Quasi-Dense Motion Stereo for 3D View Morphing

David Jelinek Camillo J. Taylor
GRASP Laboratory
Computer and Information Science Dept.
University of Pennsylvania
3401 Walnut St, RM 335C
Philadelphia, PA 19104-6229
Phone: (215) 898–0376
Fax: (215) 573–2048
email: {davidj2,cjtaylor}@grasp.cis.upenn.edu

Abstract

This paper presents a novel approach to image based rendering. The problem of constructing a new view based on a set of snapshots taken from known locations is formulated as a morphing task. Appropriate morphing functions are constructed by considering a set of feature points in the scene which correspond to intensity discontinuities in the original pictures. The locations of these features in the scene are recovered automatically from the image data by employing epipolar plane image analysis. One of the main advantages of the proposed method is that it correctly reproduces parallax in the most perceptually salient regions of the image, the areas corresponding to intensity edges. Results obtained by applying the proposed technique to actual image data sets are presented.

1 Introduction

The goal of most image based rendering systems can be stated as follows: given a set of pictures taken from various vantage points synthesize the image that would be obtained from a novel viewpoint. In this paper, the problem of producing a novel view is formulated as an image morphing task and the central objective is to automatically construct morphing functions that map the original photographs onto the novel view. These morphing functions are constructed by considering the motion of a set of features in the scene that correspond to intensity discontinuities in the original imagery. The 3D locations of these features are obtained automatically from the input image sequence by employing the Epipolar Plane Image (EPI) analysis techniques proposed by Bolles, Baker, and Marimont [1]. One of the main advantages of the proposed method is that it correctly reproduces parallax in the most perceptually salient regions of the image, the areas corresponding to intensity edges.

For the purposes of this discussion, previous approaches to the image based rendering problem can be divided into three categories. The first set of approaches are based on the plenoptic sampling approach described by Levoy and Hanrahan [7] and Gortler et al [3] . In these schemes novel views are reproduced by sampling the appropriate rays from the input images. Shum and He [17] describe an interesting and effective approach for extending these

techniques to immersive environments using a sampling system based on cocentric mosaics. The method proposed in this paper differs from the techniques in this category by attacking the view generation problem as a morphing task rather than a plenoptic sampling problem.

The second category of approaches consists of techniques which proceed by constructing a detailed geometric model of the scene in the form of per pixel depth information for every image in the data set. Laveau and Faugeras [6], Pollefeys et al [13] , and Werner et al [19, 4] all propose stereo based techniques for recovering the required depth or disparity maps from the input image data. Other researchers assume that the depth maps can be obtained from auxiliary range sensors [14, 12]. Once these depth maps have been obtained it is a relatively straightforward matter to produce a novel view of the scene by computing where each of the pixels in the original views will appear in the novel image. The Layered Depth Image representation proposed by Shade et al. [16] provides a particularly efficient method for rendering data sets of this form. These authors also describes a method for determining the relative depth of points in the scene by estimating the motion of various layers in the input imagery.

The problem with recovering dense depth maps from image sequences is that there are two important situations where it is exceedingly difficult for traditional correlation-based stereo algorithms to produce accurate depth estimates. The first situation corresponds to texture free regions in the scene, such as blank walls, which do not produce a sufficiently distinctive correlation signature. Occluding edges in the scene can also cause significant difficulties since the regions in the image surrounding such an edge will often contain half occluded regions which cannot be adequately matched between frames by the correlation metric. The proposed method overcomes these problems by employing an epipolar plane image analysis to recover the positions of the feature points rather than a correlation based approach and by using an interpolation scheme which produces acceptable results in texture free regions.

