ARBITRARY VIEW GENERATION FOR THREE-DIMENSIONAL SCENES FROM UNCALIBRATED VIDEO CAMERAS

Nelson L. Chang and Avideh Zakhor

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720 USA
e-mail: nlachang@eecs.Berkeley.EDU, avz@eecs.Berkeley.EDU

ABSTRACT

This paper focuses on the representation and arbitrary view generation of three dimensional (3-D) scenes. In contrast to existing methods that construct a full 3-D model or those that exploit geometric invariants, our representation consists of dense depth maps at several preselected viewpoints from an image sequence. Furthermore, instead of using multiple calibrated stationary cameras or range data, we derive our depth maps from image sequences captured by an uncalibrated camera. We propose an adaptive matching algorithm which assigns various confidence levels to different regions. Nonuniform bicubic spline interpolation is then used to fill in low confidence regions in the depth maps. Once the depth maps are computed at preselected viewpoints, the intensity and depth at these locations are used to reconstruct arbitrary views of the 3-D scene. Experimental results are presented to verify our approach.

1. INTRODUCTION

In light of recent advances in technology, virtual environments have become an important tool in engineering, design, manufacturing and many other areas. Especially important to the development of this growing field is the problem of Arbitrary View Generation (AVG) in which an intermediate view of a three dimensional (3-D) scene is interpolated from its neighboring views. Existing work in this area can be classified into three classes: in the first class, a full 3-D model of the scene is constructed by volumetric intersection and then reprojected in order to generate the desired view [1]. The main difficulty with this approach is that of registering and combining the 2-D information to generate a full 3-D model. In the second class, views are generated by exploiting certain invariants in the geometry of the problem [2]. This approach however does not correctly reconstruct points that become deoccluded.

The third class of AVG algorithms attempts to deal with occluded/deoccluded regions in the scene better than the second class while not resorting to a full 3-D representation of the first class. Generally, a set of $2\frac{1}{2}$-D surfaces is first estimated and then combined to generate the desired view. For example, Chen and Williams [3] measure range and camera transformation to establish pixel correspondence and then apply morphing to interpolate intermediate views. Similarly, Skerjanc and Liu [4] compute depth with known camera positions in order to synthesize intermediate pictures.

This work was supported by an Air Force Laboratory Graduate Fellowship, PYI-NSF grant MIP-9057466, ONR young investigator award N00014-92-J-1732, and Sun Microsystems.

Our approach to AVG falls into this third category [5]. However, unlike existing techniques, we use a sequence of images captured by a hand held, uncalibrated camcorder. Uncalibrated cameras with unknown position are used to avoid the difficult and time-consuming step of calibration and therefore increase the flexibility of the image acquisition process. Our motivation for using a sequence of video images rather than a few still images is to improve the robustness of the depth estimation step. Wide availability of video cameras in today's research and commercial environment justifies their use in place of still cameras in many applications.

Our proposed approach consists of scanning a camcorder across several trajectories of the scene in order to generate image sequences to be used in constructing the depth maps. The idea is to estimate depth only at several prespecified locations, called "reference frames," by using their neighboring captured frames. Once the depth has been computed at reference frames, the neighboring intensity frames are discarded, and only the depth and intensity at reference frames are kept as a compact representation of the scene. This representation is then used to reconstruct arbitrary views located on or off the scanning trajectories.

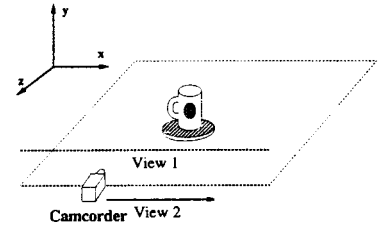

Figure 1: Experimental set up used to generate results.

In this paper, we consider a simple imaging geometry in which a camcorder is translated across the object on a line at multiple elevations, shown in Figure 1. The motivation for not choosing rotation, or a combination of rotation and translation motion, is the sensitivity of depth reconstruction to these classes of motion, especially when the motion parameters are unknown. In addition, it is well known that depth reconstruction can be more accurate when the camera translates across an object, rather than when it translates toward or away from it.

The outline of the paper is as follows. In Section 2, we discuss an adaptive approach to dense depth estimation. Section 3 describes the reconstruction algorithm. Results are presented in Section 4. The paper concludes with a discussion in Section 5.

2. COMPACT REPRESENTATION

Our overall approach in deriving the depth information at reference locations is to establish correspondence between the reference frame and each of its neighboring frames. Theoretically speaking, it is sufficient to establish correspondence with only one of the neighbors. In practice, however, it is advantageous to do so with a large number of neighboring frames in order to improve the accuracy of the resulting depth map. Note that once these neighboring frames are used in computing the depth at the reference frames, they are discarded in the reconstruction process; therefore, their use only affects the quality of the representation and not its compactness.

