# VideoPlus

Camillo J. Taylor
GRASP Laboratory, CIS Department
University of Pennsylvania
3401 Walnut Street, Rm 335C
Philadelphia, PA, 19104-6229
email: cjtaylor@central.cis.upenn.edu
Phone: (215) 898 0376
Fax: (215) 573 2048

February 23, 2000

## Abstract

This paper describes an approach to capturing the appearance and structure of immersive environments based on the video imagery obtained with an omnidirectional camera system. The scheme proceeds by recovering the 3D positions of a set of point and line features in the world from image correspondences in a small set of key frames in the image sequence. Once the locations of these features have been recovered the position of the camera during every frame in the sequence can be determined by using these recovered features as fiducials and estimating camera pose based on the location of corresponding image features in each frame. The end result of the procedure is an omnidirectional video sequence where every frame is augmented with its pose with respect to an absolute reference frame and a 3D model of the environment composed of point and line features in the scene.

By augmenting the video clip with pose information we provide the viewer with the ability to navigate the image sequence in new and interesting ways. More specifically the user can use the pose information to travel through the video sequence with a trajectory different from the one taken by the original camera operator. This freedom presents the end user with an opportunity to immerse themselves within a remote environment and to control what they see.

**Keywords:** Reconstruction, Omnidirectional Video, Pose Estimation

## 1 Introduction

This paper describes an approach to capturing the appearance and structure of immersive environments based on the video imagery obtained with an omnidirectional camera system such as the one proposed by Nayar [13]. The scheme proceeds by recovering the 3D positions of a set of point and line features in the world from image correspondences in a small set of key frames in the image sequence. Once the locations of these features have been recovered the position of the camera during every frame in the sequence can be determined by using these recovered features as fiducials and estimating camera pose based on the location of corresponding image features in each frame. The end result of the procedure is an omnidirectional video sequence where every frame is augmented with its pose with respect to an absolute reference frame and a 3D model of the environment composed of point and line features in the scene.

One area of application for the proposed reconstruction techniques is in the field of robotics since it allows us to construct 3D models of remote environments based on the video imagery acquired from a mobile. For example, the model of an indoor environment shown in Figure 6 was constructed from the video imagery acquired by the mobile robot shown in Figure 4 as it roamed through the scene.

Such a model would allow the remote operator to visualize the robots operating environment. It could also be used as the basis for an advanced human robot interface where the robot can be tasked by pointing to a location on the map and instructing the robot to move to that position. The robot would be able to automatically plan and execute a collision free path to the destination based on the information contained in the map.

Another interesting application of the proposed technique is in the field of virtual tourism. By augmenting the video clip with pose information we provide the viewer with the ability to navigate the image sequence in new and interesting ways. More specifically the user can use the pose information to travel through the video sequence with a trajectory different from the one taken by the original camera operator. This freedom presents the end user with an opportunity to immerse themselves within a remote environment and to control what they see.

The rest of this paper is arranged as follows Section 2 describes the process whereby the 3D locations of the model features and the locations of the cameras are estimated from image measurements. Results obtained by applying these techniques to actual video sequences are also presented in this section. Section 3 discusses the relationship between this research and previously published work. Section 4 briefly describes future directions of this research and section 5 presents some of the conclusions that have been drawn so far.

## 2 Reconstruction

This section describes how the 3D structure of the scene and the locations of the camera positions are recovered from image correspondences in the video sequence. The basic approach is similar in spirit to the reconstruction schemes described in [17] and [4]. The reconstruction problem is posed as an optimization problem where the goal is to minimize an objective function which indicates the discrepancy between the predicted image features and the observed image features as a function of the model parameters and the camera locations.

In order to carry out this procedure it is important to understand the relationship between the locations of features in the world and the coordinates of the corresponding image features in the omnidirectional imagery. The catadioptric camera system proposed by Nayar [13] consists of a parabolic mirror imaged by an orthographic lens. With this imaging model there is an effective single point of projection located at the focus of the parabola as shown in Figure 1.