The third category of approaches consists of techniques based on the Image Morphing approach described by Chen and Williams [2]. Seitz and Dyer [15] propose a technique for producing physically correct images by interpolating between a given pair of views. Lhuiller and Quan [8, 9] describe a view morphing technique which also seeks to produce interpolated views which correctly reproduce the motion of salient points in the scene. They describe a scheme for triangulating the input image in such a way as to respect intensity discontinuities. These papers demonstrate that it is possible to produce compelling interpolated images from a relatively sparse set of correspondences. Both of these approaches deliberately avoid the problem of estimating the actual 3D locations of the feature points that are used as correspondences. This means that the techniques can be applied to uncalibrated imagery but it also limits the systems to producing views that lie along the straight line connecting the two original images. The technique proposed in this paper eliminates this restriction by estimating the actual 3D locations of the observed feature points. This allows the system to predict where the features will appear in any viewpoint.

The rest of this paper is organized as follows: Section 2 describes the implementation of the proposed approach

in more detail. Section 3 presents experimental results. Conclusions and further work are discussed in section 4.

2 Implementation

The proposed method can be divided into four stages, each of which will be described in turn below. The first stage involves capturing a sequence of images taken at regular intervals as the camera undergoes a translation along an axis perpendicular to its optical axis. The motion control rig pictured in Figure 1 has been constructed for this purpose. It provides a means for precisely positioning the camera along a single translational axis under computer control. In this case the motion of the camera is along an axis parallel to the rows of the image which simplifies the resulting image analysis problem. The setup used here corresponds to the slider stereo configuration described by Moravec [10] and by Baker, Bolles, and Marimont [1]. The intrinsic parameters for the camera (the focal lengths and the position of the principal point in the image) are obtained using the calibration procedure proposed by Zhang [21].

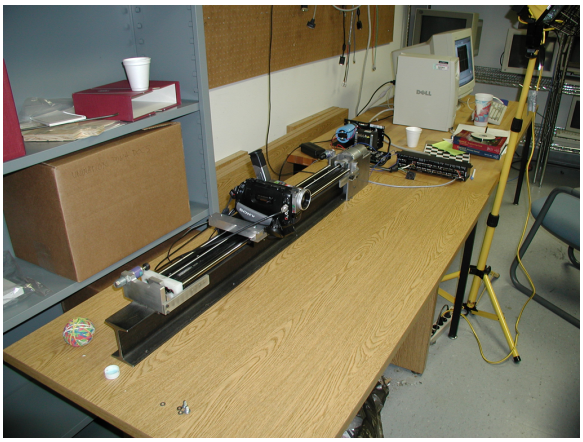


Figure 1: The slider stereo rig used to acquire imagery.

Once the images have been acquired, the second stage of processing involves rearranging the pixel data to

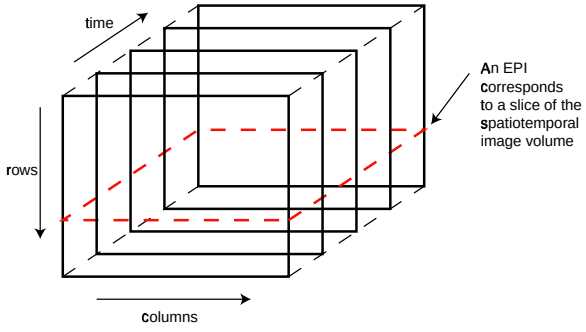


Figure 2: An epipolar plane image is formed by taking a slice of the spatiotemporal image volume formed by stacking all of the images in the data set together

Figure 3: A typical epipolar plane image obtained from the imagery is shown in.

Figure 4: The results of applying the edge enhancement procedure to the epipolar plane image shown in in Figure 3.

produce an epipolar plane image for each row in the image. As shown in Figure 2, an EPI simply corresponds to a slice of the spatiotemporal volume formed by stacking the acquired images together. An example of a typical EPI is shown in Figure 3. Note the characteristic banded structure of this image which can be explained by noting that feature points in the scene will correspond to straight lines in the epipolar plane imagery.

Each of these epipolar plane images is subjected to an analysis which seeks to extracts these straight line features. This analysis is based on the techniques described by Baker et al [1] and Yamamoto [20]. A series of filters designed to enhance edges at various orientations is run over each epipolar plane image and the resulting edge elements are linked together to form straight line segments. Line segments that are deemed long enough and straight enough are interpreted as 3D feature points. Typical results obtained by inoking this procedure are shown in Figure 4. The 3D location of the corresponding feature point in the scene is computed from the position and slope of the extracted line segments. The end result of the EPI analysis is a set of 3D points corresponding

precisely to intensity discontinuities in the image.