After correspondence between the reference frame and each of its neighbors has been achieved, the resulting depth maps at the reference frames are normalized and combined in order to form a depth map for the reference frame. In the remainder of this section, each step will be discussed in detail.

2.1. Depth Estimation

In the first step of the representation process, local dense depth maps are generated by matching the reference frame and each neighboring frame. Existing stereo matching techniques [6] cannot be used because they assume correspondence or known camera positions. Similarly, structure-from-motion algorithms [7] estimate the structure of only a small set of feature points in the scene.

We shall assume local perfect translation between every pair of images to reduce the depth estimation problem to a 1-D correspondence matching problem [8]. In this case, the epipolar lines of the two images are parallel with the scan lines of the image. For every point (i, j), the depth may be estimated as the inverse of disparity $d(i, j)$ given by

$$d(i,j) = \min_{m \in L} \left\{ \sum_{x=i-b/2}^{i+b/2} \sum_{y=j-b/2}^{j+b/2} |I_1(x,y) - I_2(x+m,y)|^2 \right\} \quad (1)$$

where L is the appropriate epipolar line.

There are some artifacts inherent both in the algorithm and the problem itself that induce incorrect disparities for certain regions. If the relative motion between two images is translational along the x axis, then an artifact known as aperture ambiguity occurs for horizontal lines. It arises because the block B used for matching is too small and does not include enough distinct features when matching. A second artifact occurs in regions of constant intensity where disparities are incorrectly matched because the block size is again too small. Other artifacts occur in occluded regions and near depth discontinuities; see [5] for more details.

It is straightforward to identify most of these artifacts and subsequently assign confidence levels to different regions in the scene. These confidence levels are important for locating the regions to ignore when combining multiple depth maps together. To detect aperture ambiguity, a gradient-based edge detector is used to locate the horizontal edges [5]. Points in the image near these edge pixels are marked as possibly spurious. To identify constant intensity regions, a small window is used to find regions where the intensity variance is lower than a prespecified threshold. A low variance suggests that the block consists of low texture and nearly constant intensity. Occluded regions consist of the unmapped points from matching two images in both directions. Performing the match in both directions also helps to validate the matches [5]. In the end, the scene will consist of low confidence regions marked according to the different artifacts: constant intensity, aperture ambiguity, occlusion, and inconsistencies in matching.

Since many real world scenes consist largely of low textured regions, the matching algorithm will produce a high percentage of low confidence regions due to constant intensity. To avoid too sparse a depth map, we attempt to improve estimates in these regions. We propose an adaptive matching approach whereby a small block size is used to match regions near boundaries and a larger block size is used to match constant intensity regions [5]. This overcomes the well-known tradeoff between good boundary localization with a small window and improved matching in low textured regions with a large window. The final result consists of fairly dense and reasonably accurate disparities.

2.2. Normalization of Initial Estimates

The depth maps from the previous stage need to be normalized so that they are all related by the same scaling factor. For this task, we propose to estimate the translation parameter between maps and scale by the reciprocal. The relationship between disparities $\Delta u_{m,i}$ and relative motion b_m may be derived [5] to get the linear least squares solution

$$\frac{b_m}{b_1} = \frac{\sum_{i=1}^{k} (\Delta u_{1,i})(\Delta u_{m,i})}{\sum_{i=1}^{k} (\Delta u_{1,i})^2} \quad (2)$$

where b_1 is assumed to be one. Then b_m is precisely the scaling factor α_m by which we need to adjust the m-th depth map. An iterative process is used to reduce the error $\|A\alpha_m - y\|_2$ to some desired amount where outlier points greater than a given error percentage are disregarded when computing α_m.

2.3. Combination of Multiple Depth Maps

Once all the depth maps have been normalized to a common scaling factor, they are combined to form a single depth map for a particular reference frame. For every point, an iterative procedure is used to analyze the statistics of the given data, throw out outliers, and reduce the data set to a more consistent one. Points outside the range median $\pm k\sigma$ are discarded. The remaining points are combined in a weighted average based on confidence levels [5]. Depth information from matching a vertically-related pair of images is also included in combination to overcome spurious estimates due to horizontal aperture ambiguity.

2.4. Cubic B-Spline Approximation

The depth map after the combination stage is fairly accurate in many regions. There are however a considerable number of low confidence regions. To fill in these regions and to make the map much denser while not sacrificing too much accuracy, nonuniform cubic B-splines are used. Every depth point in low confidence regions is interpolated by its neighboring high confidence depth vertices along the same row or column, depending on the variance of these vertices. The depth surface is treated as a tensor product, i.e. the product of 1-D functions, so the data may be processed first along one direction and then along the other which helps to simplify computations.

