or

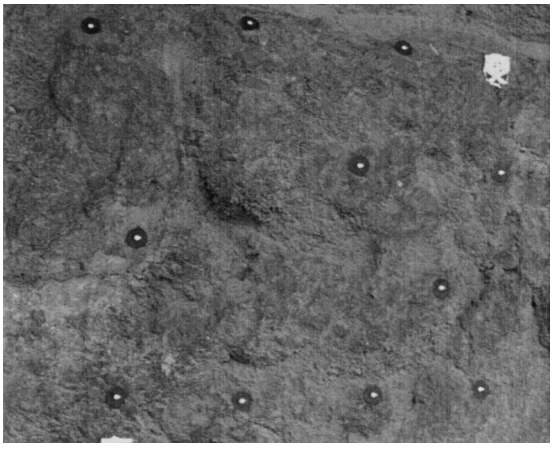

Fig. 9. One of the original views (left) of a stone of the Roman Amphitheater of Aosta, Italy.

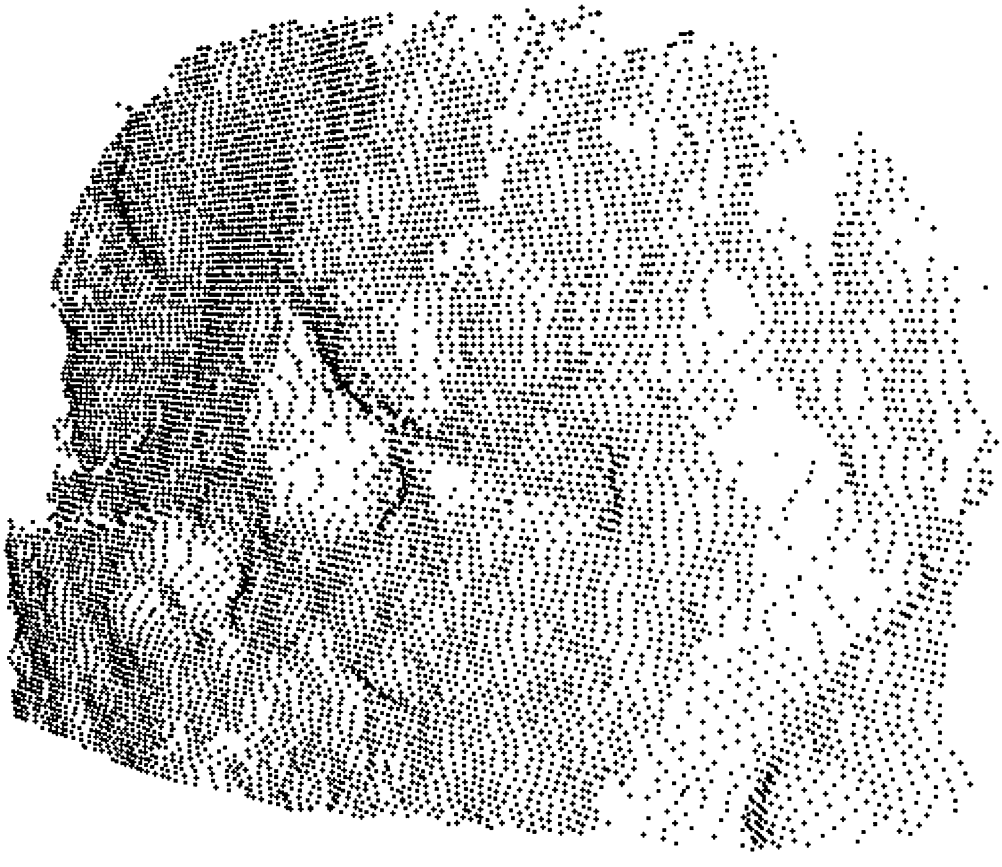

Fig. 10. A stone of the Roman Amphitheater of Aosta, Italy: 3D points extracted through area matching.