Note that unlike correlation based approaches, this method for recovering the depth of feature points in the imagery does *not* assume that corresponding points in the images will be strongly correlated. It simply exploits the fact that feature points in the scene correspond straight lines in the EPI. The resulting line fitting problem is heavily overdetermined which serves to improve the accuracy and robustness of the method. This means that the technique produces accurate estimates for the depth of occluding edges and other features that are problematic for correlation based methods. This is important because occluding edges are particularly salient features in the scene and it is important to correctly predict where these features will appear in the novel view.

In the fourth stage, the system applies a Delaunay triangulation to each of the images in the data set. The vertices of the triangles correspond to the projections of the feature points that are visible in that image. This construction has the agreeable property that the triangles tesselate the image without crossing image intensity discontinuities which means that the resulting triangles have a strong tendency to correspond to facets of actual surfaces in the scene. Even triangles that do not correspond to surface facets encompass homogenous regions in the image, such as blank walls, which can be morphed to the novel view in a convincing manner. The scheme also has an adaptive sampling property in that regions of the image that have a lot of intensity discontinuties are tessalated quite finely while other regions of the images that are less interesting are covered with fewer facets. This is appropriate from the perspective of view synthesis since we can usually get away with less refined interpolation in portions of the image that correspond to untextured areas. Figure 5 shows the results of a typical triangulation.

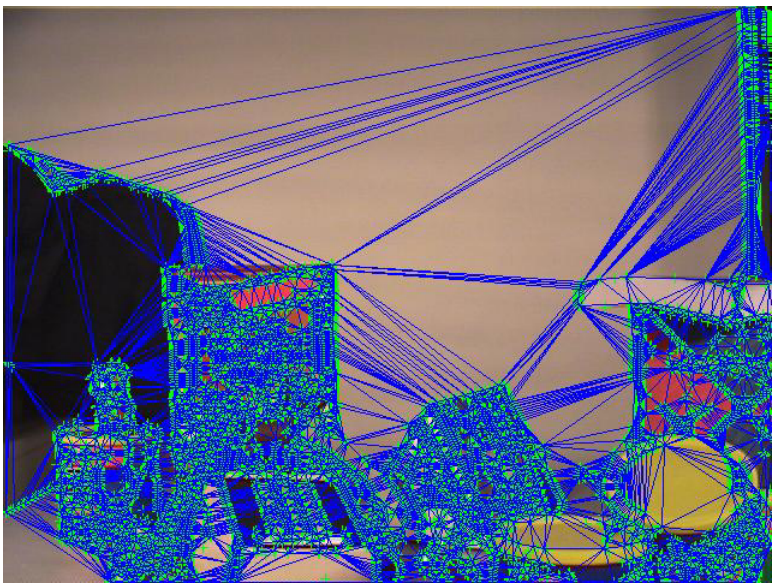


Figure 5: Results obtained by applying the Delaunay triangulation procedure to the features recovered by the EPI analysis procedure.

At the end of this stage, the system has constructed a set of triangular facets for each of the input images and the locations of the vertices of these triangles in the scene are known. At this point the system can generate novel

