Proc. 1997 Image Understanding Workshop, pp. 869-874, May 1997

Multi-Image Correspondence using Geometric and Structural Constraints

George T. Chou[*] **Seth Teller**
MIT Computer Graphics Group
545 Technology Square NE43-208
Cambridge MA 02139
gtc@graphics.lcs.mit.edu, http://www.graphics.lcs.mit.edu

Abstract

In this paper, the problem of recovering three-dimensional information from multiple images is considered. The goal is to build a system that can incrementally process images acquired from arbitrary camera positions. Our approach makes use of both the geometric constraints inherent in the camera configuration, as well as the structural relationships between image features. The correspondence problem is analyzed directly in 3D through multi-image triangulation. To address the possibilities of false features and spurious correspondence, every initial match is modeled as a hypothesis. At the core of our system is a state machine which keeps track of matching hypotheses in various states of certainty, and evolves their states in response to new evidence.

1 Introduction

Traditional multi-image stereo analysis typically assumes that input images are temporally coherent. A short baseline between consecutive images is essential for constraining the matching and tracking processes in these systems. But in the context of a large-scale reconstruction project to distill thousands of images into thousands of structures, this assumption becomes rather limiting. We wish to be able to process images in arbitrary order, without temporal constraints on the input.

Algorithms crafted without dependence on temporal coherence have a number of advantages. ¿From geometry it is well known that long baseline stereo can produce more stable and precise reconstruction than short baseline stereo. Practically, it also simplifies the image acquisition process. No longer will there be a problem with camera motion control. Moreover, since we can record images at a larger sampling interval, fewer images need to be taken. This will greatly relieve the burden of storing and processing high resolution image data.

Given these considerations, we have designed a method for the recovery of 3D structure from multiple images of an urban scene. The algorithm operates by establishing long-baseline correspondences between 3D features. However, just as in 2D, spurious matches can occur in 3D. The occurrence of false matches can be significantly reduced by supplementing geometric constraints of imaging configuration with knowledge about structural relationship of image features. Still, feature detection is by no means a flawless process. Each matching hypothesis must be supported by a sufficient number of observations before it can be confirmed. A state machine has been developed to keep track of the hypotheses.

2 Previous Work

Over the years, stereo researchers have explored countless ways to improve the performance of

[*] Funding for this research was provided in part by the Advanced Research Projects Agency of the Department of the Defense under Ft. Huachuca contract DABT63-95-C-0009.

stereo algorithms. A primary objective is to establish reliable correspondence across two or more images. The challenge is that when a large area must be searched for a match, the potential for spurious matches increases also.

In response to this problem, researchers first turned to coarse-to-fine methods [Grimson, 1981], [Terzopoulos, 1983]. In these systems, matching begins at a low resolution of the image in order to cover large displacements. Matching then proceeds to higher resolutions where results from lower resolutions are used to constrain the search. This class of method cannot deal with significant perspective distortion and occlusion present in long baseline images.

Other researchers advocated using multiple images acquired with closely spaced cameras as a way of extending the baseline of analysis while minimizing false matches [Herman and Kanade, 1986], [Baker and Bolles, 1989]. By exploiting the temporal coherence of very short baseline images, stereo correspondence can be performed accurately through incremental tracking of pixels or features. Although these methods seem to work well, they are dependent on the temporal coherence of the input for reliable feature tracking. They cannot, for example, associate images which are taken at very different times, but which contain observations of identical real-world structures.

Another approach is to utilize the structural relationship between image features to resolveg matches [Lim and Binford, 1988], [Horaud and Skordas, 1989]. It has been observed that structural properties tend to be more invariant with respect to viewing changes than local image/feature properties. The problem of correspondence then becomes a problem in finding the mapping which best preserves the structural relationship. Because these methods often assume their feature extraction process as ideal, they tend to be fragile with real images.

Recently, new algorithms capable of analyzing long baseline inputs have been proposed. Bedekar and Haralick [1996] describe a method for Bayesian triangulation and hypothesis testing. A major drawback of their work is that they do not consider the possibility of spurious matches.

