Generalized 3D Area Matching and Multicamera Reconstruction for Scene Analysis

Federico Pedersini, Augusto Sarti, Stefano Tubaro
Dipartimento di Elettronica e Informazione (DEI), Politecnico di Milano
Piazza L. Da Vinci 32, 20133 Milano, Italy
Telephone: +39-2-2399-3647, Facsimile: +39-2-2399-3413
E-mail: pedersin/sarti/tubaro@elet.polimi.it
http://www-dsp.elet.polimi.it/ispg/frame.htm

ABSTRACT: *We propose a method for close-range 3D reconstruction of objects from stereo correspondence of luminance patches which works with an arbitrary multi-camera geometry. The area matching process is performed in 3D space as it takes both perspective and radiometric distortion into account. An implementation of the technique with a calibrated set of three standard TV-resolution CCD cameras is presented and tested over real scenes.*

1. INTRODUCTION

In the past decades we witnessed a proliferation of methods and algorithms for extracting information on the structure of a 3D scene from 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.

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 applications for robotics, 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, edges and luminance patches. These three types of features tend to provide information of a different nature. For example, luminance edges are particularly suitable for 3D reconstruction because of the intrinsic precision and reliability of their localization on the image [6]. 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/back-projection of the luminance profile of small image regions tends to return much denser sets of 3D points but it is rather sensitive to the unavoidable viewer-dependent perspective and 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

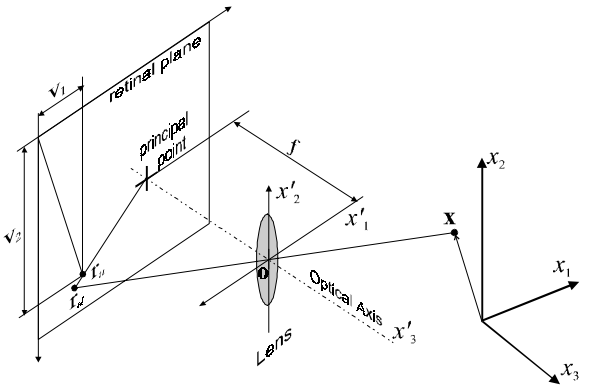


Fig. 1: Camera model.

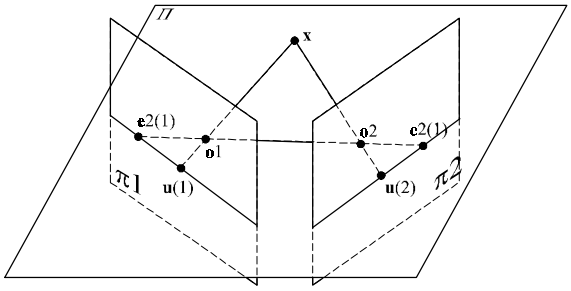


Fig. 2: Epipolar constraint.

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 takes the viewer-dependent radiometric distortion into account.

2. PRELIMINARIES

2.1 Camera Model

The camera model adopted in this paper is basically a pinhole to which a nonlinear stretching of the image plane is applied in order to take the geometric distortion of the optics into account. A pinhole model performs a projection of the object point, through the optical center of the camera, onto the retinal plane. The relationship $\mathbf{u}=\mathbf{P}\mathbf{x}$ between image coordinates $\mathbf{u}\in P^2$ and object coordinates $\mathbf{x}\in P^3$ is linear projective and is specified by a rank-3 projection matrix $\mathbf{P}$ of the form

$$\mathbf{P}=\begin{bmatrix} \mathbf{r}_1 & -\mathbf{r}_1\mathbf{O} \\ \mathbf{r}_2 & -\mathbf{r}_2\mathbf{O} \\ \frac{1}{f}\mathbf{r}_3 & -\frac{1}{f}\mathbf{r}_3\mathbf{O} \end{bmatrix}$$

where $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ are the rows of the rotation matrix $\mathbf{R}$ that describes the orientation of the camera frame in world coordinates.

2.2 Epipolar Constraint

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. Let $\mathbf{u}^{(1)}=\mathbf{P}^{(1)}\mathbf{x}\in P^2$ and $\mathbf{u}^{(2)}=\mathbf{P}^{(2)}\mathbf{x}\in P^2$ be the projective coordinates of a point $\mathbf{x}\in P^3$, as seen by two projective cameras, assuming that their projection matrices are $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$, respectively. The essential constraint can be written as

$$\left(\mathbf{u}^{(2)}\right)^T\mathbf{E}_{21}\mathbf{u}^{(1)}=0\ ,$$

where

$$\mathbf{E}_{21}=\mathbf{T}_{21}\mathbf{R}_{21}=\begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}\mathbf{R}_{21}$$

is called *essential* matrix. The above-written essential constraint is valid also when $\mathbf{u}^{(i)}$ are image coordinates. When considering n views, the essential constraint can be applied pairwise to the image coordinates of homologous points $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$, …, $\mathbf{u}^{(n)}$ as follows

$$\left(\mathbf{u}^{(i)}\right)^T\mathbf{E}_{ij}\mathbf{u}^{(j)}=0\ ,\ i,j=1,\ldots,n,\quad i>j.$$

The above property can be used as a form of point-wise multi-ocular invariance, for checking on the correctness of a match between image features.