Given a point with coordinates (u, v) in the omnidirectional image we can construct a vector, $\upsilon$, which is aligned with the ray connecting the imaged point and the center of projection of the camera system.

$$\upsilon = \begin{pmatrix} s_x(\mathrm{u}-\mathrm{c_x}) \\ s_y(\mathrm{v}-\mathrm{c_y}) \\ (s_x(\mathrm{u}-\mathrm{c_x}))^2 + (\mathrm{s_y}(\mathrm{v}-\mathrm{c_y}))^2 - 1 \end{pmatrix} \quad (1)$$

This vector is expressed in terms of a coordinate frame of reference with its origin at the center of projection and with the z-axis aligned with the optical axis of the device as shown in Figure 1.

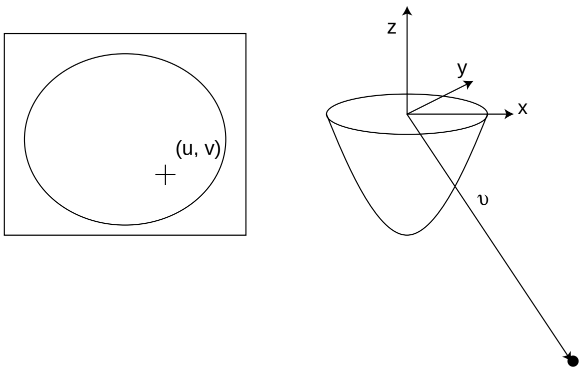

Figure 1: The relationship between a point feature in the omnidirectional image and the ray between the center of projection and the imaged point.

The calibration parameters, $s_x$, $s_y$, $c_x$ and $c_y$ associated with the imagery can be obtained in a separate calibration procedure [5]. It is assumed that these calibration parameters remain constant throughout the video sequence.

The model features considered in the current implementation are points whose location with respect to a global coordinate frame of reference can be denoted by a three vector $(X_i, Y_i, Z_i)$ [1] and vertical lines whose locations can be denoted by only two parameters $(X_i, Y_i)$ (the vertical axis corresponds to the z-axis of the global coordinate frame). Note that the vertical lines are considered to have infinite length so no attempt is made to represent their endpoints.

The position and orientation of the camera with respect to the world frame of reference during frame $j$ of the sequence is captured by two parameters, a rotation $R_j \in SO(3)$ and a translation $\mathbf{T}_j \in \mathbb{R}^3$. This means that given the coordinates of a point in the global coordinate frame, $\mathbf{P}_{iw} \in \mathbb{R}^3$ we can compute its coordinates with respect to camera frame j, $\mathbf{P}_{ij}$ from the following expression.

$$\mathbf{P}_{ij} = R_j(\mathbf{P}_{iw} - \mathbf{T}_j) \quad (2)$$

The reconstruction program takes as input a set of correspondences between features in the omnidirectional imagery and features in the model. For correspondences between point features in the image and point features in the model we can construct an expression which measures the discrepancy between the predicted projection of the point and the vector obtained from the image measurement, $\upsilon_{ij}$, where $\mathbf{P}_{ij}$ is computed from equation 2.

[1] the subscript $i$ serves to remind us that these parameters describe the position of the $i$th feature in the model.

$$\|(\upsilon_{ij} \times \mathbf{P}_{ij})\|^2 / (\|\mathbf{P}_{ij}\|^2 \|\upsilon_{ij}\|^2) \qquad (3)$$

This expression yields a result equivalent to the square of the sine of the angle between the two vectors, $\upsilon_{ij}$ and $\mathbf{P}_{ij}$.

