ELSEVIER

Signal Processing 82 (2002) 1215–1232

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Image-based surface modeling: a multi-resolution approach

Augusto Sarti *, Stefano Tubaro

Dipartimento di Elettronica e Informazione (DEI), Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy

Received 22 March 2001; received in revised form 8 April 2002

Abstract

In this article we propose a general and robust technique for modeling surfaces through the analysis of multiple image acquisitions. Our method is based on the minimization of the multi-view texture mismatch and is inherently multi-resolution, as the surface is obtained through a progressive refinement of hierarchical radial basis functions.

Keywords: Surface modeling; Stereo; Radial basis functions

1. Introduction

Classical stereometric methods for 3D data extraction from multiple views are based on the detection, matching and geometric triangulation of viewer-invariant features such as corner points or "sharp" edges. Such methods, unfortunately, are unable to generate dense clouds of 3D data, therefore it is usually quite difficult to interpolate them into a global surface that resembles the shape of the imaged object. In order to generate denser depth maps, a widely adopted solution is stereopsis, which consists of determining the correspondences between the luminance profiles of small image areas on the available views [9,17,21]. The 3D coordinates of the surface patch that originated a homologous pair of luminance profiles can, in fact, be computed through geometric triangulation, while the area matching in different views is usually based on the optimization of some similarity function (or the minimization of the texture mismatch) between luminance profiles.

The literature is rich with solutions based on this area matching approach, which basically differ in the surface representation, in the matching strategy and in the adopted regularization constraints. For example, the surface model can be in implicit or explicit form: the former is based on the evolution of a volumetric function's levelset [6,11] (an implicit function in 3D space) driven by the texture mismatch between views; the latter consists of patchworking [22] several depth maps (explicit functions) obtained through stereopsys [9,17,21].

Depth map estimation is usually based on the analysis of a limited number of views (typically two or three) and it requires that all surface points are simultaneously visible from at least two of them. Depending on the geometry of the acquisition system, the surface parametrization can be that of an elevation map (distance from a reference plane) or a perspective depth map (distance from the optical center of the reference camera). Also, besides the usual multi-view constraints such as the epipolar one,

* Corresponding author.
E-mail addresses: augusto.sarti@polimi.it (A. Sarti), tubaro@elet.polimi.it (S. Tubaro).

0165-1684/02/$ - see front matter
PII: S0165-1684(02)00244-X

it usually requires some additional constraints to improve the estimation's robustness, such as ordering constraints [17] or smoothness constraints (locally constant depth [9], or locally planar surface [21]). It is also possible to account for lens distortion and non-lambertian surface reflectivity [21].

All such solutions, however, need an initial approximation of the object surface in order to prevent the algorithm from encountering relative minima, which is, in fact, a major limitation. The 3D modeling approach that we propose in this article, on the contrary, is able to estimate the surface geometry effectively and efficiently without producing outliers and without needing any initial model. The method, in fact, begins with modeling a very simple (low-resolution) surface and applies plastic deformations to it by changing a limited number of parameters until the projection of textures on it agrees at best. This "modeling bootstrap" relies on a multi-resolution surface shaping strategy. Every time the resolution increases, previously missing details are added in such a way to improve the similarity between projected textures. In order to do so, the surface is modeled as a hierarchical radial basis function (RBF) network [3] made of 2D gaussian functions positioned on regular hexagonal grids of progressively increasing density.

The article is organized as follows: after a brief introduction to some basic facts of multiple-view geometry, included in Section 2, we will discuss the parametric surface representation adopted in this article in Section 3. This will enable us to introduce, in Section 4, our approach to image-based 3D modeling. Section 5 will illustrate some results obtained from real image acquisitions. Finally, in Section 6 we will draw some conclusions on the proposed strategy, discuss future perspectives of this solution and propose possible improvements.

2. Some general concepts surface and multiple-view geometry

The problem of estimating the 3D geometry of a scene from two or more of its views has received a considerable attention in the computer vision literature (see, for example, [10]). The basic idea is to start from two or more views of the scene and use the image coordinates of corresponding image features to determine the 3D position (see Fig. 1) of the scene features that generated them. The mathematical tools that are required to successfully perform this task come from projective geometry (the camera model is a perspective projection); differential topology (surfaces are modeled as manifolds); and multi-view geometry (model shaping is based on stereometry).

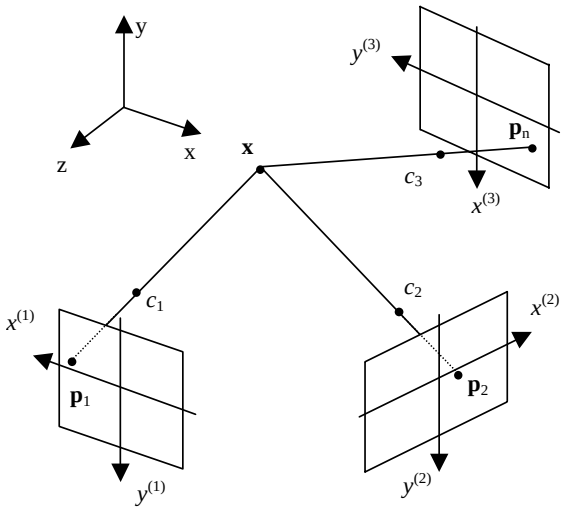


Fig. 1. Multicamera stereo geometry: $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ are image points that correspond to a same point $\mathbf{x}$ in the 3D space. The optical rays associated to corresponding image points intersect in $\mathbf{x}$. c_1, c_2 and c_3 are the optical centers of the cameras.

In this section we will briefly discuss some of the basic concepts that will prove useful in the following sections. These concepts are here collected mostly for the sake of setting a homogeneous notation and vocabulary.

2.1. The camera model

We assume that the cameras perform a perspective projection of the 3D world on the image plane (pinhole camera). Considering that the image coordinates are expressed in pixel units (see Fig. 2), the projective coordinates $\mathbf{u} = [u_1, u_2, u_3]^{\mathrm{T}}$ of the image point (the image coordinates are $x = u_1/u_3$, $y = u_2/u_3$) are obtained as $\mathbf{u} = \mathbf{P}\mathbf{x}$ where $\mathbf{P}$ is a 3×4 projection matrix and $\mathbf{x} = [x_1, x_2, x_3, 1]^{\mathrm{T}}$ are to the projective coordinates of the corresponding 3D point. When the reference frame corresponds to that of the projective

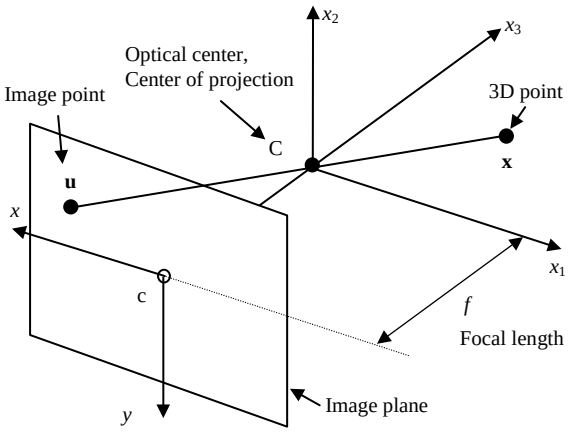


Fig. 2. A scheme for the image formation process.

camera (see Fig. 2) the projection matrix takes on a very simple form $\mathbf{P} = \mathbf{K}\mathbf{P}_0$, where

$$\mathbf{K} = \begin{bmatrix} f/d_x & 0 & x_{\text{off}} \\ 0 & f/d_y & y_{\text{off}} \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{P}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (1)$$

f is the focal length; d_x and d_y are the horizontal vertical size of the camera pixel; and x_{off} and y_{off} are the horizontal and vertical offsets between the image center and the camera's principal axis.

If we apply a rigid motion (rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$) to the camera, the projection matrix becomes $\mathbf{P} = \mathbf{K}\mathbf{P}_0\mathbf{G}$, where

$$\mathbf{G} = \begin{bmatrix} \mathbf{R}^{\mathbf{T}} & -\mathbf{R}^{\mathbf{T}}\mathbf{t} \\ 0 \;\; 0 \;\; 0 & 1 \end{bmatrix}. \quad (2)$$

The perspective projection operated by a real lens, however, is less ideal than the one described above. In fact, lens distortion (a non-linear stretching of the image plane) needs to be accounted for. In most cases, lens distortion is practically radial [14], and can be modeled by a power series of the form

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

truncated to the third or the fifth order [25], where r_{u} and r_{d} are the distances from the principal point of the distorted and undistorted image points.

In what follows we will assume that the relative positions, orientations and the internal parameters (e.g. focal length, lens distortion coefficients) of the camera(s) during the acquisition session are known (calibrated acquisition system [19,25]), and that lens distortion has been compensated for [21] through image warping or through a coordinate distortion function [21].

2.2. Local charts on surfaces

Modeling surfaces means shaping 2D manifolds, which are differential topological entities that locally look like a vector space. Basically, a manifold can be seen as an atlas made of a collection of "local" maps that can be "gently" flattened. More precisely, a subset M of R^3 is a smooth manifold of dimension 2 if for each point $x \in M$ there is a neighborhood $W \subset M$ that can be smoothly and invertibly mapped onto a subset $U \subset R^2$ in a one-to-one fashion. The mapping ψ from $W \subset M$ to $U \subset R^2$ is thus called a "system of coordinates" (or "coordinate map") on W, and its inverse ψ^{-1} is called a "parametrization" [26].