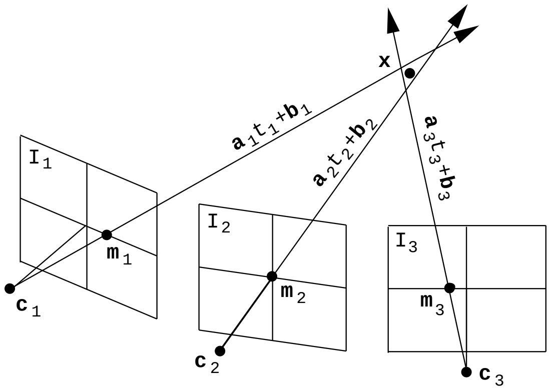
x
$\mathbf{a}_1t_1+\mathbf{b}_1$
$\mathbf{a}_2t_2+\mathbf{b}_2$
$\mathbf{a}_3t_3+\mathbf{b}_3$
I_1
I_2
I_3
m_1
m_2
m_3
c_1
c_2
c_3

Figure 1: Multi-image triangulation

Collins [1996] present a space-sweep approach to multi-image matching. The problem with this method is that it uses a constant threshold for rejecting false matches, and so does not handle underlying causal factors in a generic fashion.

3 Multi-Image Triangulation

The basic principle underlying the recovery of three-dimensional information from two-dimensional images is triangulation. Suppose we are given the corresponding image positions $\mathbf{m}_i$ of a 3D point $\mathbf{x}$ projected onto a set of images I_i. We can compute the 3D position of the point by finding the intersection of rays projected, respectively, from camera $\mathbf{c}_i$ and passing through the image feature $\mathbf{m}_i$ (Figure 1).

Typically the rays will not intersect precisely at one point. However, a well-fitting point $\mathbf{x}$ can be estimated with the least squares method. Our goal is to minimize the sum of squared distances of the rays to point $\mathbf{x}$:

$$D(\mathbf{x}) = \sum_i (\mathbf{a}_i t_i + \mathbf{b}_i - \mathbf{x})^T (\mathbf{a}_i t_i + \mathbf{b}_i - \mathbf{x}) \quad (1)$$

where $\mathbf{a}_i$ is the direction of the ray i, and $\mathbf{b}_i$ is an arbitrary point on the ray (usually taken to be the camera position $\mathbf{c}_i$).

Setting $dD(\mathbf{x})/d\mathbf{x} = 0$, we get

$$\begin{aligned} &\sum_i (\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I})^T (\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I})\, \mathbf{x} \\ = &\sum_i (\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I})^T (\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I}) \mathbf{b}_i \end{aligned} \quad (2)$$

We note that this is a linear system, $\mathbf{Ax} = \mathbf{b}$. Using singular value decomposition the matrix $\sum_i (\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I})^T(\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I})$ can be decomposed into

$$\sum_i (\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I})^T(\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I}) = \mathbf{UWU}^T \qquad (3)$$

where $\mathbf{U}$ is an orthonormal matrix satifying $\mathbf{U}^T = \mathbf{U}^{-1}$, and $\mathbf{W}$ is a diagonal matrix containing the singular values. The least squares estimate $\hat{\mathbf{x}}$ is then

$$\hat{\mathbf{x}} = \mathbf{UW}^{-1}\mathbf{U}^T\left(\sum_i (\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I})^T(\mathbf{a}_i\mathbf{a}_i^T - \mathbf{I})\mathbf{b}_i\right) \qquad (4)$$

The residual of the intersection process is given by $D(\hat{\mathbf{x}})$ in Equation (1).

4 Matching via Triangulation

Suppose we hypothesize that a set of image features is in correspondence. A direct method for testing the hypothesis would be to apply multi-image triangulation on the established feature set, and examine the residual of the least square computation. If the residual is greater than a certain threshold, there is no single 3D point near which all of the rays pass, and the correspondence hypothesis can be rejected.

However, we cannot hastily accept any intersection of rays as a match. Figure 2 illustrates a case in point. In the figure, the rays of vertices a_1 and a_2 intersect with the rays of vertices b_1 and b_2 by accident. By themselves, the accidental intersections could be interpreted as a line floating in front of two buildings. This is clearly incorrect, and situations like this are not uncommon. Whenever multiple images are shot with a camera revolving around some region in space, there will be many rays crossing very close together.

Additional observation of the the features is needed for resolving this ambiguity. Certain structural properties, for example adjacency, are invariant with respect to large changes in viewing direction. Connectivity of vertices is a useful structural property in this regard. For two vertices to match, we require that at least two of their incident edges must match also. In the case of Figure 2, we test whether two incident edges of a_1 match with two incident edges of b_1, and can quickly reject this configuration as a spurious intersection. To follow this strategy, we need to determine vertex connectivity information from the images.

