3-D Scene Representation as a Collection of Images

S. Laveau* O.D. Faugeras
INRIA. 2004, route des Lucioles. B.P. 93. 06902 Sophia-Antipolis. FRANCE.
e-mail: Stephane.Laveau@sophia.inria.fr, Olivier.Faugeras@sophia.inria.fr

Abstract

In this paper, we address the problem of the prediction of new views of a given scene from existing weakly or fully calibrated views called *reference views*. Our method does not make use of a three-dimensional model of the scene, but of the existing relations between the images. The new views are represented in the reference views by a viewpoint and a retinal plane, i.e. by four points which can be chosen interactively. From this representation and from the constraints between the images, we derive an algorithm to predict the new views. We discuss the advantages of this method compared to the commonly used scheme: 3-D reconstruction-projection.

1 Introduction

The problem we solve in this paper is the following. Suppose we are given N views of a static scene obtained from different viewpoints, perhaps with different cameras. These viewpoints we call *reference viewpoints* since they are all we know of the scene. We would like to decide if it is possible to predict another view of the scene taken by a camera from a viewpoint which is arbitrary and a priori different from all the reference viewpoints. One method for doing this would be to use these viewpoints to construct a three-dimensional representation of the scene and reproject this representation on the retinal plane of the virtual camera. In order to achieve this goal, we would have to establish some sort of calibration of our system of cameras, fuse the three-dimensional representations obtained from, say, pairs of cameras thereby obtaining a set of 3-D points, the scene. We would then have to approximate this set of points by surfaces, a segmentation problem which is still mostly unsolved, and then intersect the optical rays from the virtual camera with these surfaces. We do not claim that there does not exist any simpler way of using the three-dimensional representation than the one we just sketched, but this is just simply not our point.

Our point is that it is possible to avoid entirely the explicit three-dimensional reconstruction process: the scene is represented by its images and by some basically linear relations that govern the way points can be put in correspondence between views when they are the images of the same scene-point. These images and their algebraic relations are all we need for predicting a new image. This approach is similar in spirit to the one that has been used in trinocular stereo. Hypotheses of correspondences between two of the images are used to predict features in the third. Related to these ideas are those developed in the photogrammetric community under the name of ***transfer*** methods which find for one or more image points in a given image set, the corresponding points in some new image set. If the camera geometries are known, transfer is done in a straightforward fashion by 3-D reconstruction and reprojection. If the camera geometries are unknown, this can still be done by methods based on using projective invariants. As a third source of correspondence, people interested in recognition and pose estimation have recognized recently that the variety of views of the same object can be expressed as the combinations of a small number of views [10].

2 The Approach

We make heavy use of elementary projective geometry. The reader who is unfamiliar with these notions is referred to the recent computer vision literature on the subject [7, 3]. Given a pair of images, we call them weakly calibrated when the epipolar geometry between the images is known. When we consider N>2 views, we use $\mathbf{F}_{ij}$ to denote the fundamental matrix between views i and j (in that order). e_{ij} is the epipole in view i with respect to view j.

2.1 Two reference images

Let us consider first the case where two views are available. The knowledge of the fundamental matrix is sufficient to obtain a set of point correspondences between the two views. We can set up the third view by selecting two corresponding points e_{13} and e_{23} in the two images. These points may or may not be the images of a real point of the scene, but they must satisfy the epipolar constraint . We say that this point is the new point of view. In the weakly-calibrated situation, the notion of perpendicularity does not exist and we cannot define the retinal plane as perpendicular to the viewing direction and must therefore define this plane directly from the two views. Since a plane is de-

*This work was partially supported by DRET contract No 91-815/DRET/EAR and by the EEC under Esprit project 6448, Viva

1051-4651/94 $04.00 © 1994 IEEE

fined by three points, we can select them interactively like we did for the optical center, making sure that the three pairs of corresponding points each satisfy the epipolar constraint.

These three pairs of points, plus a fourth one which is determined as explained in [5] will be considered as being the images of a projective basis of the retinal plane of the virtual camera. We call them control points. In order to construct the new image, we need the following two ingredients:

- A set of point correspondences between the two reference views.
- A way to compute the intersection of the optical rays of the virtual camera with its retinal plane.