One very simple example of surface parametrization is the elevation map with respect to a reference plane, usually perpendicular to the viewing direction. Another typical choice is the depth map, where the depth is measured as the distance from surface and the optical center of a "reference" camera (perspective depth map). In this second case, the parametrization is simply the mapping from the image plane to the object surface, i.e. the prolongation of the optical ray from the image point to the object surface. In general, we could choose any parametrization that best fits the viewing conditions and the object's topology. An elevation map is suitable for a limited class of $2\frac{1}{2}$D surfaces viewed from a distance $d \gg f$, while a perspective depth map is more suitable for $2\frac{1}{2}$D surfaces acquired from a closer viewpoint. This last parametrization, in fact, is the one that guarantees the maximum consistency between images and surface topology as depth is measured along optical rays, and images are acquired in such a way to capture the surface geometry at best (see Fig. 3).

2.3. Transferring and comparing luminance profiles

As already said in Section 2.2, a perspective depth map guarantees the maximum consistency between images and surface topology. In fact, with reference to

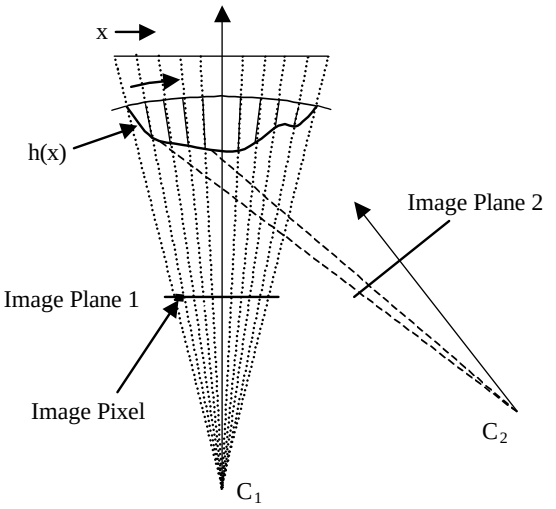


Fig. 3. A 2D sketch of the perspective depth map and of the point correspondence between two views.

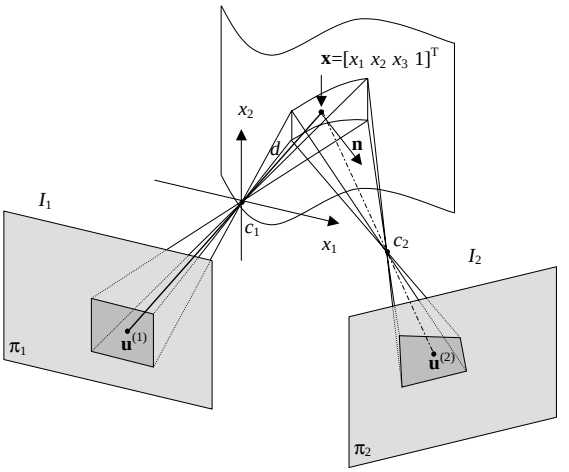


Fig. 4. Luminance transfer: the texture on image 1 is back-projected onto the object surface and re-projected onto the image plane 2.

Fig. 3, we can see that it guarantees a one-to-one correspondence between image points of the reference view and surface points. In fact, adopting a multi-camera system for surface reconstruction, it seems logical to choose the most central view as a reference for this parametrization.

In order to characterize the mechanism that allows us to estimate the correct surface depths using the available views, we need to define some measure of the mismatch between luminance profiles acquired from different viewpoints. This requires a detailed characterization of the luminance transfer between image planes (See Fig. 4).

Let us consider a point $\mathbf{x} = [x_1, x_2, x_3, 1]^{\mathrm{T}}$ on the surface of the object to be reconstructed and let $\mathbf{u}^{(1)} = [u_1^{(1)}, u_2^{(1)}, u_3^{(1)}]^{\mathrm{T}} = \mathbf{P}_1\mathbf{x}$ and $\mathbf{u}^{(2)} = [u_1^{(2)}, u_2^{(2)}, u_3^{(2)}]^{\mathrm{T}} = \mathbf{P}_2\mathbf{x}$ be the projective coordinates of this point, as viewed from two cameras (see Fig. 1). The corresponding image coordinates will thus be

$$(x^{(1)}, y^{(1)}) = \left(\frac{u_1^{(1)}}{u_3^{(1)}}, \frac{u_2^{(1)}}{u_3^{(1)}}\right)$$

and

$$(x^{(2)}, y^{(2)}) = \left(\frac{u_1^{(2)}}{u_3^{(2)}}, \frac{u_2^{(2)}}{u_3^{(2)}}\right).$$

Using the first camera as a reference, the surface can be parametrized by a smooth perspective depth map of the form $d(x^{(1)}, y^{(1)})$, where d measures the distance between surface point and c_1.

The 3D coordinates of the point $\mathbf{x} = [x_1, x_2, x_3, 1]^{\mathrm{T}}$ can be quite easily written as a function of the image coordinates (x, y) and the depth d as

$$x_1 = \frac{xd}{\sqrt{x^2 + y^2 + f^2}}, \quad x_2 = \frac{yd}{\sqrt{x^2 + y^2 + f^2}},$$

$$x_3 = \frac{fd}{\sqrt{x^2 + y^2 + f^2}}.$$

When the focal length f is much greater than the image coordinates x and y, the above expressions become

$$\mathbf{x} = \left[\frac{xd}{f}, \frac{yd}{f}, d, 1\right]^{\mathrm{T}}.$$

It is now possible to find a closed-form expression of the image coordinates $(x^{(2)}, y^{(2)})$ on π_2 as a function of the image coordinates $(x^{(1)}, y^{(1)})$ on π_1 and the depth d. In fact, assuming that $\mathbf{u}^{(1)} = \mathbf{K}_1\mathbf{P}_0\mathbf{x}$ and that $\mathbf{u}^{(2)} = \mathbf{K}_2\mathbf{P}_0\mathbf{G}\mathbf{x}$ (where the reference frame of point $\mathbf{x}$ is attached to camera 1 and $\mathbf{G}$ describes the rigid motion from camera 1 to camera 2), we can write

$$\mathbf{u}^{(2)} = \mathbf{K}_2\mathbf{P}_0\mathbf{G}\begin{bmatrix}\mathbf{X}\\ 1\end{bmatrix} = \mathbf{K}_2[\mathbf{R} \quad -\mathbf{R}^{\mathrm{T}}\mathbf{t}]\begin{bmatrix}\mathbf{X}\\ 1\end{bmatrix}$$

$$= \mathbf{K}_2\mathbf{R}\mathbf{X} - \mathbf{K}_2\mathbf{R}^{\mathrm{T}}\mathbf{t},$$

where $\mathbf{X} = [x_1\ x_2\ x_3]^{\mathrm{T}}$, whose components are written above as a function of $x^{(1)}$, $y^{(1)}$ and d. The resulting expression after normalization with respect to the third component, provides the (nonlinear) relationship between stereo-corresponding coordinates.

Let I_1 and I_2 be the luminance profiles of the two images. Assuming that the surface reflectivity is perfectly Lambertian [19] we can write

$$I_1(x^{(1)}, y^{(1)}) = I_2(x^{(2)}, y^{(2)}). \tag{3}$$

where the image coordinates $(x^{(2)}, y^{(2)})$ on π_2 can be written as a function of the corresponding image coordinates $(x^{(1)}, y^{(1)})$ on π_1 and the depth d as shown above. This luminance constancy constraint can be used to compute the depth d. However, relying on this constraint is very risky, as several pixels on the epipolar line could have the same luminance. In order to reduce this risk, we need to make a shape regularization assumption in the neighborhood of $(x^{(1)}, y^{(1)})$. The simplest way to proceed would be, for example, to assume that the depth is constant in a neighborhood $\wp$ of $(x^{(1)}, y^{(1)})$. This depth constancy constraint (zero-order smoothness assumption) allows us to estimate the depth d through the minimization of a cost function [11] of the form

$$C(d) = \int\int_{\wp} [I_1(x, y) - I_2(x^{(2)}(x, y, d), y^{(2)}(x, y, d))]^2 \,\mathrm{d}x\,\mathrm{d}y. \tag{4}$$

This can be done to estimate the depth of all surface points that are visible on both images.

A better regularization constraint is represented by the tangent plane constancy constraint (first-order smoothness), which consists of assuming that the surface is planar in a neighborhood $\wp$ of $(x^{(1)}, y^{(1)})$. Let us assume that the surface is locally described by a plane of equation $\mathbf{s}^{\mathrm{T}}\mathbf{x} = 0$, where $\mathbf{s} = [s_1\ s_2\ s_3\ 1]^{\mathrm{T}}$, and let S be the back-projection of $\wp$ onto the surface. The projective image coordinates $\mathbf{u}^{(2)} = \mathbf{P}_2\mathbf{x}$ of a point $\mathbf{x} \in S$ can be written as a function of $\mathbf{u}^{(1)} = \mathbf{P}_1\mathbf{x}$, through a collineation (a linear projective mapping)

$$\mathbf{u}^{(2)} = \mathbf{K}(\mathbf{s})\mathbf{u}^{(1)}, \quad \mathbf{u}^{(1)} = [x^{(1)} \quad y^{(1)} \quad 1]^{\mathrm{T}}, \quad (x^{(1)}, y^{(1)}) \in \wp. \tag{5}$$

The collineation $\mathbf{K}(\mathbf{s})$ (a 3×3 invertible matrix) can be written in closed form as a function of the (three) parameters of the plane (and of the camera parameters), as shown in [21]. This allows us to estimate the tangent plane $\mathbf{s}$ through the minimization of a cost function of the form

$$C(\mathbf{s}) = \int\int_{\wp} [I_1(\mathbf{u}) - I_2(\mathbf{K}(\mathbf{s})\mathbf{u})]^2 \,\mathrm{d}\mathbf{u}. \tag{6}$$

Indeed, this second way of proceeding requires the estimation of three parameters instead of one, but provides more information (a differential surface description of order 1). This, by the way, could be exploited to reduce the density of the depths to be computed.