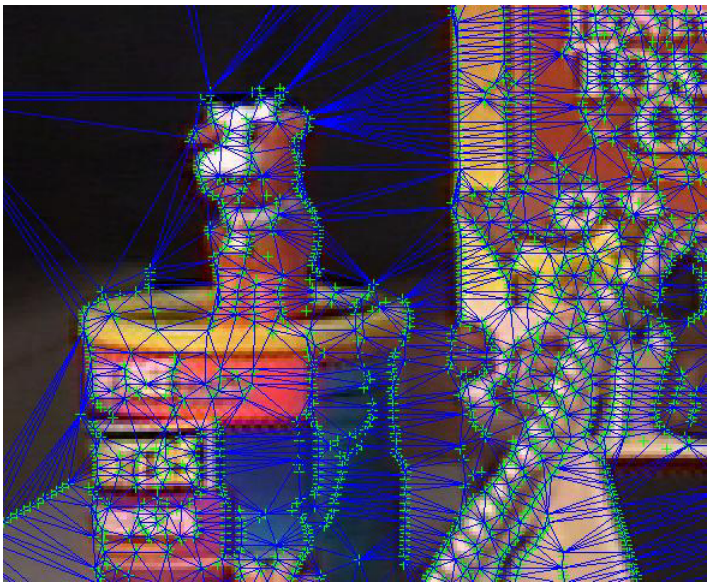


Figure 6: A closeup view of a section of the triangulated image. Notice that the triangulation produces facets corresponding to homogenous regions of the image that can be effectively morphed to the novel view. The majority of these facets correspond reasonably well to actual surfaces in the scene.

views of the scene by supplying these triangles to a standard rendering pipeline and using the original images as texture maps. The rendering system correctly accounts for the parallax induced as a result of the motion of the virtual viewpoint and hidden surface removal reproduces the majority of the occlusion and disocclusion events that would be observed as the camera moves.

Note that in this framework any of the original images could be morphed to the novel viewpoint. In the current implementation the simple expedient of choosing the closest original viewpoint as a basis for morphing is employed with the idea that minimizing the difference in position between novel and original viewpoints will minimize the errors introduced by the viewpoint morphing operation.

3 Experimental Results

The proposed technique has been implemented and applied to actual image sequences. Figure 7 shows two of the images taken from a sequence of 500. With this setup, it is possible to produce plausible novel renderings of the scene from all viewpoints contained within a box approximately 15 cm by 15 cm by 75 cm centered around the original linear camera trajectory. Figure 8 shows a selection of novel images produced using the technique described in this paper.

Figures 9 and 10, 11 and 12, and 13 and14 show samples of the original and novel imagery produced on another sequence.

Using an OpenGL implementation on a standard PC *without* the benefit of hardware acceleration the system was able to render novel views involving thousands of triangular facets in under a second. It is expected that

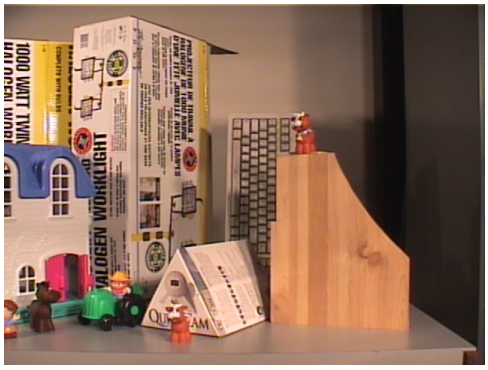
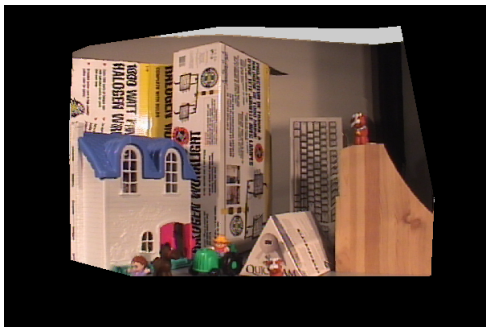
a.  b. 

Figure 7: Two of the input images taken from a sequence of 500.

a.  b. 

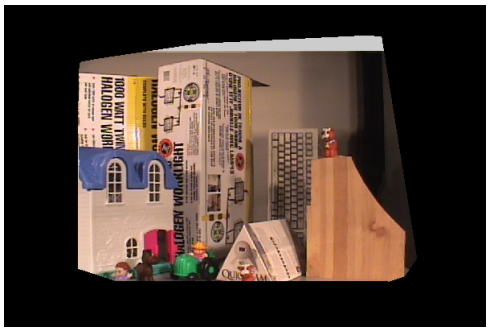
c.  d. 

