Robust Area Matching

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 general and robust solution to the problem of close-range 3D reconstruction of objects from stereo correspondence of luminance profiles. The method does not require a particular camera geometry, and can be implemented with an arbitrary number of CCD cameras. Its robustness can be mainly attributed to the physicality of the matching process, which is performed in the 3D space, while taking both geometric and radiometric distortions into account. Extensive tests have been performed over a variety of real scenes, using a calibrated trinocular camera system.

Introduction

A crucial problem of many stereometric methods for the automatic measurement and reconstruction of close-range objects is the robustness of the matching process between homologous features.

The image features that are usually considered for stereo matching are luminance edges and luminance patches. These two types of features tend to provide information of a rather different nature. Edge matching is generally very accurate and reliable, but it usually generates a sparse set of 3D data. Conversely, the matching/back-projection of the luminance profile of small image patches tends to provide 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.

The approach we propose in this paper represents 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 employs a calibrated [1,2] set of three or more standard TV-resolution CCD cameras, which provides enough redundancy for removing possible matching ambiguities. The robustness of the approach can be attributed to the *physicality* of the matching process, which is actually performed in the 3D 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 takes into account the viewer-dependent radiometric distortion.

Preliminaries

In this paper the camera is modeled basically as a pure perspective projection onto an image plane to which a nonlinear stretching is applied in order to take the geometric distortion of the optics into account. The relationship $\mathbf{u}=\mathbf{P}\mathbf{x}$ between image coordinates $\mathbf{u}\in\mathcal{P}^2$ and object coordinates $\mathbf{x}\in\mathcal{P}^3$ 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.

Two projective views of the same point in 3D space, are bound to comply with the so-called "epipolar" (or "essential") constraint, according to which the two optical rays (lines that connect the object point with the optical centers of the two projective cameras) are coplanar. Let $\mathbf{u}^{(1)}=\mathbf{P}^{(1)}\mathbf{x}\in\mathcal{P}^2$ and $\mathbf{u}^{(2)}=\mathbf{P}^{(2)}\mathbf{x}\in\mathcal{P}^2$ be the projective coordinates of a point $\mathbf{x}\in\mathcal{P}^3$, 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, $\mathbf{R}_{21}$ and $\mathbf{T}_{21}$ are the rotation matrix and the translation vector (in skew-symmetric matrix form) that describe the change of reference frame between camera 1 to camera 2. The essential constraint remains valid also when $\mathbf{u}^{(i)}$ are scaled in such a way for its first two components to become image coordinates. When considering n views, the epipolar constraint can be applied pairwise to the image coordinates of homologous points $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$ as follows

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

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

Area matching requires a comparison between the actual luminance profile of a patch with the one that we obtain by *transferring* luminance profiles of other views through a specific 3D surface model. Let $\mathcal{S}$ be a surface patch in object space, obtained by back-projecting a reference image patch of any of the views onto the plane $\mathbf{s}^T\mathbf{x}=0$, and let $\mathcal{S}^{(i)}$ be its i-th view. The transfer of projective coordinates from the j-th view to the i-th view through the plane $\mathbf{s}^T\mathbf{x}=0$, can be expressed as a homography (an invertible linear projective transformation) of the form

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

where $\mathbf{M}_{ij}(\mathbf{s})$ 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.

Area Matching

As 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.

Our approach to 3D area matching consists of modeling the object's surface as a patchwork of smaller surfaces, each of which is determined through a matching of luminance profiles of homologous image regions in different views. The determination of the position and the orientation of a 3D patch is done in such a way to maximize a similarity measure (*correlation*) between the actual view within the image patch and a *transferred* version of the other views through the 3D patch.

Let us consider a patch $\mathcal{S}$ in object space, which lies on a certain parametric surface. This patch is a good approximation of the object surface when the the back-projection of the luminance profiles onto the 3D patch are maximally similar.

The similarity function can be either computed on the planar surface in 3D space, or on 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. This choice is reasonable as the surface can be assumed as being smooth enough to be well-described by its *tangent bundle*. Therefore, 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 $\mathcal{M}$ 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 $\mathcal{M}$, 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 $\mathcal{M}$.