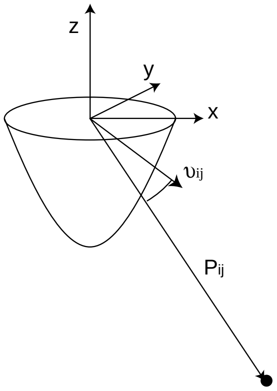

Figure 2: Given a correspondence between a point feature in the omnidirectional image and a point feature in the model we can construct an objective function by considering the disparity between the predicted ray between the camera center and the point feature, $\mathbf{P}_{ij}$, and the vector $\upsilon_{ij}$ computed from the image measurement.

For correspondences between point features in the image and line features in the model we consider the plane containing the line and the center of projection of the image. The normal to this plane, $\mathbf{m}_{ij}$ can be computed from the following expression.

$$\mathbf{m}_{ij} = R_j(\mathbf{v}_i \times (\mathbf{d}_i - \mathbf{T}_j)) \qquad (4)$$

Where the vector $\mathbf{v}_i$ denotes the direction of the line in space and the vector $\mathbf{d}_i$ denotes an arbitrary point on the line. For vertical lines the vector $\mathbf{v}_i$ will be aligned with the z axis $(0, 0, 1)^T$ and the vector $\mathbf{d}_i$ will have the form $(X_i, Y_i, 0)^T$.

The following expression measures the extent to which the vector obtained from the point feature in the omnidirectional imagery, $\upsilon_{ij}$, deviates from the plane defined by the vector $\mathbf{m}_{ij}$.

$$(\mathbf{m}^T \upsilon_{ij})^2 / (\|\mathbf{m}_{ij}\|^2 \|\upsilon_{ij}\|^2) \qquad (5)$$

A global objective function is constructed by considering all of the correspondences in the data set and summing the resulting expressions together. Estimates for the structure of the scene and the locations of the cameras are obtained by minimizing this objective function with respect to the unknown parameters, $R_j$, $\mathbf{T}_j$, $X_i$, $Y_i$ and $Z_i$.

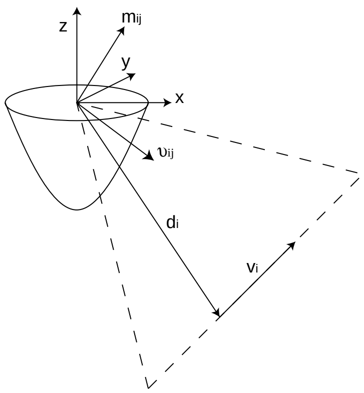

Figure 3: Given a correspondence between a point feature in the omnidirectional image and a line feature in the model we can construct an objective function by considering the disparity between the predicted normal vector to the plane containing the center of projection and the model line, $\mathbf{m}_{ij}$, and the vector, $\upsilon_{ij}$, computed from the image measurement.

An initial estimate for the orientation of the camera frames, $R_j$, can be obtained by considering the lines in the scene with known orientation such as lines parallel to the x, y, or z axes of the environment. If $\upsilon_1$ and $\upsilon_2$ represent the vectors corresponding to two points along the projection of a line in the image plane then the normal to the plane between them in the cameras frame of reference can be computed as follows $\mathbf{n} = \upsilon_1 \times \upsilon_2$. If $R_j$ represents the rotation of the camera frame and $\mathbf{v}$ represents the direction of the line in world coordinates then the following objective function represents the fact that the normal to the plane should be perpendicular to the direction of the line in the coordinates of the camera frame.

$$(\mathbf{n}^T R_j \mathbf{v})^2 \qquad (6)$$

An objective function can be created by considering all such lines in an image and summing these penalty terms. The obvious advantage of this expression is that the only unknown parameter is the camera rotation $R_j$ which means that we can minimize the expression with respect to this parameter in isolation to obtain an initial estimate for the camera orientation.

The current implementation of the reconstruction system allows the user to specify constraints that relate the features in the model. For example the user would be able to specify that two or more features share the same z-coordinate which would force them to lie on the same horizontal plane. This constraint is maintained by reparameterizing the reconstruction problem such that the z-coordinates of the points in question all refer to the same variable in the parameter vector.