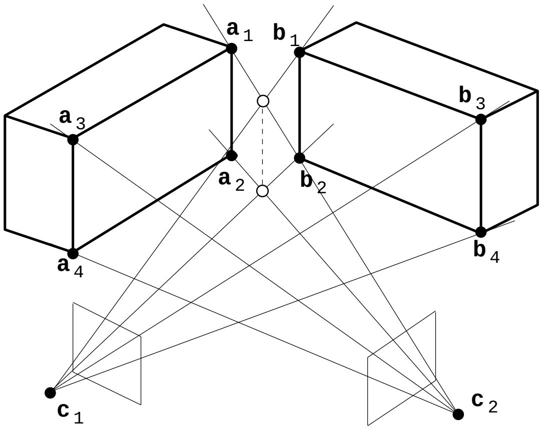

Figure 2: Two spurious matches

Unfortunately, no existing feature extraction algorithm is perfect. We may never be certain that a feature detected from an image is not an artifact of the extraction process. For instance, occlusion often generates incidental features like T-junctions. Since T-junctions are not intrinsic to any real 3D object, their presence can confuse the matching process. Due to these complications, every matching hypothesis should begin with a low degree of certainty.

4.1 The Data Structures

We list here five types of data structures that are relevant to our algorithm. The first two are image features, and the last three are matching hypotheses in increasing states of certainty.

- *2D lines* – are extracted by fitting lines to the output of an edge detector. The system constructs 2D lines only for the purpose of vertex detection.
- *2D vertices* – are located by intersecting 2D lines that form an L-junction. Vertices are the key features used in the correspondence process. Each vertex is described by: 1) a

label, 2) an image position, 3) the number of incident lines, and 4) any connected adjacent vertices.

- *2D hypothesis* – is a list of matched 2D vertices with a combined baseline too short to produce a reliable 3D estimate. Each 2D hypothesis is described by: 1) a label, 2) the number of contributing vertices, 3) a list of matched 2D vertices, and 4) baseline information.
- *3D hypothesis* – is a list of matched 2D vertices with a 3D estimate, but a number of observations insufficient for confirmation as a 3D model. Each 3D hypothesis is described in the same way as a 2D hypothesis, with two elements of additional information: 5) a 3D estimate of the feature's position, and 6) an estimate of the reconstruction error in 3D.
- *3D element* – is a confirmed 3D hypothesis. Each 3D model is described in exactly the same way as a 3D hypothesis.

4.2 The Matching Algorithm

The algorithm maintains a set of hypotheses, and evolves the state of each after insertion of each image.

After the features (lines and vertices) of a new image have been extracted, the algorithm tries to find confirming evidence for existing hypotheses among any newly observed 2D features. The algorithm first attempts to reduce reconstruction error for any 3D element for which a new observation is found. Next, hypotheses are processed in order of most to least evolved, beginning with 3D hypotheses, then 2D hypotheses, and finally unmatched 2D vertices. For each, confirmatory evidence is sought among any newly identified features.

For every existing element/hypothesis/vertex:

1. For every new vertex, we project a ray from the new camera position through the new vertex.
2. For 3D elements and 3D hypotheses, we find the shortest distance between the ray and the estimated 3D position of the element/hypothesis. If this distance is sufficiently small, we check to see if at least two incident edges of the element/hypothesis match edges incident to the new vertex.

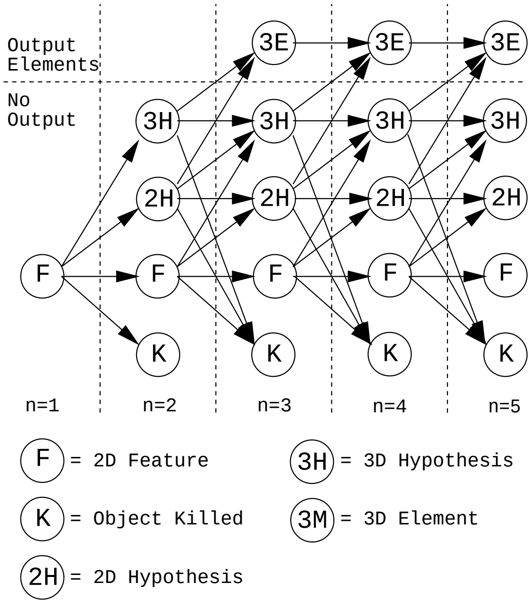

Figure 3: State evolution diagram

3. For 2D hypotheses and unmatched vertices, we find the residual resulting from intersecting this ray with rays from all matched vertices in the 2D hypothesis, or the single ray of the unmatched vertex. If this residual is sufficiently small, we check for matching adjacent edges as in Step 2.
4. We link the current model/hypothesis/vertex with the new vertex that has the best score.