Figure 8: Examples of the novel views obtained by employing the proposed technique.

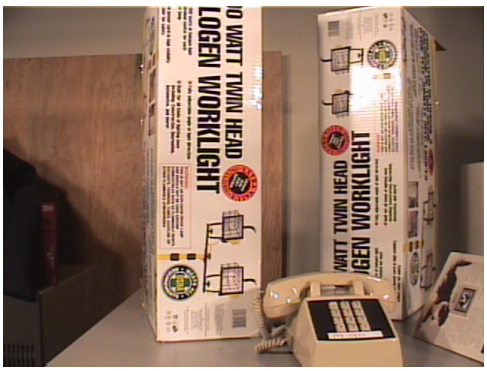
a.  b. 

Figure 9: Two of the input images taken from a sequence of 500.

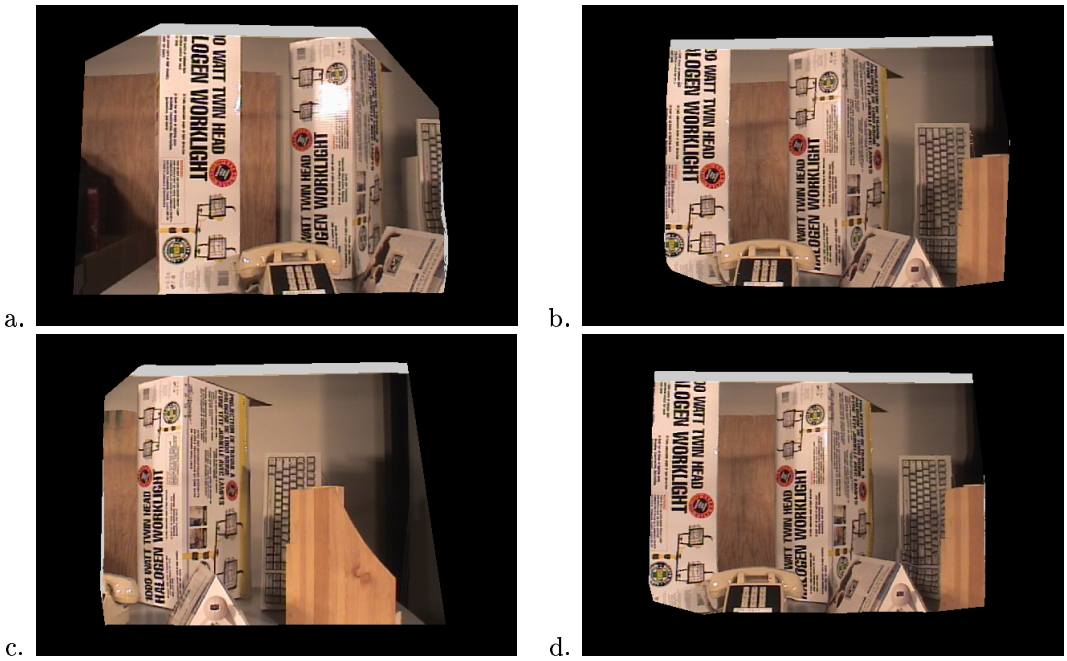


Figure 10: Examples of the novel views obtained by employing the proposed technique.