Once the depth map for each reference frame has undergone spline approximation, we are left with $2\frac{1}{2}$-D surface estimates at different locations around the scene. The

final step in the representation process is to estimate the relative camera motion between reference frames using an approach like [7]. Once the relative motion between all reference frames is known, a geometric relationship may be constructed among the different reference frames. This enables us to select the reference frames needed to use in the reconstruction stage.

In the end, the representation of the object consists of the intensity-depth pair at each reference location along with the relative motion among reference frames. Once these data have been derived, they may be stored in a database for later reconstruction.

3. RECONSTRUCTION OF VIEWS

Once we have generated the representation for a particular 3-D object, we may choose to reconstruct the view of the object at some specified viewpoint. Assume that the center of one reference frame coincides with the origin of the coordinate system and that the desired viewpoint is known with respect to this origin. The reconstruction algorithm consists of the following: First the appropriate reference frame(s) are chosen. Then initial estimates of the desired view are constructed by applying motion parameters to each reference frame. Finally, the estimates are combined into a single image, interpolating when necessary.

3.1. Selection of Appropriate Reference Frame(s)

Given the relative position and orientation of the desired view, it should be a straightforward task to determine which reference frames to use. One way of deciding is to include those frames with the smallest motion in norm relative to the view.

Another consideration is the number of reference frames. If the specified view is very close to one of the reference frames, then we may choose to use only that single frame. However, at least two reference frames are needed to properly reconstruct the desired view to reduce noise and to recover occluded regions in the scene.

3.2. Generation of View Estimates

The notion of applying motion parameters to a frame has been addressed in conventional computer vision literature [8]. Let (u_1, v_1) be the projection of a point in the scene onto the image plane. Suppose the frame of reference undergoes a rigid transformation (R,T) given by $R = [r_{i,j}]$ and $T = (\Delta x, \Delta y, \Delta z)'$ where both rotation R and translation T are in terms of the world coordinates. Then the new image coordinates are given by

$$u_2 = \frac{(r_{1,1}u_1 + r_{1,2}v_1 + r_{1,3})Z + \Delta x}{(r_{3,1}u_1 + r_{3,2}v_1 + r_{3,3})Z + \Delta z} \quad (3)$$

$$v_2 = \frac{(r_{2,1}u_1 + r_{2,2}v_1 + r_{2,3})Z + \Delta y}{(r_{3,1}u_1 + r_{3,2}v_1 + r_{3,3})Z + \Delta z} \quad (4)$$

where the focal length f is assumed to be 1.

The points of the reference frame arrays are considered not as discrete independent points, but rather as vertices of a deformable wire mesh [5] to overcome possible inconsistencies after transformation. Neighboring points in the reference frame are viewed as connected to one another. A view estimate is generated by applying equations (3) and (4) to the collection of points and examining not only the new coordinates of every point, but also the ordering in the mesh. In this manner, the ordering of points may be better preserved and inconsistencies of spurious background points appearing among foreground points in the transformed data are not as prevalent. Regions behind moving objects may become uncovered after view transformation. In this case, interpolation between consecutive points according to the mesh may be included.

3.3. Combination of Reconstructed Data

For each point, a small region around the point is considered. Outliers in the depth domain are thrown out until the variance in the intensity of the points in the region is approximately uniform. The motivation is that the points are expected to possess similar depth and intensity in the same neighborhood. This step further rules out discrepancies among the data.

During reconstruction, "holes" may be created when no points fall within a region. This condition arises because of uncovered regions in the scene, i.e. deoccluded regions, and because of sparse depth information. Generally, introducing more reference frames helps to reduce the size of these holes. For the remaining holes, the region around each point is grown until a sufficient number of points exists within the region [5].

4. RESULTS

We shall now examine some results using the techniques described above. The object of interest is a mug placed atop a stool. A CCD camcorder is moved by hand to follow trajectories at two different elevations to generate an image sequence for each trajectory, similar to the set up drawn in Figure 1. Each frame is 640×480 pixels large and consists of intensity only. We attempted to make the motion roughly translational along the x axis to demonstrate that neither a calibrated set up nor a track is needed. Moreover, no special lighting was used to film the scene; specularities of the stool and the lid of the mug are very apparent in the images.

Figure 2: Example of reference frame (intensity).

For the first set of results, the desired view is roughly halfway between two reference frames along the same horizontal trajectory; one reference frame is shown in Figure 2. This desired view is perhaps the one most prone to errors due to the large occluded regions. Note that there is roughly a maximum of 120 pixel disparity between the two reference frames.