More generally, we could choose a more complex parametric model for the surface shape whose region of support $\wp$ could be limited to a neighborhood of modest size (local approach), or it could be arbitrarily large, up to the size of the entire viewed surface (global approach).

If we express the depth as a function of a limited set of surface parameters $\mathbf{p} = (\alpha_1, \ldots, \alpha_N)$ (see Section 3 for a more detailed description), we can define a cost function of the form

$$C(\mathbf{p}) = \int\int_{\wp} [I_1(x, y) - I_2(x^{(2)}(x, y, \mathbf{p}), y^{(2)}(x, y, \mathbf{p}))]^2 \,\mathrm{d}x\,\mathrm{d}y. \tag{7}$$

The surface parameters that minimize the luminance mismatch can thus be estimated as

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}}[C(\mathbf{p})]. \tag{8}$$

In general, the CCD sensors of the cameras used for image acquisition do not exactly have the same physical and electrical characteristics, therefore the luminance constancy constraint cannot be taken for granted. In order to overcome this difficulty, we can modify this constraint as

$$I_1(x^{(1)}, y^{(1)}) = a_{12} + g_{12} \cdot I_2(x^{(2)}(x^{(1)}, y^{(1)}, \mathbf{p}), y^{(2)}(x^{(1)}, y^{(1)}, \mathbf{p})),$$

where a_{12} and g_{12} represent a differential offset and a differential gain between the two cameras [1]. In order to keep this problem into account the cost function (7) can be rewritten as

$$C(\mathbf{p}) = \int\int_{\wp} [(I_1(x, y) - \overline{I_1}) - g_{12}(I_2(x^{(2)}(x, y, \mathbf{p}), y^{(2)}(x, y, \mathbf{p})) - \overline{I_2})]^2 \,\mathrm{d}x\,\mathrm{d}y, \tag{9}$$

where $\overline{I_1}$ and $\overline{I_2}$ are the mean values of the luminance profiles, while g_{12} can be estimated by comparing the views of some reference object.

This last model-based approach to parametric surface estimation is the focus of this article, therefore in the next section we will develop it further.

3. Modeling surfaces with RBF networks

In the previous section we described in a general fashion how to approach the surface modeling problem through a global minimization process. The question that we would like to answer now is how to define a suitable parametric surface model for this purpose, and how to perform an image-based estimation of its parameters through a procedure of practical usability.

One of the characteristics that we would like our parametric surface model to have is a simple control of its smoothness. In fact this characteristic would enable a control of the surface resolution, and favor a certain robustness in the parameter estimation.

Although the literature is rich with parametric surface models with controlled smoothness (see for example [2,13,23,24]), our interest in radial basis functions (RBFs) [2] arises from the fact that they can also be organized hierarchically (HRBF), which makes them good candidates for multi-resolution methods. Among the RBFs, we are interested in radially symmetric gaussian functions, as they are circularly symmetric in spite of their separability. Let $\mathbf{s} = [x\ y]^{\mathrm{T}}$ be the image coordinates of a point on the reference image plane π_1. The surface model can be expressed in parametric form by expressing the depth map as a weighed linear combination of gaussians of the form

$$\begin{aligned} d(\mathbf{s}) &= \sum_{m=1}^{M} w_m G(\mathbf{s}; \mathbf{s}_m, \sigma_m); \\ G(\mathbf{s}; \mathbf{s}_m, \sigma_m) &= \frac{1}{\pi\sigma_m^2} \mathrm{e}^{-|\mathbf{s}-\mathbf{s}_m|^2/\sigma_m^2}, \end{aligned} \tag{10}$$

where M is the number of gaussians; the image points $\mathbf{s}_m,\ m = 1,\ldots,M$, provide the depth rays along which the centers of the gaussians are positioned; while w_m and σ_m represent their weights (magnitudes) and spreads (standard deviations).

If the gaussians are scattered on a regular and separable (rectangular) grid on the image plane and the maximum viewing angle β on the CCD is not very wide (so that $\tan\ \beta \cong \beta$), then all the gaussians can be assumed to have the same spread ($\sigma_m = \sigma$). In what follows, we will always assume that this is true, although removing this hypothesis would not generate unbearable complications, as it would require the adoption of a mildly space-varying spread. In fact, a regular grid on the image plane would project onto the viewing sphere as distorted (non-uniform) grid. Anyway, we can assume this model to be completely specified by the set of weights $w_m,\ m = 1,\ldots,M$. As far as the grid density is concerned, the choice will have to be made in such a way to capture all the details of interest [4,5].

3.1. Hierarchical Radial Basis Function (HRBF) networks

The nonlinear optimization problem set forth in Section 2.3 exhibits potential problems of convergence to a global minimum. This encourages us to search for a solution that allows us to tackle the parametric estimation problem in a progressive (multi-resolution) fashion. One such solutions would allow us to subdivide the optimization process in several steps; to "navigate" the parameter space in a more efficient fashion; and to adapt the total number of parameters to the level topological complexity of the imaged surfaces [5].

A multi-resolution representation of our parametric surface is provided by Hierarchical Radial Basis Function (HRBF) networks. In this case, the surface representation is organized in layers of different resolution, each built on a set of uniformly distributed gaussians with the same spread. The gaussians are usually positioned on a regular and separable (rectangular) grid. From one layer to the next one the grid density increases (normally it doubles in both directions) in order to capture finer details. This hierarchy of basis functions can be specified by rewriting Eq. (10) as

$$\begin{aligned} d(\mathbf{s}) &= \sum_{l=1}^{L} \sum_{k=1}^{K_l} w_{lk} G(\mathbf{s}; \mathbf{s}_{lk}, \sigma_l); \\ G(\mathbf{s}; \mathbf{s}_{lk}, \sigma_l) &= \frac{1}{\pi\sigma_l^2} \mathrm{e}^{-(\mathbf{s}-\mathbf{s}_{lk})^2/\sigma_l^2}, \end{aligned} \tag{11}$$

where L is the number of resolution layers; K_l is the number of Gaussians in the lth layer; and σ_l is the standard deviation of the gaussians of that layer. In order to obtain a good surface representation, the value of σ_l should be chosen according to the grid density. A good choice for the spread is proposed in [4] as

$$\sigma_l = 1.465 \cdot \Delta_l, \tag{12}$$

where Δ_l is the grid step (along x or y) for the layer l.

3.2. RBF networks as surface interpolators

Let us consider a smooth function $d(\mathbf{s})$, whose values are known only in a limited set of points $(\mathbf{s}_k,\ k = 1,\ldots,K)$, arbitrarily scattered on a certain domain. The value of $d(\mathbf{s})$ in an arbitrary point $\mathbf{s}$ of that domain can be estimated as a weighed sum of the known samples $\mathbf{s}_k$

$$d(\mathbf{s}) = \frac{\sum_{\mathbf{s}_k \in A(\bar{\mathbf{s}})} d(\mathbf{s}_k) w(|\mathbf{s} - \mathbf{s}_k|)}{\sum_{\mathbf{s}_k \in A(\bar{\mathbf{s}})} w(|\mathbf{s} - \mathbf{s}_k|)}.$$

The support $A(\mathbf{s})$ of the function's sampling is, in principle, the set of all the K points where d is known. In practice, however, its extension may be limited to the so-called *receptive field* of $\mathbf{s}$, i.e. to the set of the closest points to $\mathbf{s}$. Among the possible choices of weight functions $w(\cdot)$, gaussians are those that yield the maximum a posteriori (MAP) estimate of $\mathbf{s}$ from its sampling set $\mathbf{s}_k$ [11]

$$d(\mathbf{s}) = \frac{\sum_{\mathbf{s}_k \in A(\mathbf{s})} d(\mathbf{s}_k) \exp(-|\mathbf{s} - \mathbf{s}_k|^2/\sigma_e^2)}{\sum_{\mathbf{s}_k \in A(\mathbf{s})} \exp(-|\mathbf{s} - \mathbf{s}_k|^2/\sigma_e^2)},$$

where the spread σ_e is chosen according to the surface's roughness. We will see later that an RBF approach to surface interpolation becomes particularly useful when using a local approach to depth estimation. Such methods, in fact, produce a set of depths scattered on a regular grid, therefore the RBF interpolator enables a fast analytic computation of all the other points of the support region [11]:

$$\begin{aligned} &\hat{d}(\mathbf{s}) = \sum_{k=1}^{K} w_k G(\mathbf{s}; \mathbf{s}_k, \sigma_k); \\ &G(\mathbf{s}; \mathbf{s}_k, \sigma_k) = \frac{1}{\pi\sigma_k^2} \mathrm{e}^{-|\mathbf{s}-\mathbf{s}_k|^2/\sigma_k^2}; \\ &w_k = d(\mathbf{s}_k)\Delta^2, \end{aligned} \tag{13}$$

where Δ is the grid step of the RBF network.