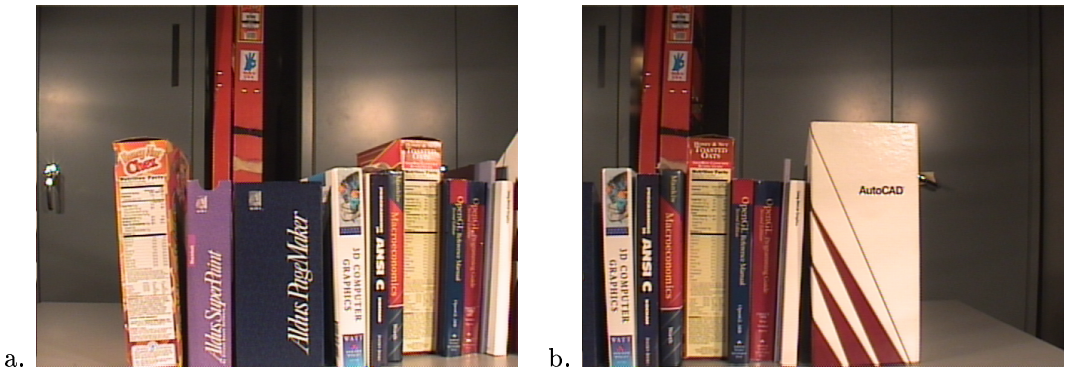


Figure 11: Two of the input images taken from a sequence of 500.

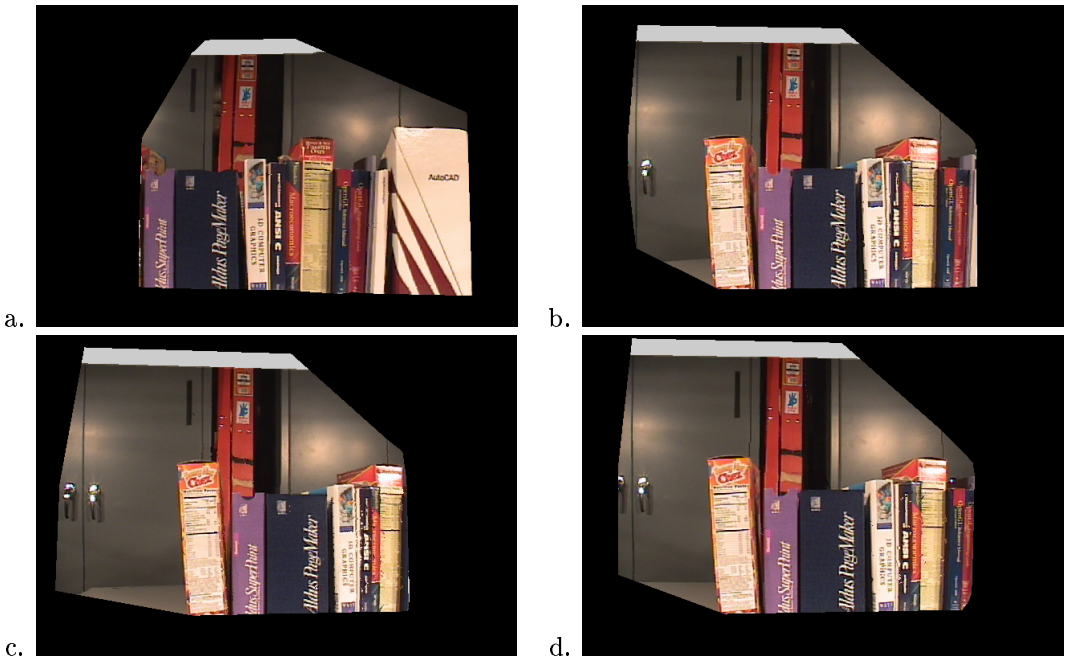


Figure 12: Examples of the novel views obtained by employing the proposed technique.

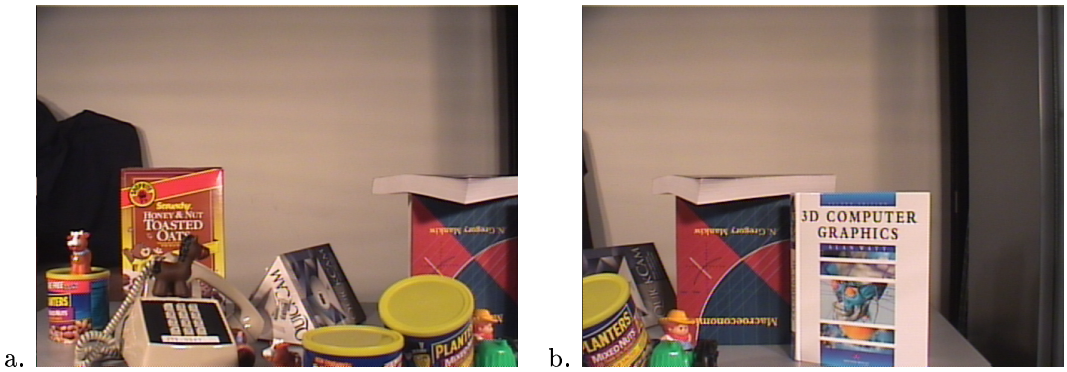


Figure 13: Two of the input images taken from a sequence of 500.