The first ingredient is obtained through the use of standard stereo algorithms. The reason for the second ingredient should be pretty clear. For each point correspondence (m^1, m^2) between the two reference views, we consider the two image lines $\langle e_{13}, m^1 \rangle$ and $\langle e_{23}, m^2 \rangle$. They are the images of the optical ray from the virtual optical center to the scene point whose images are m^1 and m^2. Therefore, we need to compute the intersection of this optical ray with the retinal plane. This problem has been solved by several authors [8, 9]. More precisely, if p^1 is the image of the point of intersection of the optical ray defined by the two image lines $\langle e_{13}, m^1 \rangle$ and $\langle e_{23}, m^2 \rangle$ with the retinal plane $\mathcal{R}$ of the virtual camera, p^1 can be computed as the intersection of $\langle e_{13}, m^1 \rangle$ and $\mathbf{H}_{21}^T . \langle e_{23}, m^2 \rangle$, $\mathbf{H}_{12}$ being the homography defined by the 4 pairs of corresponding points $(m_1^1, m_2^1, m_3^1, m_4^1)$ and $(m_1^2, m_2^2, m_3^2, m_4^2)$.

Having built these points, their projective coordinates in the projective basis formed by the four reference points can be read directly from the reference images. This allows us to construct a collection of points in a projective plane defined by the four reference points: the virtual retinal plane. If the choice of the four points has been made in an arbitrary fashion, it is seen that the image is obtained as a distortion of the "real" image by an unknown planar projective transformation. It turns out that a straightforward implementation of these ideas does not work well because of the appearance of gaps in the predicted image caused by the irregular distribution in 3 of the pixels in 1 and 2 as seen from the new viewpoint. This is why we develop in the next section an alternative approach based on the same principles.

2.2 More than two reference views

We now assume that we are given $N > 2$ reference views and the complete epipolar geometry between these views. It has been shown elsewhere [6] that this depends in fact only on $18 + 11(N-3)$ parameters.

The procedure for predicting a new view from a viewpoint which is different from the existing N is then very similar to the two views case. The epipolar geometry between the reference views 1 and 3, and 2 and 3 being known, we can in fact propagate this information very simply. For example, the image e_{34} of the new viewpoint is obtained by intersecting the epipolar lines $l_{e_{14}}^3$ of e_{14} and $l_{e_{24}}^3$ of e_{24} in the third reference view. Similar computations can be done for the points m_i^1, m_i^2.

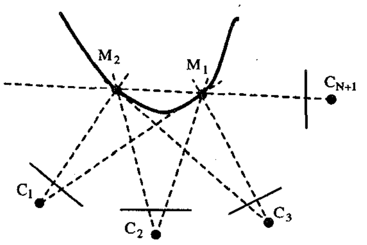

Figure 1: Ambiguities: multiple solutions..

2.3 Strongly Calibrated Images

In this paragraph, we assume that our system is fully calibrated. The algorithm described previously is still valid. But since we know the internal parameters of the cameras, we can use them to eliminate the unknown projective transformation mentioned previously. A very intersting point of this method compared with the reconstruction is that we need only two strongly calibrated images. The images of the control points in every other reference views can be inferred from the weak calibration.

3 Implementation of the Ray-tracing like algorithm

The solution to the problem discussed in section 2.1 is well known in image synthesis: the scanning must take place directly in the target image.

Given a point m^3 in $\mathcal{R}_3$, we can draw p^1 and p^2 in $\mathcal{R}_1$ and $\mathcal{R}_2$ since they have the same projective coordinates. The projections of the optical ray $\langle C_3, m^3 \rangle$ in $\mathcal{R}_1$ and $\mathcal{R}_2$ (O^1 and O^2) are $\langle p^1, e_{13} \rangle$ and $\langle p^2, e_{23} \rangle$.

The problem now is first to find where the scene points are located on the optical ray and, second, among the possibly several points, which one leads to the correct interpretation, i.e. is the closest to the virtual camera.

3.1 Physical points

If $m^1 \in O^1$ is the image of a physical point M, there are constraints on its correspondent m^2 in $\mathcal{R}_2$. First, m^2 lies on O^2 because M belongs to the line $\langle C_3, m^3 \rangle$. Second, M being a physical point, m^2 lies on $d_{12}(O^1)$ the image of O^1 by the disparity map. m^2 is one of the intersections between $d_{12}(O^1)$ and O^2

3.2 Disambiguating whenever possible

If $d_{12}(O^1)$ and O^2 do not meet, it means that the point is occluded in either 1 or 2 or is not correlated and therefore no information can be obtained. If $d_{12}(O^1)$ and O^2 meet once, there is no ambiguity.

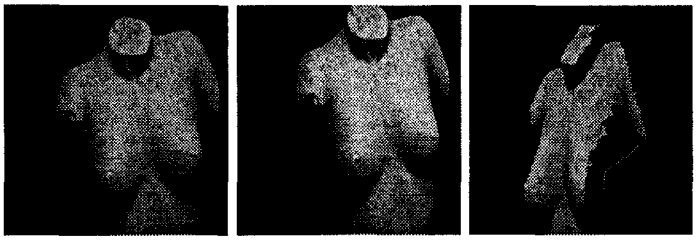

Figure 2: Face images of the mannequin (left images) and predicted side view of the manequin (right image)

If $d_{12}(O^1)$ and O^2 meet in more than one point, the optical ray intersects an object in the physical world in more than one point too (see figure 1).