In order to improve the quality of the results and speed up the interpolation process, once again we can adopt a multi-resolution approach based on a HBRF network:

$$\begin{aligned} &\hat{d}(\mathbf{s}) = \sum_{l=1}^{L} \hat{d}_l(\mathbf{s}) = \sum_{l=1}^{L} \sum_{k=1}^{K_l} w_{lk} G(\mathbf{s}; \mathbf{s}_{lk}, \sigma_l); \\ &G(\mathbf{s}; \mathbf{s}_{lk}, \sigma_l) = \frac{1}{\pi\sigma_m^2} \mathrm{e}^{-(\mathbf{s}-\mathbf{s}_{lk})^2/\sigma_l^2}; \\ &w_{lk} = r_{lk}(\mathbf{s}_k)\Delta_l^2; \\ &r_{lk}(\mathbf{s}_k) = \begin{cases} d(s_k) & \text{for } l = 1; \\ d(s_k) - \hat{d}_{l-1}(\mathbf{s}_k) & \text{for } l \neq 1. \end{cases} \end{aligned} \tag{14}$$

At each resolution level l interpolation (13) is carried out through an updating process based on the residuals $r_{lk} = d(\mathbf{s}_k) - \hat{d}_{l-1}(\mathbf{s}_k)$. In fact, the true interpolated value is obtained by adding, layer by layer, the missing details at the corresponding resolution.

4. Multi-resolution depth modeling

Although our approach to multi-resolution modeling has a wide range of validity in terms of camera-based acquisition systems, its development has been done with reference to a specific calibrated multi-camera system like the one of Fig. 5. The system is based on three digital cameras mounted on a

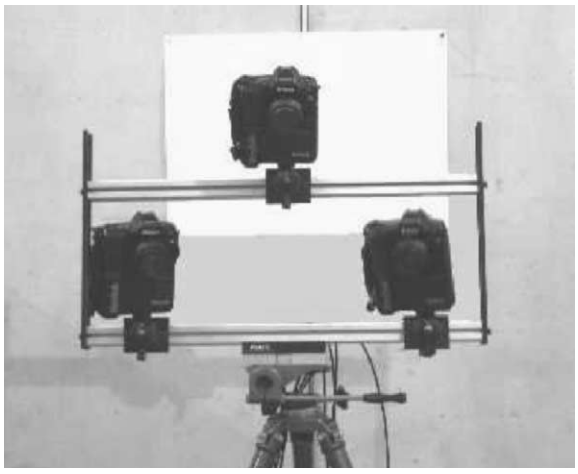


Fig. 5. A prototype trinocular acquisition system.

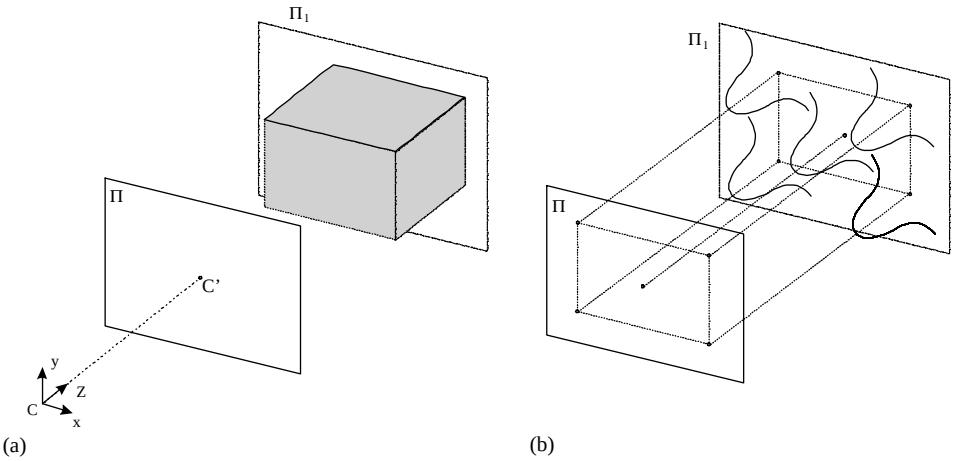


Fig. 6. Illustration of the surface parametrization mechanism in the case of elevation map. (a) The plane Π represents the reference image plane [10] and is shown in front of the optical center (C) only for convenience. The plane Π_1 is parallel to Π, all the elevations can be more conveniently referred to this plane. The gray volume identifies the space that contains the object of interest. (b) Arrangement of the gaussians on the first layer of the HRBF network.

rigid frame in a triangular configuration with convergent optical axis. The adopted resolution is *1524* by *1012* pixels with 8 or 10 bit per color component (R,G,B). Some experiments have been also conducted with another system, with similar geometry, but based on TV-resolution analog CCD cameras connected to a PC with frame grabber.

The choice of a trinocular system is motivated by the fact that three views give us extra redundancy in the identification of stereocorresponding elements. This improves the robustness of the 3D reconstruction process with respect to a simple binocular system [10,19,21], and sometimes they also allow us to overcome self-occlusion problems. Camera calibration is performed just before the acquisition campaign and it consists of acquiring and analyzing several trinocular views of a planar target in different random positions [19,25]. The upper (central) image is used as a reference view, and all the estimated 3D information will be refereed to it.

4.1. Model-based surface parametrization

As already said above, our approach to image-driven surface estimation is based on a hierarchical (multi-resolution) parametric surface model. This type of representation is provided by an HRBF network and the parameters that describe it are estimated through a comparison between the luminance profiles on the three available images.

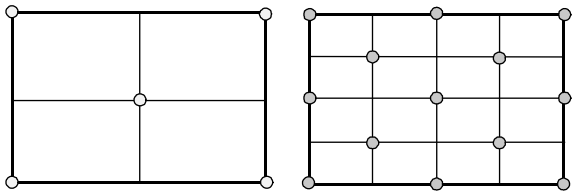


Fig. 7. Grids for positioning the gaussians in the first and the second layers of the HRBF network. The grid density quadruples between consecutive layers.

As already said in Section 3, the gaussians of typical HRBF networks are usually arranged on a separable (rectangular) grid whose axes are oriented like the reference image axis. The grid geometry that we adopted, however, is non-separable. In fact, the grids are hexagonal and their density quadruples from one level to the next (see Figs. 6 and 7).

The choice of a hexagonal grid is motivated by the fact that it provides a slightly better packing of the gaussian functions compared to an equivalent square grid [7]. Moreover, this distribution of gaussians turns out to be more suitable for surface representation at the lowest levels of resolution (the first few network

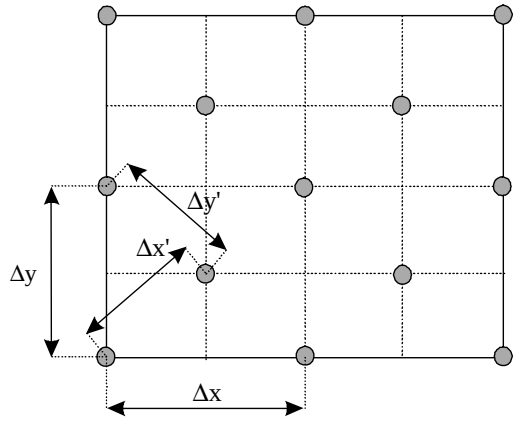


Fig. 8. An hexagonal grid can be seen as a rectangular grid rotated by a 45° angle.

layers). In fact, with this choice, some gaussians always end up in the middle of the region where the subject (and the most important details) usually are.

A regular hexagonal grid, however, can be seen as a square grid rotated of 45° (see Fig. 8), therefore the spread of (12) can still be used, as long as we define Δ_l as the distance between points along the diagonal direction (see Fig. 8).

As shown in Fig. 6, the region of support of the HRBF network is chosen such a way to select only the portion of the image in which the object under analysis is contained. Moreover it is necessary for the selected portion of the scene to be visible on all the three available images.

The total number of Gaussians N_{lt} and the number N_{ll} of those that lie on one side of the border of the surface's support region can be expressed in closed form as a function of the layer's index l as

$$\left.\begin{aligned} N_{ll} &= 2^{l-1}+1 \\ N_{lt} &= \frac{(2N_{ll}-1)^2+1}{2} \end{aligned}\right\} l \geqslant 1.$$

Our HRBF network is made of up to 8 layers, and the number of Gaussians quadruples from one to the next. This means that the linear resolution (along each axis) doubles from a layer to the next. We chose 8 as a maximum number of layers because in all our experiments (see Section 5) a grid of this density was always able to represent all the significant details of the imaged object. In order to give an idea of the numbers involved, Table 1 collects the values of N_{ll} and N_{lt} that correspond to the layer indices of interest, including index 0. This trivial layer is made of a single radial function with infinite spread. If our surface parametrization were based on an elevation map, this gaussian would correspond to a plane. Of course, if the region of interest was viewed under a small angle (tele-zoom lens), a perspective projection could be safely approximated with an affine camera (see Fig. 6, where $f \to \infty$), therefore even with a perspective depth map a gaussian with infinite spread would correspond to a plane. In general, however, the affine camera assumption is not acceptable, therefore using a perspective depth map makes a gaussian of infinite spread look like a portion of a spherical surface. In conclusion, our best choice for the surface parametrization is still the perspective depth map $d(x, y)$ discussed in Section 2.2 (see Fig. 9).