The ability to specify these relationships is particularly useful in indoor environments since it allows the user to exploit common constraints among features such as two features belonging to the same wall or multiple features lying on a ground plane. These constraints reduce the number of free parameters that the system must recover and improve the coherence of the model when the camera moves large distances in the world.

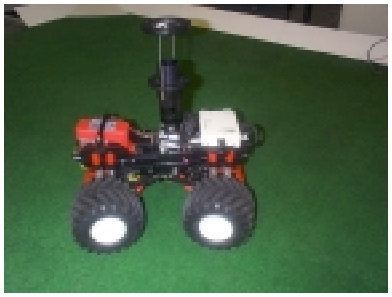

Figure 4: Mobile platform equipped with an omnidirectional camera system that was used to acquire video imagery of an indoor environment.

Using the procedure outlined above we were able to reconstruct the model shown in Figure 6 from 14 images taken from a video sequence of an indoor scene.

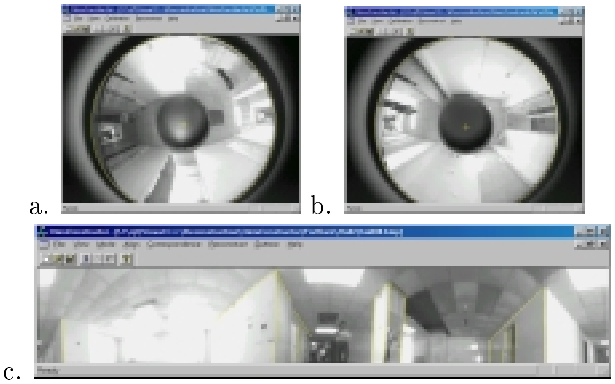

Figure 5: Two of the omnidirectional images from a set of 14 keyframes are shown in a and b. A panoramic version of another keyframe is shown in c.

The polyhedral model is constructed by manually attaching surfaces to the reconstructed features. Texture maps for these surfaces are obtained by sampling the original imagery.

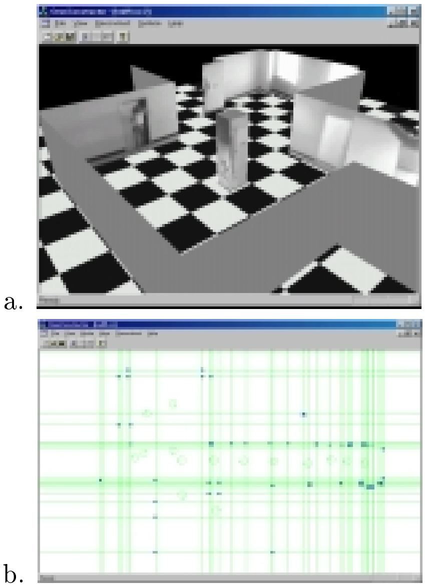

Figure 6: a. 3D model of the environment constructed from the data set shown in Figure 5. b. Floor plan view showing the estimated location of all the images and an overhead view of the feature locations. The circles correspond to the recovered camera positions while the dots and crosses correspond to vertical line and point features.

An important practical advantage of using omnidirectional imagery in this application is that the 3D structure can be recovered from a smaller number of images since the features of interest are more likely to remain in view as the camera moves from one location to another.

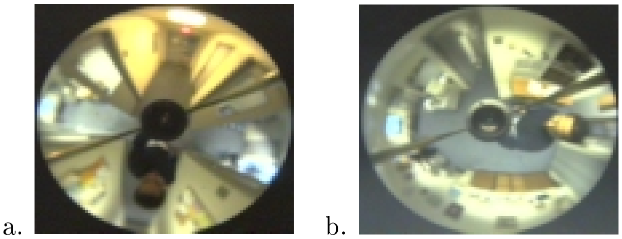

Figure 7: Two images taken from a video sequence obtained as the camera is moved through an office complex.