We want to sort the corresponding points M_1, M_2 with respect to the distance to the optical center C_{N+1} in the world. Let us choose arbitrarily a view i in which M_1 and M_2 are seen. They are physical points, therefore they are in front of the focal plane $\mathcal{F}_i$ of this reference view. If C_{N+1} is in front of the focal plane, the order of M_1, M_2 and C_{N+1} will be preserved by perspective projection, whereas it will be inversed if C_{N+1} is behind $\mathcal{F}_i$. If we are fully calibrated, We know the 3-D position of C_{N+1} and $\mathcal{F}_i$. If we are weakly calibrated, but we have identified the plane at infinity, then we can determine the vanishing point v_∞ of the line passing through C_{N+1}, M_1 and M_2. Then, if v_∞ is between C_{N+1} and M_1 and M_2, C_{N+1} is behind the focal plane. This reasoning seems to be somehow related to the difficult problems raised in [4]. Note that it is always possible to disambiguate in the strongly calibrated case.

3.3 Generalization to the case of an arbitrary number of views

Suppose now that we are given N views of the scene and wish to predict an $(N+1)$st view. The user chooses the control points in say views 1 and 2. The algorithm proceeds as follows:

- For each image i do
  - Compute the disparity function $d_{i\,i+1}$ between views i and $i+1$.
  - From $\mathbf{F}_{i-1\,i+1}$ and $\mathbf{F}_{i\,i+1}$ compute the control points of image $i+1$ from the control points in image i and $i-1$.
- For each pixel m^{N+1} in image $N+1$ do
  - Compute O_i in every image i.
  - Iteratively, scan O_i and its image in $i+1$ to find the physical points M on the optical ray. If possible, disambiguate.

4 Experimental Results

We took 2 front images of our mannequin and we predicted what a side view would be (Figure 2). The occlusions are well dealt with as can be seen on the left breast and on the neck. The strip on the right is an error due to false matches given by our correlation. We are currently working on improving our correlation algorithms to avoid these false matches. Of course, the unseen areas in the reference views are not visible in the transferred image (the right breast, the right part of the neck). More results can be found in [5].

5 Conclusion

We have proposed a method for representing a 3-D scene which does not involve an explicit reconstruction. It rather considers the scene as a collection of images related by simple algebraic relations. We have shown elsewhere [2] that these relations allow us to compute 3-D information about the scene. We show here that they allow the prediction of an image of the scene from an arbitrary viewpoint. We believe that representing 3-D data as images could be as powerful as using a complete 3-D model. One advantage over existing methods of reconstruction and projection is that we do not need calibration for all reference views, but only for two of them.

References

[1] Eamon B. Barett, Michael H. Brill, Nils N. Haag, and Paul M. Payton. *Invariant Linear Methods in Photogrammetry and Model-Matching*, chapter 14. MIT Press, Cambridge, MA, 1992.

[2] O.D. Faugeras. What Can be Seen in Three Dimensions with an Uncalibrated Stereo Rig ? In Giulio Sandini, editor, *Proc. European Conference on Computer Vision*, pages 563–578, Santa Margherita Ligure, Italy, 1992.

[3] O.D. Faugeras. *Three-Dimensional Computer Vision: a geometric viewpoint.* MIT Press, 1993.

[4] R. I. Hartley. Cheirality Invariants. In *Proc. DARPA Image Understanding Workshop*, pages 745–753, Washington, DC, April 1993.

[5] Stéphane Laveau and Olivier Faugeras. 3-D Scene Representation as a Collection of Images and Fundamental Matrices. Technical Report 2205, INRIA, Sophia-Antipolis, France, February 94. available by ftp at ftp.inria.fr.

[6] Q.-T. Luong and T. Viéville. Canonic Representations for the Geometries of Multiple Projective Views. In *3rd E.C.C.V., Stockholm*, 1994.

[7] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision.* MIT Press, 1992.

[8] L. Robert and O.D. Faugeras. Relative 3D Positioning and 3D Convex Hull Computation from a Weakly Calibrated Stereo Pair. In *Proc. First International Conference on Computer Vision*, pages 540–544, Berlin, Germany, May 1993.

[9] Amnon Shashua. Projective Depth: A Geometric Invariant for 3D Reconstruction From Two Perspective / Orthographic Views. In *Proc. First International Conference on Computer Vision*, pages 583–590, 1993.

[10] Shimon Ullman and Ronen Basri. Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.