Table 1
Number of gaussians for each resolution level. Level 0 is made of a single radial function with infinite spread

Resolution level (l)	No. of gaussians on the border (N_{ll})	Total number of gaussians (N_{lt})
0	1	1
1	2	5
2	3	13
3	5	41
4	9	145
5	17	545
6	33	2113
7	65	8321
8	129	33025

4.2. Estimating surface parameters

The geometric configuration of the acquisition system that we used to test our modeling technique is illustrated in Fig. 10. The spherical surface A represents the layer 0 of the HRBF network used to model the surface. Its distance from the reference image ("Up" image/camera) is chosen in correspondence to the point of minimum distance from the principal axes, which is likely to fall close to the object center.

Layer by layer, the parameters of the HRBF can be estimated through the minimization of a specialized version of the cost functions (6)–(9). The cost

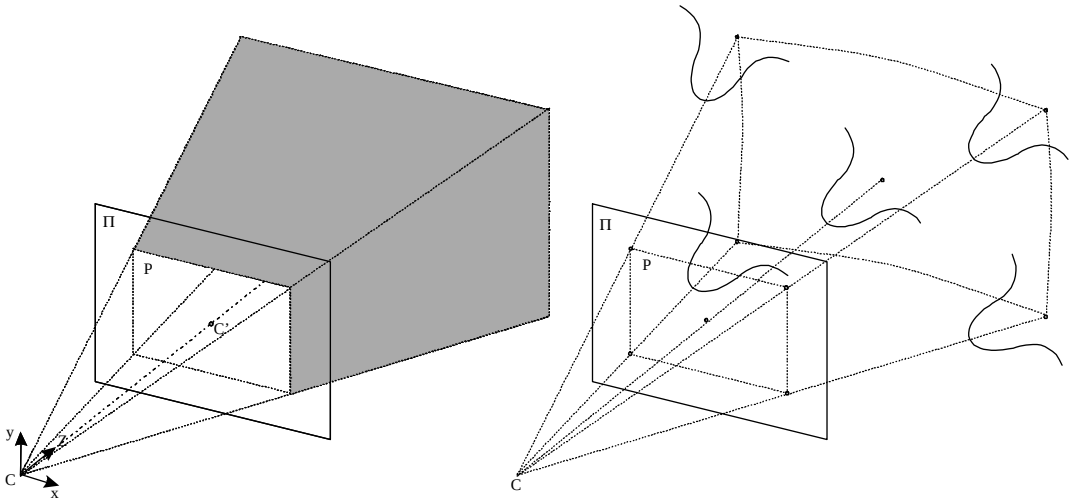


Fig. 9. Illustration of the surface parametrization mechanism in the case of perspective depth map. (a) The gray volume identifies the space that contains the object of interest. (b) Arrangement of the gaussians on the first layer of the HRBF network.

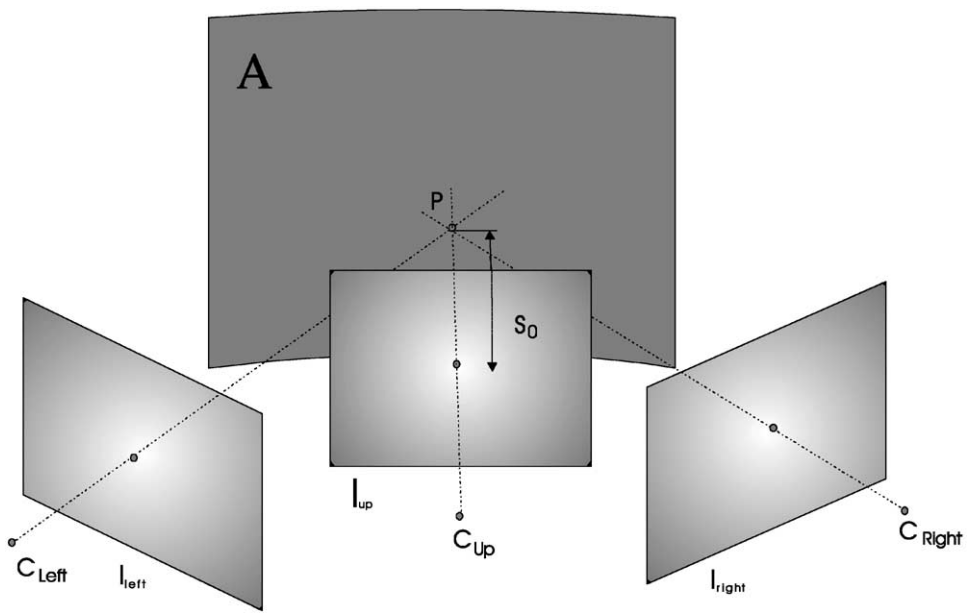


Fig. 10. Geometrical arrangement of the camera system. Surface A represents the layer 0 of the HRBF network used to model the object under analysis. Its distance from the reference image ("Up" image/camera) is chosen in such a way to include the point of minimum distance from the principal axes, which is likely to fall close to the object center.

function used for this system, in fact, is

$$
\begin{aligned}
&C(w_{l1},\ldots,w_{lN_{lt}})\\
&\quad=\int\int |(I_{\mathrm{up}}(\mathbf{s}_{\mathrm{up}})-\overline{I_{\mathrm{up}}})\\
&\qquad - g_{\mathrm{ul}}(I_{\mathrm{left}}(\mathbf{s}_{\mathrm{left}})-\overline{I_{\mathrm{left}}})|^2\,\mathrm{d}\mathbf{s}_{\mathrm{up}}\\
&\qquad + \int\int |(I_{\mathrm{up}}(\mathbf{s}_{\mathrm{up}})-\overline{I_{\mathrm{up}}})\\
&\qquad - g_{\mathrm{ur}}(I_{\mathrm{right}}(\mathbf{s}_{\mathrm{right}})-\overline{I_{\mathrm{right}}})|^2\,\mathrm{d}\mathbf{s}_{\mathrm{up}}. \qquad (15)
\end{aligned}
$$

The integral is computed over the whole relevant area (a square block) of the reference image. $\overline{I_{\mathrm{up}}}, \overline{I_{\mathrm{left}}}, \overline{I_{\mathrm{right}}}$ are the average luminances of the "Up", "Left" and "Right" images, computed on the relevant area of the Up image and on the corresponding (transferred) areas on the other two images. The gains $g_{\mathrm{ul}}, g_{\mathrm{ur}}$ are estimated by analyzing several triplets of views of test objects. As the up view is the reference, the coordinates $\mathbf{s}_{\mathrm{up}}$ (expressed in pixel) assume only integer values. The relationships between corresponding coordinates in different views ($\mathbf{s}_{\mathrm{up}}$ and $\mathbf{s}_{\mathrm{left}}$; $\mathbf{s}_{\mathrm{up}}$ and $\mathbf{s}_{\mathrm{right}}$) are defined by (5). The optimal parameters are estimated through the minimization of a highly nonlinear cost function over the parameter space

$$
(\hat{w}_{l1},\ldots,\hat{w}_{lN_{lt}}) = \underset{(w_{l1},\ldots,w_{lN_{lt}})}{\arg\min}\; C(w_{l1},\ldots,w_{lN_{lt}}), \qquad (16)
$$

which is performed with a downhill simplex algorithm [16]. Of course, the estimation of the parameters of layer l is done assuming the weights of the Gaussian functions included in the previous layers are assumed as known. This means that at the next layer we will search only for a refinement of a known lower-resolution surface (see Section 5 for comments on real data).

4.3. A local approach to parameter estimation

The global optimization process described in the previous subsection works well (see experimental results in Section 5) as long as the dimensionality of the parameter space is not excessive. In our experiments we found that a global minimization is feasible up to layer 4 of the HRBF, which has *145* weights (see Table 1) to be estimated. Beyond that layer, the number of parameters becomes unmanageable. In order to overcome this difficulty, from layer 5 on, the algorithm switches to a "local mode", in which progressively smaller image regions are considered. We found that a good choice for the shape of the local region is the one of Fig. 11, where the outermost square defines the area of the reference image where luminance profiles are compared. Notice, however, that only the weights of the internal five gaussians are optimized using (15) and (16).

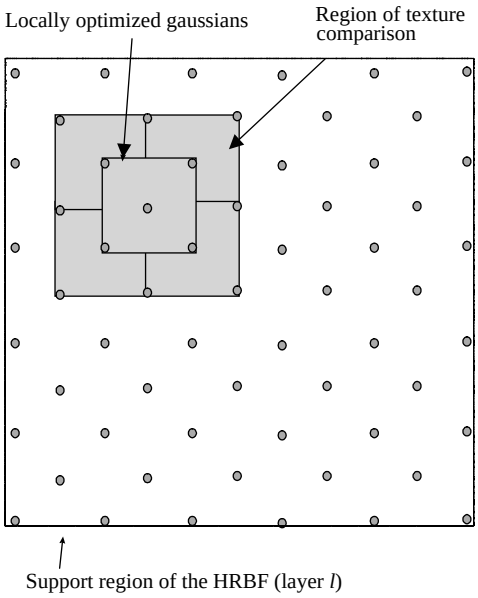


Fig. 11. Shape of the support region of the local RBF surface model on the reference image. The outermost square defines the region where luminance profiles are compared, although only the weights of the internal five gaussians are optimized.