Each new vertex can be matched with more than one hypothesized object. Thus, a spurious match will not affect other objects in the system. After each new match is identified, we test for these possible state transitions (Figure 3):

- 3D hypothesis → 3D element
  if the number of observations is sufficient.
- 2D hypothesis → 3D element
  if the baseline is long enough and the number of observations is sufficient.

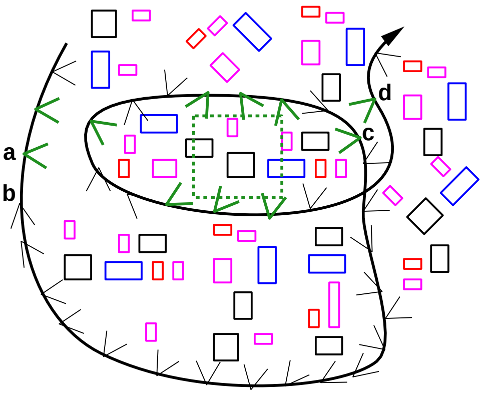

Figure 4: Images are associated spatially, not temporally.

- 2D hypothesis → 3D hypothesis if the baseline is long enough.
- 2D feature → 3D hypothesis if the baseline is long enough.
- 2D feature → 2D hypothesis if the baseline is not long enough.

If a hypothesis lingers longer than permitted without confirmatory evidence, it is "killed" or deleted from the set of active hypotheses.

5 Image Insertion

Above, we specified the processing to be done for each newly inserted image. Rather than insert the images in temporal order (the order in which they were acquired), we process images in groups according to whether they are suspected to have observed the same region of absolute 3D space (Figure 4). That is, given a set of images annotated with estimates of 6-DOF pose, we fix our attention on a region of 3D space (the dashed box in Figure 4), then identify those images possibly containing observations of this region from a distance less than some absolute threshold (typically, one hundred meters). In the figure, this set of images is represented by bold wedges. These images are inserted in arbitary order and processed as described above, producing a stateful set of feature hypotheses. The region of interest is then moved; any 3D elements no longer in the region of interest are output, and the set of relevant images is coherently updated to contain observations of the new region of interest.

6 Conclusion

In this paper, we describe a method for matching images acquired from arbitrary camera positions. Rather than processing images in temporal order, we process images by grouping them according the 3D regions they observe. The method operates by hypothesizing 3D features, then seeking confirmatory evidence for these features in successively inserted images. This incremental approach seeks to evolve feature hypotheses by amassing a sufficiently large number of observations which agree on a feature's position to within a sufficiently small tolerance or reconstruction error.

References

[Baker and Bolles, 1989] H.H. Baker and R.C. Bolles. Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface. *International Journal of Computer Vision*, 3:33-49, 1989.

[Bedekar and Haralick, 1996] A.S. Bedekar and R.M. Haralick. Finding Corresponding Points Based on Bayesian Triangulation *Proc. IEEE Computer Vision and Pattern Recognition*, San Francisco, CA, 1996, pp61-66.

[Collins, 1996] R.T. Collins. A Space-Sweep Approach to True Multi-Image Matching *Proc. IEEE Computer Vision and Pattern Recognition*, San Francisco, CA, 1996, pp358-363.

[Grimson, 1981] W.E.L. Grimson. A Computer Implementation of a Theory of Human Stereo Vision. *Phil. Trans. Royal Soc. London*, B292:217-253, 1981.

[Herman and Kanade, 1986] M. Herman and T. Kanade. Incremental Reconstruction of 3D Scenes from Multiple Complex Images. *Artificial Intelligence*, 30(3):289-341.

[Horaud and Skordas, 1989] R. Horaud and T. Skordas. Stereo Correspondence Through

Feature Groupings and Maximal Cliques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(11):1168-1180, 1989.

[Lim and Binford, 1988] H. S. Lim and T. O. Binford. Structural Correspondence in Stereo Vision. *Proc. DARPA Image Understanding Workshop.* Cambridge, MA, 1988, pp794-808.

[Terzopoulos, 1983] D. Terzopoulos. Multilevel Computational Processes for Visual Surface Reconstruction. *Computer Vision, Graphics, Image Processing.* 24:52-96, 1983.