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

$C_j^{(i)}(\mathbf{s}) = \int\limits_{S^{(i)}} \left| I^{(i)}(\mathbf{u}^{(i)}) - I_j^{(i)}(\mathbf{M}_{ji}(\mathbf{s})\mathbf{u}^{(i)}) \right|^2 d\mathbf{u}^{(i)}$ is the mean square error associated to the transfer from camera j top camera i, $\mathbf{s}$ is the vector that characterizes the plane that the patch lies on, and $\mathbf{p}$ is a vector of parameters that includes gains and offsets (radiometric corrections) that appear in the expression of the luminance transfer. Minimizing this cost function corresponds to looking for the solution that best satisfies (intrinsically) 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.

The above area matching process is based on the minimization of a highly nonlinear cost function, therefore we can expect the process to be rather incline to terminate in correspondence to 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 drastic 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 for by a reduction of computational efficiency. The method consists of performing area matching many times (with narrow thresholds), every time starting from a different one of many parallel planes that regularly slice the whole object's volume. At the end of the process we can *merge* all the estimates and perform surface interpolation.

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

Examples of Application

Some experiments of 3D scene reconstruction have been carried out on several test scenes. The first test we performed was for measuring the accuracy of the reconstruction of a flat textured object, placed at about 1.2 m of distance from the camera system. The surface reconstruction resulted to be flat with 0.1 mm of standard deviation.

Another reconstruction experiment concerned a stone of the Roman Amphitheater of Aosta, Italy. Also in this case, a quantitative evaluation of the quality of the results has been possible: we found that our reconstruction results agreed with the measurements taken with classical photo-rammetric methods. In Fig. 2 one of the original views of the object is shown. In Fig. 3 the points extracted through area matching are seen from a different viewpoint. In Fig. 4 a virtual view of the object after surface interpolation and texture mapping has been obtained.

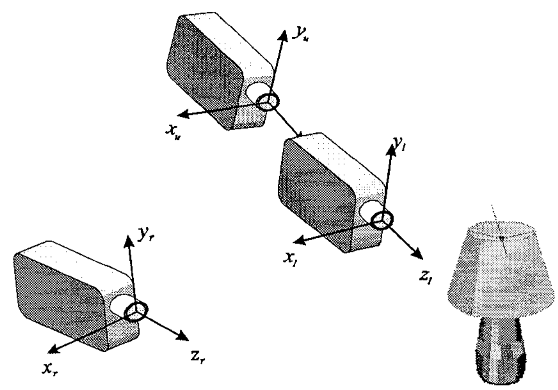

Fig. 1: Trinocular acquisition system.

Fig. 2: one of the views of a stone of the Roman Amphitheater of Aosta (fiducial marks added for photogrammetric comparison)

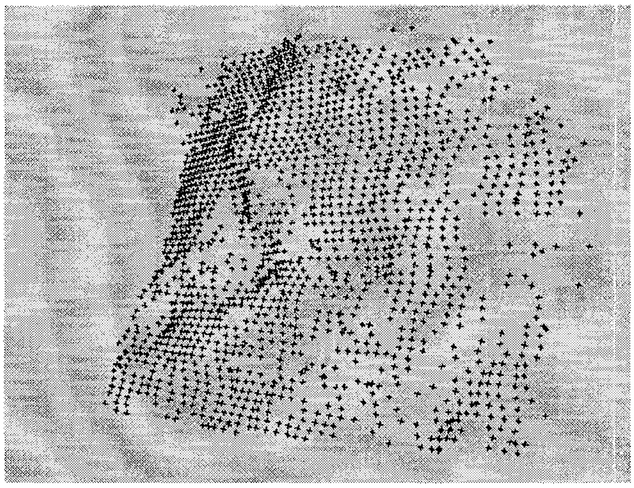

Fig. 3: a perspective view of the 3D points extracted through area-matching

Fig. 4: a virtual view of the reconstructed surface after interpolation and texture mapping.

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.

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 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.