In order to determine the surface refinement associated to layer 5 and the next ones, we will individually estimate the sets of five weights of all the local windows centered in the grid points. As the windows overlap each other and the surface shapes are individually estimated, the only shape information that we will preserve for each surface patch will be the depth of the center of the patch. The complete shape of the HRBF layer will then be rebuilt through the HRBF-based interpolation process described in Section 3.2.

One problem that arises when using a local estimation approach is that, among the incremental depths

of the new layer, we can find a number of outliers. This can be attributed to several causes: lack of luminance gradient; presence of areas with non-lambertian reflectance; occlusions or self-occlusions in one of the views; etc. These outliers, however, can be usually detected because they exhibit a high cost function (see (15) and (16)) and/or they cause unexpected and large depth changes with respect to the previous network layer. This detection process can thus be conducted through simple thresholding. Since the surface refinement is organized in an incremental fashion and the interpolation process has a smoothing action on the resulting surface, the detected outliers can be simply removed from the data to be interpolated.

4.4. Occlusion management

As underlined in the previous section, the information redundancy of trinocular acquisitions guarantees a certain robustness in the depth estimation process. On the other hand, it requires the whole surface to be simultaneously visible on all three views, which seems to increase the risk of occlusion occurrence. For example, when acquiring a trinocular front view of a human face with the acquisition system described above, there are always surface areas (e.g. the sides of the nose) which are seen either on the "Up-Right" image pair, or on the "Up-Left" image pair, but will appear as occluded on the third view. When needed, in order to overcome such problems, obvious solutions consist of switching back to a binocular approach and/or redefining the camera setup (camera positions, orientations, and focal lengths).

However, if the occlusions are not too extended, we can take advantage of the trinocular geometry in order to increase the estimation robustness and, at the same time, overcome occlusion problems. This can be done by switching between the trinocular and the binocular estimation modes only when needed. In particular, the estimation of the first layers of the HRBF network can just be done in a trinocular fashion, while ignoring the (modest) occlusion problems. When the occlusions are expected to locally disrupt the shape estimation (usually at the last one or two network layers), we can perform three minimizations per surface patch (with the same reference view): one corresponding to the whole triplet of views; one corresponding to the "Up-Left" pair; and one corresponding to the "Up-Right" pair. The depth information that is retained is the one associated to the minimum cost. Of course, the whole layer surface is reconstructed through RBF-based interpolation as discussed in Section 4.3.

5. Experimental results

We tested the proposed algorithm for image-based 3D shape estimation on different sets of real images, acquired with either digital photocameras or TV videocameras. In this section we will show the results of two such sets of views of a human subject. One is a close view of a human face and the latter is view of a head-and-shoulder scene. The human face, in fact, exhibits very fine shape details (but very little texturing) and strong occlusion problems, while the second exhibits a very complex depth distribution.

In order to acquire the first subject ("Elena", see Fig. 12), we used digital photocameras (1524 by 1012 pixels). We acquired two triplets of views: the first (Fig. 12a) with natural light, and the second (Fig. 12b) with "structured" illumination to artificially enhance the surface texturing [8]. Of course, we used the textured triplet for shape estimation, and the "natural" triplet for texture- mapping purposes [18]. As anticipated in the previous Section, we chose the "Up" image as the reference view. As we can see, some surface areas (bottom portions of nose and chin) are not clearly visible, therefore they will be only partially reconstructed. In addition, the sides of the nose are only visible on two of the three images, which makes the triplet suitable for testing our occlusion-management method.

In order to speed-up the reconstruction process, we selected a square region of 640 by 640 pixels on the reference image, which frames the subject's face.

The perspective views of the depth maps $d(x, y)$ relative to layers 1 to 7 are shown in Figs. 14–17. At each level, the surface is sampled on a regular grid corresponding to the pixel centers of the reference image. The resulting 3D point cloud is then Delaunay-triangulated (see Fig. 13) and rendered through a standard shading technique [12].

In Figs. 14 and 15 we can see perspective views of the estimated surface $d(x, y)$, as the resolution (layer index) increases. Considering the resolution of the

(a)

(b)

Fig. 12. One of the images of the original triplets of views of the subject "Elena": (a) acquisition with natural light; (b) acquisition with "structured light" to enhance the local texturing.

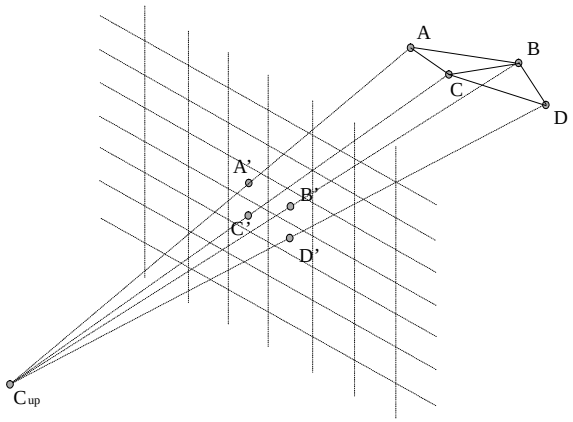
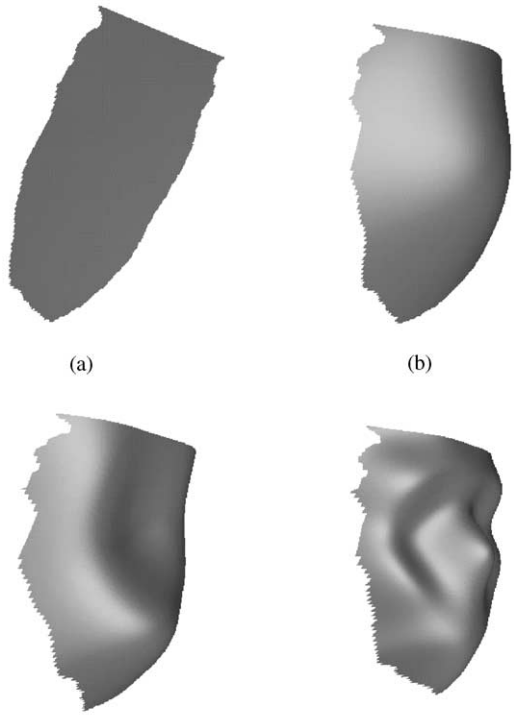


Fig. 13. Triangulation of the 3D points corresponding to the center pixel locations.



(a) (b) (c) (d)

Fig. 14. Reconstructed surface of subject "Elena": (a) first layer, with 5 gaussians; (b) second layer, with 13 gaussians; (c) third layer, with 41 gaussians; (d) fourth layer, with 145 gaussians.

local details, the reconstruction process was stopped at the 7th layer. Figs. 16 and 17 show how the management of the occlusions can improve the quality of the reconstructed surface. Two different views of the reconstructed 3D model after texture mapping are shown in Fig. 18.

Notice that, thanks to the addition of artificial texturing, the images used for building the HRBF network are very rich of luminance details, therefore there is no need to perform any outlier removal in the local estimation approach (see Section 4.3).

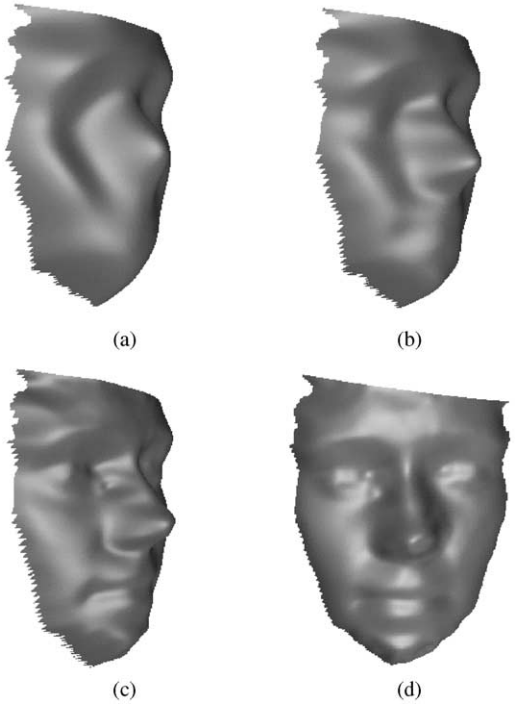


Fig. 15. Reconstructed surface of the subject "Elena": (a) the fifth layer, with 545 gaussians; (b) sixth layer, with 2113 gaussians; (c) and (d) two views of the seventh layer, with 8321 gaussians.

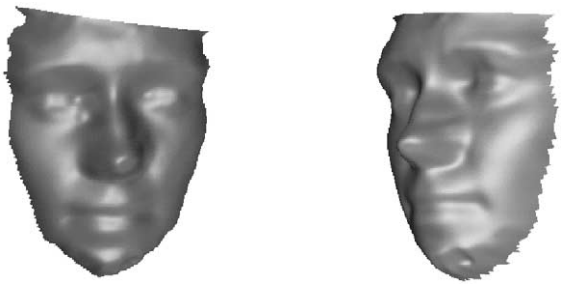


Fig. 16. Two perspective views of the reconstructed surface (HRBF network with 7 layer). It is possible to see some artifacts on the sides of the nose due to occlusion problems.