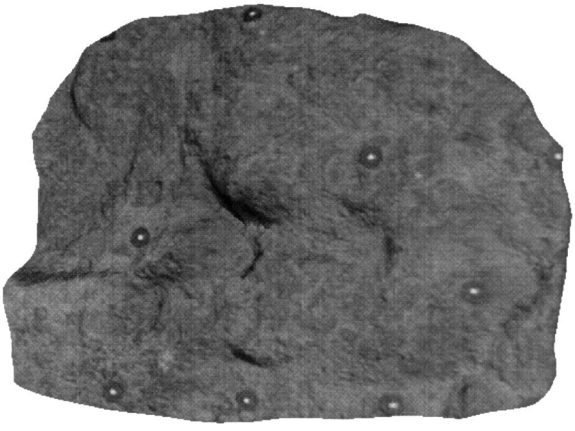

Fig. 11. A stone of the Roman Amphitheater of Aosta, Italy: 3D reconstruction after texture mapping.

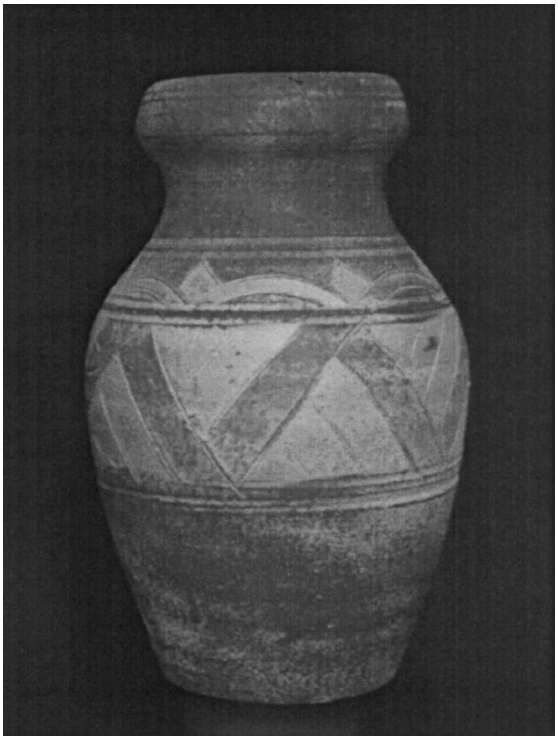

Fig. 12. Original view of the *terra-cotta* vase used for a 3D reconstruction experiment.

measurements and the photogrammetric data is to be attributed to the fact that the later consists of a sparser set of points, which required additional interpolation. Moreover, the precision of the available photogrammetric data not far from 1 mm either.

A third experiment was performed on the *terra-cotta* vase of Fig. 12, placed at about 1 m from the camera system, whose baseline was about 0.5 m. An initial rough estimate of the object surface was previously made