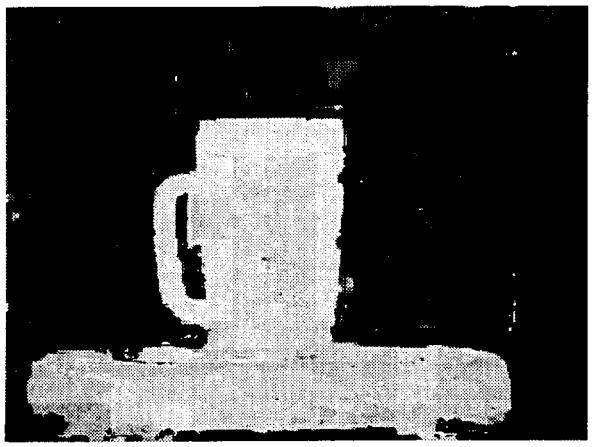

Figure 3: Example of reference frame (depth) filled in by splines.

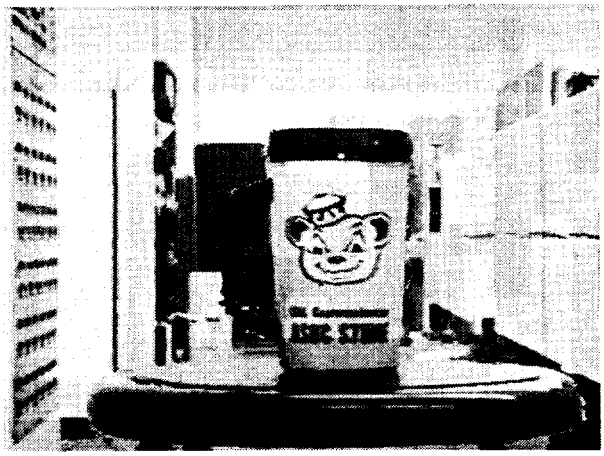

Figure 4: Reconstructed view along horizontal trajectory.

Figure 3 shows the corresponding depth map obtained by using the proposed matching algorithm. The mug and stool are estimated well and do not contain many spurious depths. There is a gradual change in depth as expected for a hallway scene. Artifacts are prevalent in the top left portion of the stool; this is primarily due to the specularities of the surface. Also, there are problems in recovering the handle of the mug accurately mainly because intensity-based matching schemes perform poorly for background regions that can be seen through foreground regions.

The reconstructed view is shown in Figure 4. The image quality is good for the most part. The horizontal edges, e.g. top of the door, top of the mug, specularities in front of the stool, and the drawers, have been reconstructed quite well. The proposed algorithms take care of problems in occluded regions: There are only a few errors to the right of the mug and near the mug handle.

To generate a view not originally scanned by the camcorder, two frames from different elevations are chosen as reference frames. The desired view is roughly the midpoint on the vertical trajectory relating the two views.

The reconstructed view in Figure 5 is a reasonable estimate of the desired view. The most noticeable artifact occurs around the upper left portion of the stool caused by specularities that result in spurious depths. This problem may be overcome by using a larger number of frames to form the combined depth map; we are currently investigating this issue.

Figure 5: Reconstructed view along vertical trajectory.

5. DISCUSSION

We have proposed an approach for representing and reconstructing stationary 3-D objects. The results in the previous section seem to indicate that this approach is very worthwhile. Future work in this area includes considering more general imaging geometry and examining the optimum positions of reference frames required for a given scene. The area of arbitrary view generation and its application to virtual environments seems very fertile and this research serves as a good starting point.

6. REFERENCES

[1] C. H. Chien and J. K. Aggarwal, "Identification of 3D objects from multiple silhouettes using quadtrees/octrees," *Computer Vision, Graphics and Image Processing*, vol. 36, no. 2–3, pp. 256–273, Nov.-Dec. 1986.

[2] A. Shashua, "Projective structure from two uncalibrated images: Structure from motion and recognition," Tech. Rep. 1363, MIT AI Laboratory, Sept. 1992.

[3] S. E. Chen and L. Williams, "View interpolation for image synthesis," in *Proceedings of SIGGRAPH*, New York, 1–6 Aug. 1993.

[4] R. Skerjanc and J. Liu, "A three camera approach for calculating disparity and synthesizing intermediate pictures," *Signal Processing: Image Communication*, vol. 4, pp. 55–64, 1991.

[5] N. L. Chang, "View reconstruction from uncalibrated cameras for three-dimensional scenes," Master's thesis, University of California at Berkeley, 1994.

[6] U. R. Dhond and J. K. Aggarwal, "Structure from stereo—a review," *IEEE Trans. Sys. Man Cyber.*, vol. 19, no. 6, pp. 1489–1509, 1989.

[7] R. Szeliski and S. B. Kang, "Recovering 3D shape and motion from image streams using non-linear least squares," Tech. Rep. CRL 93/3, Digital Equipment Corporation: Cambridge Research Lab, March 1993.

[8] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1991.