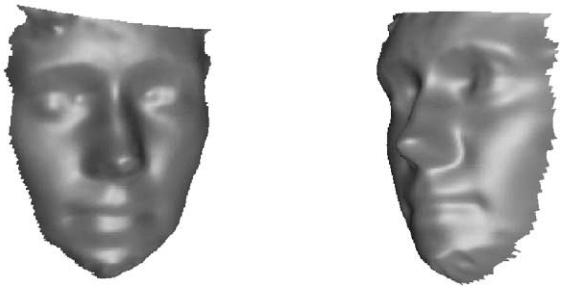


Fig. 17. Two perspective views of the reconstructed surface (HRBF network with 7 layer). In this case, the occlusion management strategy (see Section 4.4) is able to remove the artifacts of Fig. 16.

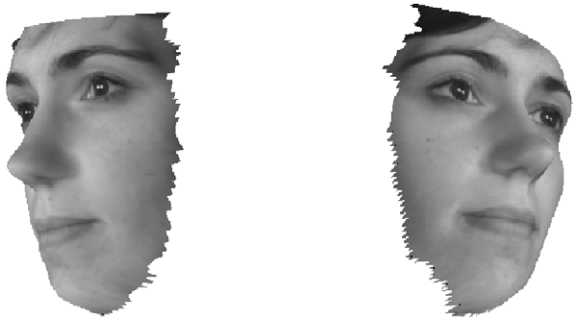


Fig. 18. Perspective views of the reconstructed 3D model after texture mapping.

The second test triplet (subject "Ludo", Fig. 19) was acquired with three TV-resolution (720 by 576 pixels) video-cameras for the ACTS-PANORAMA research project supported by the European Commission. Also in this case we chose the "Up" image as the reference view. The image area of interest (whose details appear in all three images) is a central region of 420 by 420 pixel. This is the area where we performed shape estimation. Figs. 20–25 show the reconstructed surfaces from the third to the eighth resolution levels of the HRBF network. In Fig. 26 we can see the 3D point-cloud corresponding to the last resolution layer (local approach), where some outliers are clearly visible. Such depth errors can be attributed to the lack of local texturing on the available images, but most of them can be detected and removed through statistical thresholding (see Section 4.3). The result of this outlier removal process is shown in Fig. 27. As we can see, the algorithm proved capable of a correct estimation of the surface shape even in rather complex situations. One has to bear in mind, however, that a scene with several objects (and mutual occlusions) will always be modeled with a single HRBF network,

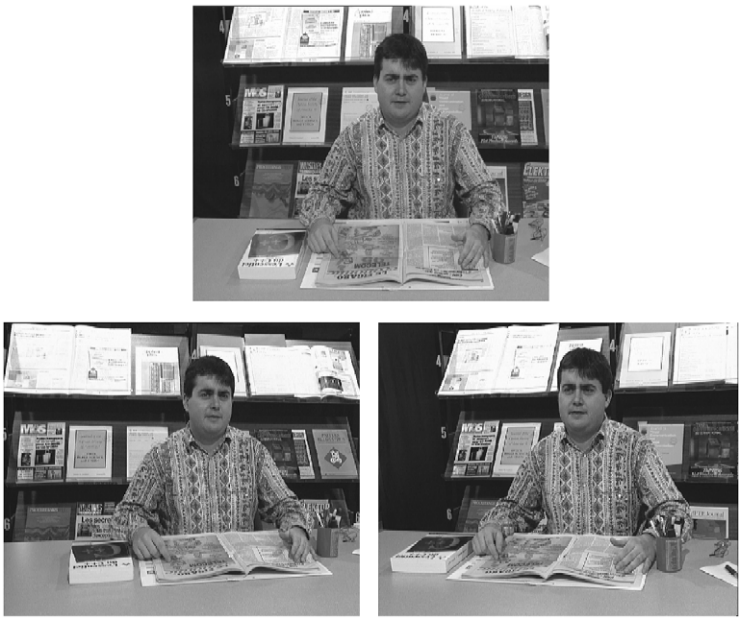


Fig. 19. Original views of subject "Ludo".

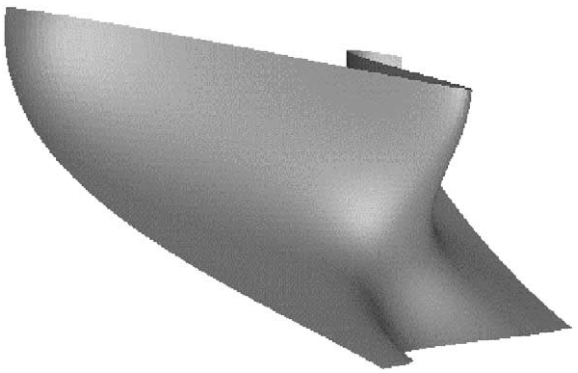


Fig. 20. Subject "Ludo": reconstructed surface at the third layer.

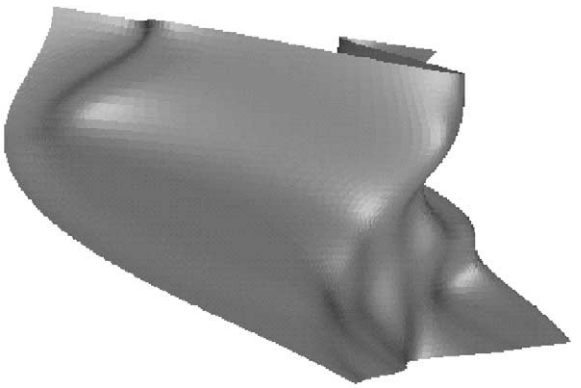


Fig. 21. Subject "Ludo": reconstructed surface at the fourth layer.

therefore a separation of the object requires additional post-processing [20].

In both examples of application ("Elena" and "Ludo") the simulations have been carried out on a PC with a Pentium III processor (300 MHz) and a Linux operating system. The program code was written in C and C++. With this platform, the complete process for an image triplet took about 2 h to be completed but significant improvement in the processing time can be still obtained through code optimization.

6. Conclusions

In this article we proposed a general and robust method for close-range 3D reconstruction of surfaces

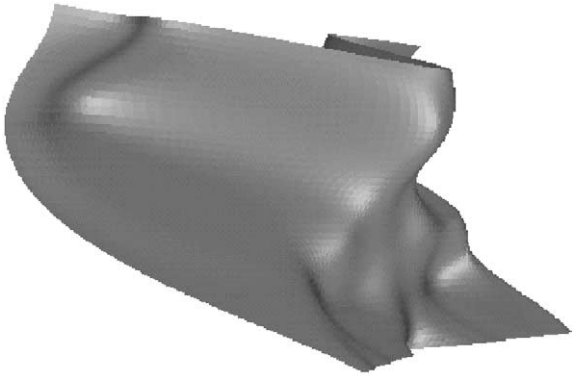


Fig. 22. Subject "Ludo": reconstructed surface at the fifth layer.

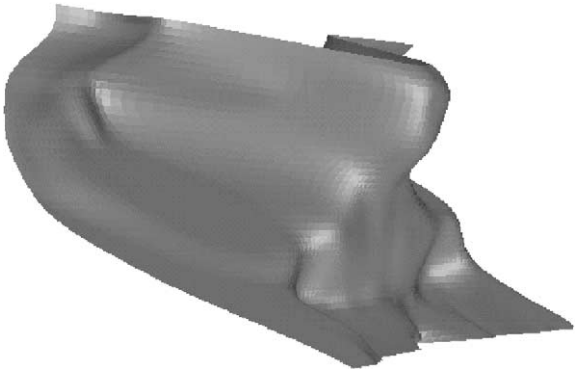


Fig. 23. Subject "Ludo": reconstructed surface at the sixth layer.

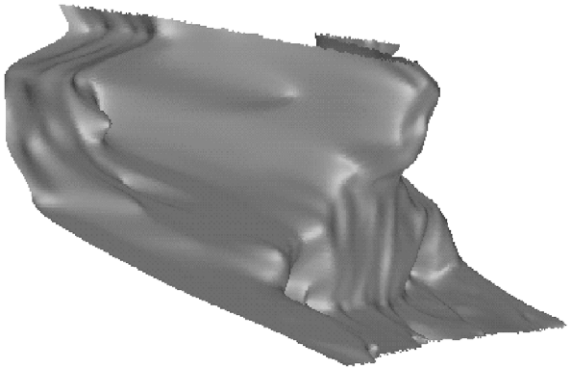


Fig. 24. Subject "Ludo": reconstructed surface at the seventh layer.

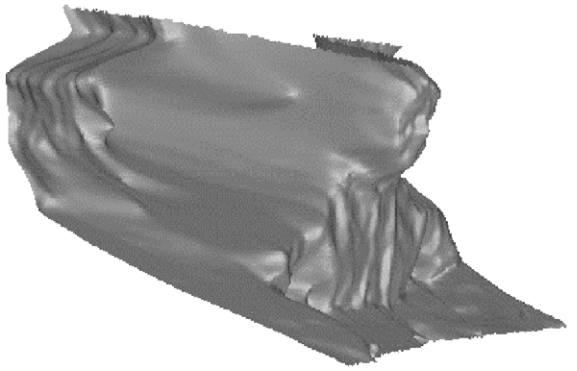


Fig. 25. Subject "Ludo": reconstructed surface at the eight layer.

through multi-resolution area matching. The method is based on the progressive refinement of a parametric surface, described by a variable set of radial functions organized as a HRBF network.