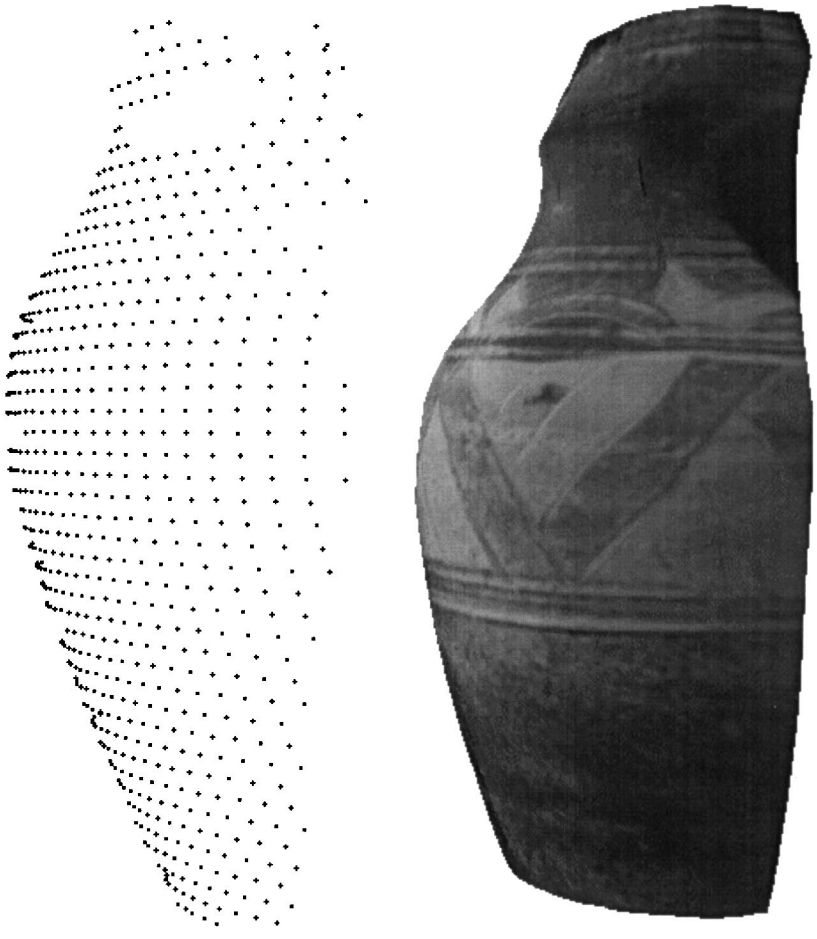

Fig. 13. 3D points extracted through area matching and 3D reconstruction with texture mapping of the *terra-cotta* vase.

through edge-matching [5,21] and boundary reconstruction at object rims [24,31]. The reconstruction results are shown in Fig. 13.

Finally, an experiment of 3D reconstruction of a speaker behind a desk in a tele-conferencing environment was performed. The acquisition was made with a trinocular camera system at CCETT, France, within the ACTS "PANORAMA" Project (see Fig. 14). No initial reconstruction was used for area matching. Instead, progressive-scan initialization was performed. The results of the area matching procedure are visible in Figs. 15 and 16. As we can from Fig. 15 the quality of the reconstruction greatly benefits from the fact that the geometric distortion is included in the model.

## 7. Conclusions

In this article we proposed and illustrated a general and robust approach to the problem of close-range 3D reconstruction of objects from stereo-correspondence of luminance profiles. The method is independent on

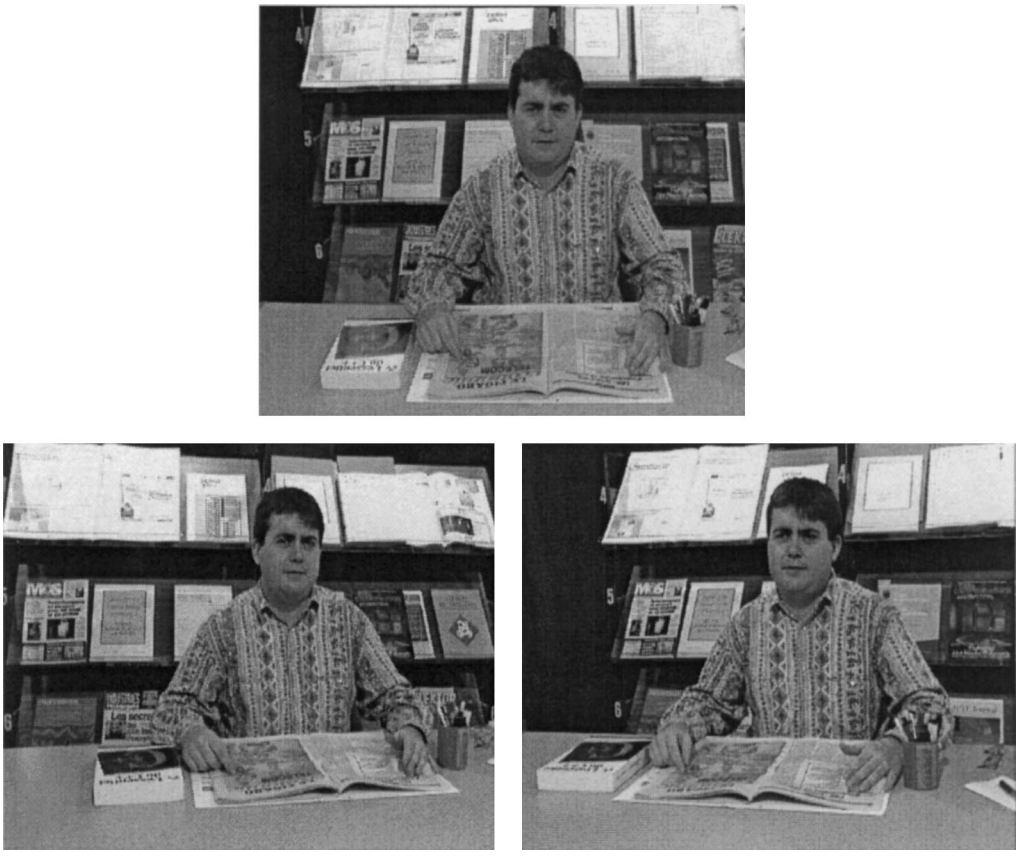

Fig. 14. Original views (top, left and right) of a speaker in a teleconferencing environment.