2.3 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. Let S be a patch in object space, obtained by back-projecting a reference patch of any of the views on the surface $\mathbf{s}^T\mathbf{x}=0$, and let $S^{(i)}$ be its i-th view. The transfer of projective coordinates from the j-th view to the i-th

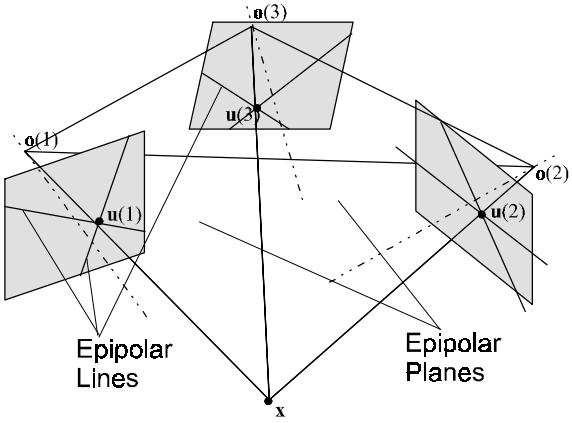


Fig. 3: Trinocular invariance.

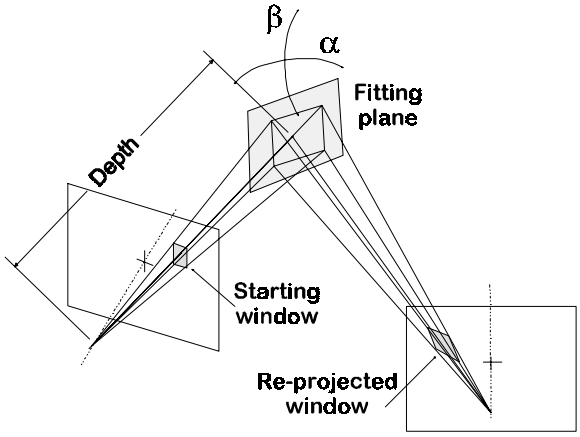


Fig. 4: Luminance transfer.

view through the plane $\mathbf{s}^T\mathbf{x}=0$, can be expressed as a homography of the form

$$\mathbf{u}^{(i)} = \mathbf{M}_{ij}(\mathbf{s})\mathbf{u}^{(j)} = 0 \ ,$$

where $\mathbf{M}_{ij}$ (**s**) is a 3 by 3 matrix which depends on the parameters of the plane over which the patch lies. This homography allows us to express the luminance transfer from the *j*-th view to the *i*-th view as

$$I_j^{(i)}(\mathbf{u}^{(i)}) = g_j^{(i)} I^{(i)}(\mathbf{M}_{ji}(\mathbf{s})\mathbf{u}^{(i)}) + \Delta_j^{(i)}$$

where $g^{(i)}_j$ is a correction factor (*gain*) that accounts for electrical differences in the camera sensors, while $\Delta^{(i)}_j$ is an additive radiometric correction (*offset*) which accounts for non-Lambertian effects of the surface reflectivity (reflection's migration with the viewpoint). Notice that the Lambertian component of the surface reflectivity does not appear in the above expression as it is the same for all views.

3. AREA MATCHING

In general, we can look at correlation-based 3D reconstruction methods as those that determine a 3D surface whose projective views result as close as possible to the actual views. This inverse problem can be thought of as that of determining a surface which maximizes a similarity measure (*correlation*) between actual views and *transferred* versions of the other views. In order to do so, it is necessary to proceed with a *local* approach, under specific stereometric constraints. Acting *locally* means describing the whole surface as a patchwork of smaller surfaces, each of which determined through a matching of luminance profiles of homologous image regions in different views. 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 *S* in object space, which lies on a generic 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 texture on the surface patch. In conclusion, area matching consists of looking for the parameters of the surface that the patch lies upon, which maximize a measure of the similarity between the back-projected corresponding image regions.

The similarity function 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 the previous Section, the transfer between points in different views, can be easily modeled when the patch is planar. On the other hand, when the surface is assumed as being reasonably smooth, 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 we can characterize the luminance transfer as done in the previous Section. We will thus look simultaneously for position and orientation of a locally planar 3D patch that originated the corresponding image areas.

Let us assume that the portion of the object surface *M* that we want to reconstruct is being imaged by a set of projective cameras that actually *see* the whole surface without occlusions. In order to determine the tangent bundle of the imaged portion of *M*, we need to find a way of *scanning* its surface. 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 *M*.