The parameters that describe the basis function in the first few resolution layers of the network are obtained through a global optimization process that minimizes the luminance differences between image pairs, which takes into account the shape-dependent luminance transfer function. This global multi-resolution approach enables the construction of the object surface without any initial model (modeling bootstrap). In order to increase the layer density further, a global estimation approach is computationally too heavy, therefore the estimation of the RBF parameters of the next layers is performed with a local approach.

The algorithm has been tested on a variety of real image triplets producing significant results.

In order to produce complete object models, it is possible to assemble several such perspective depth maps through a process of surface registration followed by fusion. We are currently investigating and developing solutions for this purpose [22]. We are also working on alternative multi-resolution methods for image-based shape modeling based on the evolution of the levelset of a volumetric function [6]. Possible further developments concern the inclusion of a radiometric surface model in the definition of the cost function, in order to improve the resilience against non-lambertian radiometric reflectance.

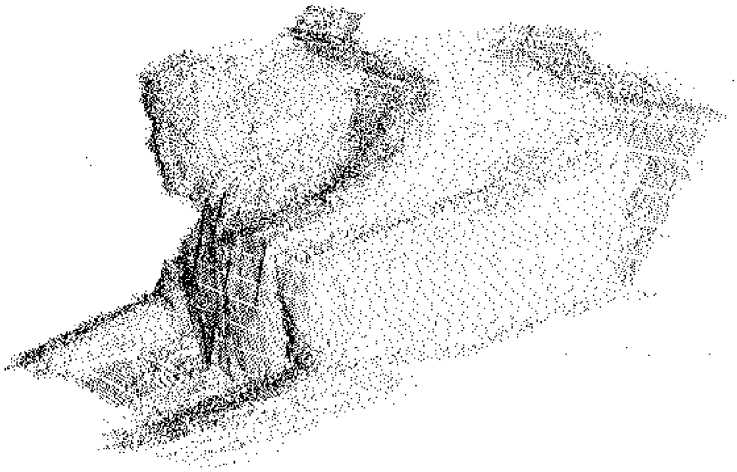


Fig. 26. Subject "Ludo". 3D point-cloud generated at the eighth layer using a local approach. Due to the limited luminance details there are several outliers.

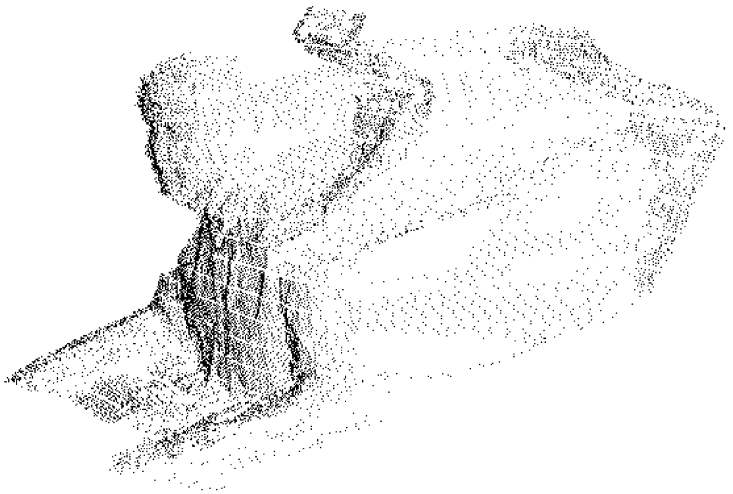


Fig. 27. Subject "Ludo". 3D point-cloud at the eighth layer after outlier removal.

References

[1] F. Argenti, L. Alparone, Coarse to fine least square stereo matching for 3D reconstruction, Electron. Lett. 26 (12) (June 1990) 812–813.

[2] B.J.C. Baxter, The interpolation theory of radial basis function, PhD Thesis, Cambridge University, August 1992.

[3] N.A. Borghese, S. Ferrari, Hierarchical RBF networks and local parameters estimate, Neurocomputing 19 (1998) 259–283.

[4] N.A. Borghese, S. Ferrari, Hierarchical RBF Networks in function approximation, Neurocomputing 19 (1–3) (1998) 259–283.

[5] N.A. Borghese, G. Ferrigno, G. Baroni, R. Savarè, S. Ferrari, A. Pedotti, AUTOSCAN: a flexible and portable scanner of 3D surfaces, IEEE Comput. Graph. Appl. 18 (3) (May/June 1998) 38–41.

[6] A. Colosimo, A. Sarti, S. Tubaro, Image-based multiresolution implicit object modeling, International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging (ICAV3D 2001), May 30–June 01, 2001, Mykonos, Greece.

[7] J.H. Conway, N.J.A. Sloane, Sphere Packings, Lattices and Group, Grundlehren der matematischen Wissenshaften, Vol. 290, Springer, Berlin, 1988.

[8] N. D'Apuzzo, Automated photogrammetric measurement of human faces, International Archives of Photogrammetry and Remote Sensing, Hakodate, Japan 32 (B5) (1998) 402–407.
[9] L. Falkenhagen, Hierarchical Block-Based Disparity Estimation Considering Neighbourhood Constraints, International Workshop on SNHC and 3D Imaging, September 5–9, 1997, Rhodes, Greece.
[10] O. Faugeras, Three-Dimensional Computer Vision: a Geometric Viewpoint, MIT Press, Cambridge, MA, 1993.
[11] O. Faugeras, R. Keriven, Complete dense stereovision using level set methods, ECCV-1998, Vol. I, Friburg, Germany, June 1998, pp. 379–393.
[12] J.D. Foley, A. van Dam, S.K. Feiner, J.F. Hughes, Computer Graphics: Principles and Practice, 2nd Edition, Addison-Wesley, Reading, MA, 1996.
[13] R. Franke, Scattered data interpolation: tests of some methods Math. Comput. 38 (1982) 121–200.
[14] L. Levi, Applied Optics—a Guide to Optical System Design, Wiley, New York, 1968.
[16] J. Nelder, R. Mead, A simplex method for function minimization, Computer J. 7 (1965) 308–313.
[17] Y. Otha, T. Kanade, Stereo by intra- and inter-scanline search using dynamic programming, IEEE Trans. PAMI 7 (2) (1985) 139–154.
[18] F. Pedersini, L. Piccarreta, A. Sarti, S. Tubaro, Estimation of radiometric parameters for a realistic rendering of 3D models, International Conference on Image Processing, ICIP-99, October 25–28, 1999, Kobe, Japan, pp. 376–380.
[19] F. Pedersini, A. Sarti, S. Tubaro, Multicamera systems: calibration and applications IEEE Signal Process. Mag. 16 (3) (May 1999) 55–65.
[20] F. Pedersini, A. Sarti, S. Tubaro, Visible surface reconstruction with accurate localization of object boundaries, IEEE Trans. Circuits Systems Video Technol. 10 (2) (March 2000) 278–291.
[21] P. Pigazzini, F. Pedersini, A. Sarti, S. Tubaro, 3D area matching with arbitrary multiview geometry, Signal Process. 14 (1–2) (1998) 71–94.
[22] A. Sarti, S. Tubaro, Multiresolution implicit surface fusion. International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging (ICAV3D 2001), May 30–June 01, 2001, Mykonos, Greece.
[23] D. Shepard, A two-dimensional interpolation function for irregularly spaced data, Proceedings of the 23rd National Conference of the ACM, New York, 1968.
[24] T. Sigitani, Y. Iiguni, H. Maeda, Image interpolation by using radial basis function networks, IEEE Trans. Neural Networks 10 (2) (March 1999) 381–390.
[25] R. Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision using off-the-shelf TV cameras and lenses, IEEE J. Robot. Automat. 3 (4) (August 1987) 323–344.
[26] F.W. Warner, Foundations of Differentiable Manifolds and Lie Groups, Graduate texts in mathematics, Vol. 94, Springer, Berlin, 1983.

Augusto Sarti was born in Rovigo, Italy, in 1963. He received the "laurea" degree (Summa cum Laude) in Electrical Engineering in 1988 and a doctoral degree in Electrical Engineering and Information Sciences in 1993, both from the University of Padova, Italy. He worked for one year for the Italian National Research Council, doing research on digital radio systems. He then spent two years doing research on nonlinear system theory at the University of California, Berkeley. Dr. Sarti is currently an Associate Professor at the Politecnico di Milano, Milan, Italy and his research interests are mainly in digital signal processing and, in particular, in video coding, image analysis for 3D scene reconstruction, audio processing and synthesis.

Stefano Tubaro was born in Novara, Italy, in 1957. He completed his studies in Electrical Engineering in 1982. He joined the Politecnico di Milano, Dipartimento di Elettronica e Informazione in 1984, doing research on voice and image coding. In 1986 he joined the CSTS (Study Center for Space Telecommunications) of the CNR (National Research Council). Since 1991 he has been an Associate Professor of Electrical Communications at the Politecnico di Milano. His current research interests are mainly on digital image processing and, in particular, on video sequence coding at low bit-rate through motion estimation and image segmentation. He also works on stereovision for 3D scene reconstruction applied to remote manipulation, autonomous vehicle guidance and telepresence.