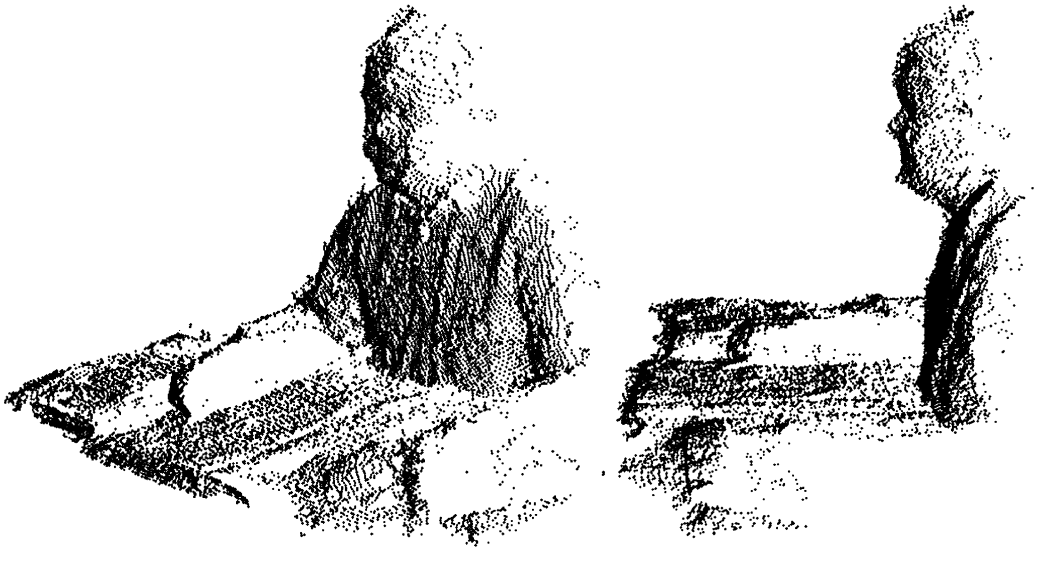

Fig. 15. Speaker in a teleconferencing environment: 3D points extracted through area matching.

Fig. 16. Speaker in a teleconferencing environment: 3D reconstruction after texture mapping. The reconstruction errors in correspondence to the arms are due to lack of information caused by object occlusion.

the geometry of the acquisition system which could be a set of $n$ cameras with strongly converging optical axes.

The robustness of the approach can be mainly attributed to the *physicality* of the matching process, which is virtually performed in the 3D space. In fact, both 3D location and local orientation of the surface patches are estimated, so that the geometric distortion can be accounted for. The method takes into account the viewer-dependent radiometric distortion as well.

The method has been implemented by using a calibrated set of three standard TV-resolution CCD cameras system, and thoroughly tested on a variety of real scenes with satisfactory results.

The experiments conducted shown that in some conditions it is possible to adopt a multi-resolution approach to area matching, and some of the best reconstruction results have been obtained that way. We are currently working on a robust multi-resolution version of the method proposed in this article. We are also working on the integration of the proposed area matching method with egomotion techniques for full-3D reconstruction of objects.