In order to determine the local surface patch that maximizes the similarity between actual image and its transferred version from the other views we minimize a MSE-like cost function of the form

$$C(\mathbf{s},\mathbf{p}) = \sum_i \sum_{j>i} C_j^{(i)}(\mathbf{s})$$

where

$$C_j^{(i)}(\mathbf{s}) = \int_{S^{(i)}} \left| I^{(i)}\left(\mathbf{u}^{(i)}\right) - I_j^{(i)}\left(\mathbf{M}_{ji}(\mathbf{s})\mathbf{u}^{(i)}\right)\right|^2 d\mathbf{u}^{(i)}$$

is the cost associated to the transfer from camera *j* top camera *i*, **s** is the vector that characterizes the plane that the patch lies on, and **p** is a vector of parameters that includes gains and offsets (radiometric corrections) that appear in the expression of the luminance transfer. Notice that minimizing the cost function defined above intrinsically corresponds to looking for the solution that best satisfies the multiocular invariance constraint of the previous Section, provided that a minimum number of three cameras is being employed. This justifies the fact that a trinocular camera systems largely outperforms a binocular system in terms of matching correctness [3,4].

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.

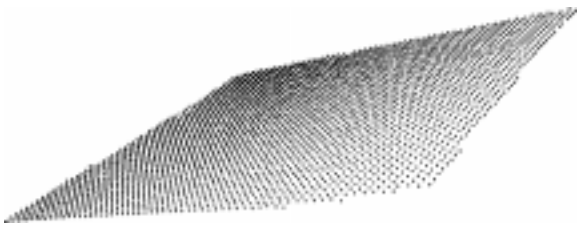
2.1 Relative Minima

The above area matching process is based on the minimization of a highly nonlinear cost function, therefore we can expect the process to be quite sensitive to the presence of relative minima. In order to avoid this problem, we can adopt several strategies, depending on the type of surface to be reconstructed.

The simplest strategy consists of starting from 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.

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 reduction of computational efficiency. The method consists of performing area matching *n* times (with narrow thresholds), each time starting from a different one of many parallel planes that scan the whole object volume. At the end of the process we can *merge* all the estimates and perform surface interpolation.

(a)



(b)

Fig. 5: Original views of the flat textured object and a view of the points reconstructed from them through area matching.

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 [5] and obtain a first rough approximation of the object

Fig. 6: Original views of a trinocular teleconference scene.

surface. At this point, the area matching process can start over with a smaller patch size and a reduced search space.

4. EXAMPLES OF APPLICATIONS

Some experiments of 3D scene reconstruction have been made on a variety of test scenes. The adopted acquisition system is a calibrated [1,2] set of three standard TV-resolution CCD cameras [3,4].

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. 5*a* 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 by 8 pixel. The reconstructed surface resulted as being flat within 0.15 mm of standard deviation (see Fig. 5*b*). Considering patches of larger size it is possible to increase this precision even further.

A second experiment concerned the 3D reconstruction of a speaker behind a desk in a tele-conferencing environment. The acquisition was made with a trinocular camera system at CCETT, France, within the ACTS "PANORAMA" Project (see Fig.6). No initial reconstruction was used for area matching. Instead, progressive-scan initialization was performed. The results of the area matching procedure are visible in Fig. 7. As we can see, the quality of the reconstruction greatly benefits from the fact that the geometric distortion is included in the model.

5. 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 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.

6. REFERENCES

[1] R.Y. 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 Autom.*, Vol. RA-3, No. 4, pp. 323-344, 1987.

[2] F. Pedersini, S. Tubaro, F. Rocca: “Camera Calibration and Error Analysis. An Application to Binocular and Trinocular Stereoscopic Systems”. 4th Intl. Workshop on *Time-Varying Image Processing and Moving Object Recognition*, Florence, Italy, 1993.
[3] F. Pedersini, S. Tubaro: “Accurate 3D reconstruction from trinocular views through integration of improved edge-matching and

Fig. 7: Two virtual views of the estimated 3D points, obtained from the three views of Fig. 1 through area matching.

area-matching techniques.” *VIII European Sig. Proc. Conf.*, Sept. 10-13, 1996, Trieste, Italy.
[4] F. Pedersini, A. Sarti, S. Tubaro: “A multi-view trinocular system for automatic 3D object modeling and rendering.” *XVIII Intl. Congr. for Photogrammetry and Remote Sensing*, July 9-19, 1996, Vienna, Austria.
[5] J.L. Mallet. “Discrete Smooth Interpolation”. *ACM Trans. on Graphics*, Vol. 8, No. 2, pp. 121-144, 1989.
[6] N. Ayache, *Artificial vision for Mobile Robots*, MIT Press, 1991.