Once the locations of a set of model features have been reconstructed from the image measurements obtained from a set of keyframes in the sequence, these features can be used as fiducials and the location of the camera at every frame in the sequence can be recovered.

For example, if frame number 1000 and frame num-

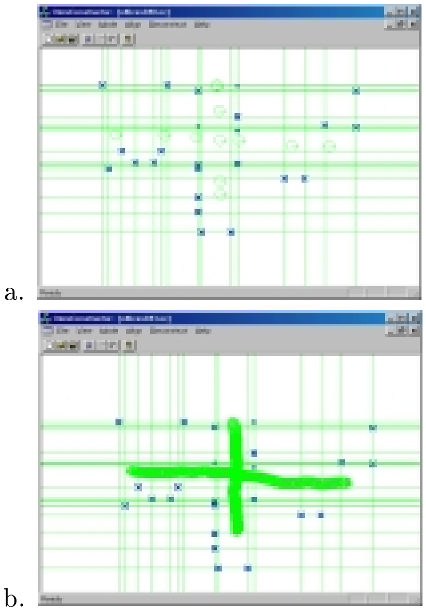

Figure 8: a. A floor plan view of the office environment showing the locations of the features recovered from 11 keyframes in the video sequence. b. Based on these fiducials the system is able to estimate the location of the camera for all the intervening frames by tracking point features through the sequence.

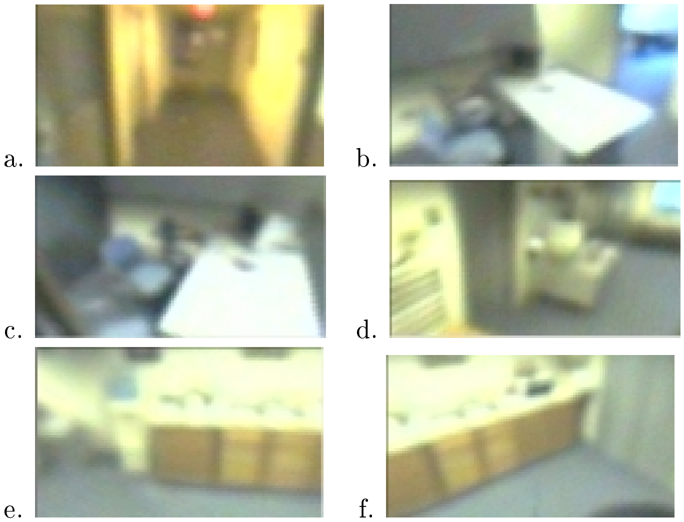

Figure 9: Views generated by the system as the user conducts a virtual tour of the environment.

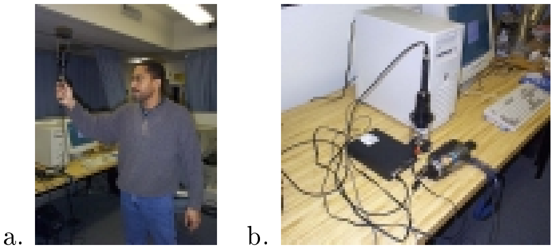

Figure 10: a. The video imagery used to produce the reconstructions of the office environment was acquired using a handheld omnidirectional camera system b. The equipment used to acquire the data

ber 1500 were used as keyframes in the reconstruction process then we know where a subset of the model features appears in these frames. Correspondences between features in the intervening images and features in the model can be obtained by applying applying standard feature tracking algorithms to the data set. The current system employs a variant of the Lucas and Kanade [12] algorithm to localize and track feature points through intervening frames.

Based on these correspondences, the pose of the camera during these intermediate frames can be estimated by simply minimizing the objective function described previously with respect to the pose parameters of the camera. The locations of the feature points are held constant during this pose estimation step. Initial estimates for the camera pose can be obtained from the estimates for the locations of the keyframes that were produced during the reconstruction process.