## Notations

| | |
|---|---|
| $\mathscr{GL}_n$ | (general linear group): set of all non-singular $n \times n$ matrices |
| $\boldsymbol{H} \in \mathscr{GL}_n$ | homography matrix ($\boldsymbol{H} \in \mathscr{R}^{n \times n}$ such that $\det \boldsymbol{H} \neq 0$) |
| $\boldsymbol{I}_n \in \boldsymbol{R}^n$ | identity matrix |
| $I^{(i)}(\boldsymbol{u}^{(i)})$ | luminance profile on the $i$th view, corresponding to the projective image coordinates $\boldsymbol{u}^{(i)}$ |
| $I_j^{(i)}(\boldsymbol{u}^{(i)})$ | luminance profile on the $i$th view at $\boldsymbol{u}^{(i)}$, as transferred from the $j$th view |
| $\boldsymbol{M} \in \mathscr{R}^{n \times n}$ | $n \times n$ matrix |
| $\boldsymbol{M} \in \mathscr{SE}_{n+1}$ | rigid motion matrix, i.e. $\boldsymbol{M} = \begin{bmatrix} & \boldsymbol{R} & & \boldsymbol{T} \\ 0 & 0 & 0 & 1 \end{bmatrix}$, with $\boldsymbol{R} \in \mathscr{SO}_n$ and $\boldsymbol{T} \in \mathscr{R}^n$ |
| $\mathscr{P}^n$ | projective space of dimension $n$ |

| | |
|---|---|
| $\mathscr{P}^2$ | projective plane (image space) |
| $\mathscr{P}^3$ | projective space (object space) |
| $\boldsymbol{P} \in \mathscr{R}^{n \times (n+1)}$ | projection matrix from $\mathscr{P}^{n+1}$ to $\mathscr{P}^n$ (rank-$n$ matrix) |
| $\mathscr{R}^n$ | Euclidean space of dimension $n$ |
| $\mathscr{R}^{n \times n}$ | Euclidean space of dimension $n$ |
| $\boldsymbol{R} \in \mathscr{SO}_n$ | rotation matrix ($\boldsymbol{R} \in \mathscr{R}^{n \times n}$ such that $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \boldsymbol{I}_n$, $\det \boldsymbol{R} = 1$) |
| $\mathscr{SO}_n$ | (special orthogonal group): set of all $n \times n$ rotation matrices |
| $\mathscr{SE}_{n+1}$ | (special Euclidean group) set of all rigid motions in $\mathscr{P}^n$ |
| $\mathscr{S}$ | surface patch in object space |
| $\mathscr{S}^{(i)}$ | $i$th view of the surface patch $\mathscr{S}$ in object space |
| $\boldsymbol{T} \times \boldsymbol{S} = (\boldsymbol{T}\times)\boldsymbol{S}$ | vector product in matrix form, $\boldsymbol{T} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$, $\boldsymbol{T}\times = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}$ |
| $\boldsymbol{u} \in \mathscr{P}^2$ | point in image space |
| $\boldsymbol{u}^{(i)} \in \mathscr{P}^2$ | projective coordinates of an image point of the $i$th camera |
| $\boldsymbol{v}^{(i)} \in \mathscr{P}^2$ | normalized projective coordinates of an image point of the $i$th camera (the first two coordinates are image coordinates) |
| $x \in \mathscr{R}^n$ | vector with $n$ elements |
| $x \in \mathscr{P}^n$ | point in projective space (vector with $n+1$ elements) |
| $x \in \mathscr{P}^3$ | point in object space |
| $\boldsymbol{0}_n \in \boldsymbol{R}^n$ | vector with zero components |
| $\pi$ | image or retinal plane |

In general, lowercase bold letters are used for vectors and uppercase bold letters are used for matrices. When it is not obvious to understand which reference frame a vector is referred to, a superscript between parentheses is used to specify it. Matrix transposition is denote by a superscript T. The superscript $-\mathrm{T}$ is used for combined inversion and transposition.