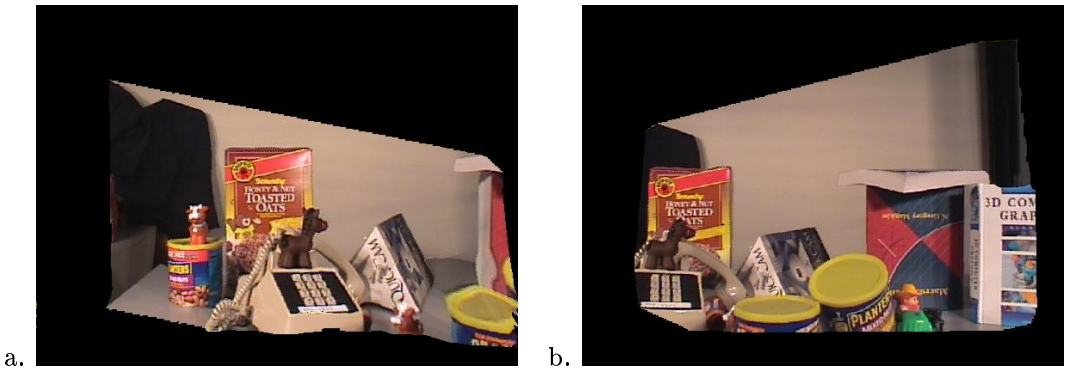


Figure 14: Examples of the novel views obtained by employing the proposed technique.

with appropriate hardware support for texture mapping it would be possible to generate novel images at frame rate which would allow a user to modify his or her viewpoint interactively.

(MPEG movies which more fully demonstrate the capabilities of the proposed method can be found at: `http://www.cis.upenn.edu/~davidj2/mpeg.html`.)

4 Conclusions

This paper describes an approach to image based rendering which proceeds from the idea that novel views of the scene can be constructed by morphing the original images to the novel viewpoint. These morphing functions are constructed by considering the motion of a set of point features corresponding to intensity discontinuities in the original imagery. The locations of these feature points are recovered automatically from the image data by applying an epipolar plane image analysis technique. A Delaunay triangulation is used to produce a plausible set of triangular facets which are morphed from the original photographs to the novel image.

This rendering scheme hinges on the observation that human observers tend to be very sensitive to errors in reproducing the motion of edges in the images but more tolerant of errors in texture free areas. The scheme exploits this property by focusing its efforts on correctly reproducing the motion of these intensity discontinuities in the image. The result is a scheme that could be used to produce novel views of scenes without requiring dense depth maps.

The proposed scheme offers some advantages over previously proposed techniques. It does not require the large, sophisticated image data structures used by plenoptic sampling approaches, it does not require a range scanning device, and it can be accelerated using standard graphics hardware.

4.1 Future Work

There are several outstanding issues which will be addressed in future work. One of the most significant problems with the proposed approach is the fact that the scheme for locating feature points in the scene fails when the features of interest correspond to edges which are parallel to the axis of translation of the slider stereo rig. In order to overcome this problem one must consider sets of images that are not taken from collinear vantage points. This will require the development of new techniques for recovering correspondences since the standard epipolar plane image analysis methods cannot be directly applied to non-linear camera trajectories.

The current implementation essentially combines ideas from plenoptic modeling and view morphing by selecting the closest original view as the basis for the morphing procedure. It is certainly possible to imagine situations where it might be advantageous to combine image information obtained from several viewpoints to produce the final rendering. Methods for weighting the contributions from all of the triangular facets in the data set based on considerations such as the disparity between the original and novel viewpoints and the extent of the foreshortening induced by the morphing operation will be explored.