## References


[1] N. Ayache, Artificial Vision for Mobile Robots, Stereo Vision and Multisensorial Perception, MIT Press, Cambridge, MA, 1991.
[2] M. Born, E. Wolf, Principles of Optics, Pergamon Press, Oxford, 1959.
[3] A. Blake, A. Zisserman, Visual Reconstruction, MIT Press, Cambridge, MA, 1987.
[4] Bui-Tuong, Phong, Illumination for computer-generated pictures, Commun. ACM (June 1975) 311–317.
[5] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6) (1986) 679–698.
[6] L. Falkenhagen, Depth estimation from stereoscopic image pairs assuming piecewise continuous surfaces, in: Y. Paker, S. Wilbur (Eds.), Image Processing for Broadcast and Video Production, Springer series on Workshops in Computing, Springer, Great Britain, 1994, pp. 115–127.
[7] O. Faugeras, L. Robert, What can two images tell us about a third one? Internat. J. Comput. Vision 18 (5–19) (1996) 5–19.
[8] A.A. Grattarola, Volumetric reconstruction from object silhouettes: A regularization procedure, Signal Processing 27 (1992) 27–35.
[9] W. Hodge, D. Pedoe, Methods of Algebraic Geomerty, Cambridge University Press, Cambridge, 1952.
[10] B.K.P. Horn, Robot Vision, MIT Press, Cambridge, MA, 1986.
[11] R. Koch, Model-based 3D scene analysis from stereoscopic image sequences, in: ISPRS'92, Vol. 29, Part B5, Washington, October 1992, pp. 427–437.
[12] L. Levi, Applied Optics – A Guide to Optical System Design, Vol. I, Wiley, New York, 1968.
[13] Q.-T. Luong, T. Vieville, Canonical representations for the geometries of multiple projective views, Comput. Vision and Image Understanding 64 (2) (September 1996) 193–229.
[14] J.L. Mallet, Discrete smooth interpolation, ACM Trans. Graphics 8 (2) (1989) 121–144.
[15] J.L. Mundy, A. Zisserman, Projective geometry for machine vision, in: J.L. Mundy, A. Zisserman (Eds.), Geometric Invariance in Computer Vision, MIT Press, Cambridge, MA, 1992.

[16] K. Nayar, K. Ikeuchi, T. Kanade, Surface reflection: Physical and geometrical perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 13 (7) (July 1991) 611–633.
[17] J.A. Nelder, R. Mead, Comput. J. 7 (1964) 308.
[18] F. Pedersini, Analisi di scena per applicazioni di ricostruzione tridimensionale, Ph.D. thesis, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy, 1995.
[19] F. Pedersini, D. Pele, A. Sarti, S. Tubaro, Calibration and self-calibration of multi-ocular camera systems, in: Internat. Workshop on Synthetic-Natural Hybrid Coding and 3D Imaging (IWSNHC3DI'97), Rhodes, Greece, 5–9 September 1997.
[20] F. Pedersini, A. Sarti, S. Tubaro, Synthesis of virtual views using non-lambertian reflectivity models and stereo matching, in: IEEE Internat. Conf. on Image Processing, Washington DC, USA, October 1995.
[21] F. Pedersini, A. Sarti, S. Tubaro, Combined motion and edge analysis for a layer-based representation of image sequences, in: IEEE Internat. Conf. on Image Processing, Lausanne, Switzerland, 1996.
[22] F. Pedersini, A. Sarti, S. Tubaro, A multi-view trinocular system for automatic 3D object modeling and rendering, in: XVIII Internat. Congress for Photogrammetry and Remote Sensing, Vienna, Austria, 1996.
[23] F. Pedersini, A. Sarti, S. Tubaro, 3D motion estimation of a trinocular system for a full-3D object reconstruction, in: IEEE Internat. Conf. on Image Processing, Lausanne, Switzerland, September 1996.
[24] F. Pedersini, A. Sarti, S. Tubaro, 3D surface reconstruction from horizons, in: Internat. Workshop on Synthetic-Natural Hybrid Coding and Three-Dimensional (3D) Imaging (IWSNHC3DI'97), Rhodes, Greece, 5–9 September 1997.
[25] F. Pedersini, A. Sarti, S. Tubaro, Estimation and compensation of subpixel edge compensation error, IEEE Trans. Pattern Anal. Mach. Intell. 19 (10) (October 1997).
[26] F. Pedersini, S. Tubaro, F. Rocca, Camera calibration and error analysis, an application to binocular and trinocular stereoscopic systems, in: 4th Internat. Workshop on Time-Varying Image Processing and Moving Object Recognition, Florence, 1993.
[27] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.
[28] K.E. Torrance, E.M. Sparrow, Theory for off-specular reflection from roughened surface for ray reflection, J. Opt. Soc. Amer. 65 (1975) 531–536.
[29] R. Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, IEEE J. Robotics and Automation 3 (4) (August 1987) 323–344.
[30] S. Tubaro, A precise stereoscopic system with two video cameras, Eur. Trans. Telecommun. 3 (3) (May–June 1992) 275–280.
[31] R. Vaillant, O.D. Faugeras, Using extremal boundaries for 3D object modeling, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (February 1986) 157–173.