It would also be interesting to apply the technique to imagery acquired with omnidirectional camera systems [11, 18, 5]. In this case, the system could be used to interactively explore an immersive environment by morphing between omnidirectional snapshots taken from known viewpoints.

References

[1] R.C. Bolles, H.H. Baker, and D.H. Marimont. Epipolar-plane image anlysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[2] S. E. Chen and L. Williams. View interpolation from image synthesis. In *SIGGRAPH*, pages 279–288, August 1993.

[3] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael Cohen. The lumigraph. In *Proceedings of SIGGRAPH 96. In Computer Graphics Proceedings, Annual Conference Series*, pages 31–43, New Orleans, LA, August 4-9 1996. ACM SIGGRAPH.

[4] V. Hlavac, A. Leonardis, and T. Werner. Automatic selection of reference views for image-based scene representations. In *European Conference on Computer Vision*, pages 526–535, 1996.

[5] Hiroshi Kawasaki, Katsushi Ikeuchi, and Masao Sakauchi. Spatio-temporal analysis of omni image. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 577–584, 2000.

[6] S. Laveau and O.D Faugeras. 3-d scene representation as a collection of images. In *International Conference on Pattern Recognition*, pages 689–691, 1994.

[7] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of SIGGRAPH 96. In Computer Graphics Proceedings, Annual Conference Series*, pages 31–43, New Orleans, LA, August 4-9 1996. ACM SIGGRAPH.

[8] Maxime Lhuiller and Long Quan. Image interpolation by joint view triangulation. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 139–145, 1999.

[9] Maxime Lhuiller and Long Quan. Edge-constrained joint view triangulation for image interpolation. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 218–224, 2000.

[10] Hans P. Moravec. The Stanford cart and the CMU rover. *Proceedings of the IEEE*, 71(7), July 1983.

[11] Shree Nayar. Catadioptric omnidirectional camera. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 1997.

[12] Ko Nishino, Yoichi Sato, and Katsui Ikeuchi. Eigen-texture method: Appearance compression based on 3d model. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 618–624, 1999.

[13] M. Pollefeys, L Van Gool, and M. Proesmans. Euclidean 3d reconstruction from image sequences with variable focal lengths. In *European Conference on Computer Vision*, pages 31–42, 1996.

[14] Y. Sato, M.D. Wheeler, and K. Ikeuchi. Object shape and reflectance modeling from observation. In *Proceedings of SIGGRAPH 97. In Computer Graphics Proceedings, Annual Conference Series*, pages 379–387. ACM SIGGRAPH, August 1997.

[15] Steven Seitz and Charles R. Dyer. View morphing. In *Proceedings of SIGGRAPH 96. In Computer Graphics Proceedings, Annual Conference Series*, pages 31–43, New Orleans, LA, August 4-9 1996. ACM SIGGRAPH.

[16] Jonathan Shade, Steven Gortler, Li wei He, and Richard Szeliski. Layered depth images. In *Proceedings of SIGGRAPH 98. In Computer Graphics Proceedings, Annual Conference Series*, pages 231–242. ACM SIGGRAPH, August 1998.

[17] Heung-Yeung Shum and Li-Wei He. Rendering with concentric mosaics. In *SIGGRAPH*, pages 299–306, August 1999.

[18] Takuji Takahashi, Hiroshi Kawasaki, Katsushi Ikeuchi, and Masao Sakauchi. Arbitrary view position and direction rendering for large-scale scenes. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 296–303, 2000.

[19] T. Werner, R.D. Hersch, and V. Hlavac. Rendering real-world objects using view interpolation. In *International Conference on Computer Vision*, pages 957–962, 1995.

[20] Masanobu Yamamoto. *The Image Sequence Analysis of Three-Dimensional Dynamic Scenes.* PhD thesis, Electrotechnical Laboratory, Agency of Industrial Science and Technology, May 1988.

[21] Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *International Conference on Computer Vision*, pages 666–673, 1999.