Figure 7 shows two images taken from a video sequence captured in an office environment. Figure 8a shows the results obtained by applying the reconstruction procedure to a set of 11 keyframes in the video sequence. Figure 8b shows the result of estimating the location of the camera during the 1000 intervening frames in the sequence. The end result is a video sequence where every frame is augmented with an estimate of the camera pose during that instant.

Once the video sequence has been fully annotated the user is able index the data set *spatially* as well as temporally. In the current implementation the user is able to navigate through an immersive environment such as the office complex shown in Figure 7 in a natural manner by panning and tilting his virtual viewpoint and moving forward and backward.

Figure 9 show viewpoints generated by the systems as the user conducts a virtual tour of this environment. Figure 9a shows a snapshot taken in the hallway, Figure 9b and Figure 9c represent views taken while the user explored the east wing of the office complex while figures 9d, 9e and 9f correspond to images of the west wing.

The fact that the reconstruction process can be carried out entirely from the video sequence simplifies the process of data collection. Figure 4 shows a mobile platform outfitted with an omnidirectional camera system produced by Cyclovision inc.. This system was used to acquire the imagery that was used to construct the model shown in Figure 6. Note that the only sensor carried by this robot is the omnidirectional camera it does not have any odometry or range sensors. During the data collection process the system was piloted by a remote operator using an RC link.

The video data that was used to construct the models shown in Figure 8 was collected with a handheld omnidirectional camera system as shown in Figure 10. In both cases the video data was captured on a Sony Digital camcorder and transferred to a PC for processing using an IEEE 1394 Firewire link . The raw images were digitized at a resolution of 720x480 at 24 bits per pixel.

## 3 Related Work

The idea of using omnidirectional camera system for reconstructing environments from video imagery in the context of robotic applications has been explored by Yagi, Kawato, Tsuji and Ishiguro [18, 8, 7, 9]. These authors presented an omnidirectional camera system based on a conical mirror and described how the measurements obtained from the video imagery acquired with their camera system could be combined with odometry measurements from the robot platform to construct maps of the robots environment. The techniques described in this paper do not require odometry information which means that they can be employed on simpler platforms like the one shown in Figure 4 which are not equipped with odometers. It also simplifies the data acquisition process since we do not have to calibrate the relationship between the camera system the robots odometry system.

Szeliski and Shum [16] describe an interactive approach to reconstructing scenes from panoramic imagery which is constructed by stitching together video frames that are acquired as a camera is spun around its center of projection. Coorg and Teller [3] describe a system which is able to automatically extract building models from a data set of panoramic images augmented with pose information which they refer to as pose imagery

From the point of view of robotic applications, reconstruction techniques based on omnidirectional imagery are more attractive than those that involve constructing panoramas from standard video imagery since they do not involve moving the camera and since the omnidirectional imagery can be acquired as the robot moves through the environment.

The process of acquiring omnidirectional video imagery of an immersive environment is much simpler than the process of acquiring panoramic images. One would not really consider constructing a sequence of tightly spaced panoramic images of an environment because of the time required to acquire the imagery and stitch it together. However, this is precisely the type of data contained in an omnidirectional video sequence. By estimating the pose at every location in the sequence the Video Plus system is able to fully exploit the range of viewpoints represented in the image sequence.

Boult [1] describes an interesting system which allows a user to experience remote environments by viewing video imagery acquired with an omnidirectional camera. During playback the user can control the direction from which she views the scene interactively. The VideoPlus system described in this paper provides the end user with the ability to control her viewing position as well as her viewing direction. This flexibility is made possible by the fact that the video imagery is augmented with pose information which allows the user to navigate the sequence in an order that is completely different from the temporal ordering of the original sequence.

The VideoPlus system in similar in spirit to the QuickTime VR system described by Chen [2] in that the end result of the analysis is a set of omnidirectional images annotated with position. The user is able to navigate through the scene by jumping from one image to another. The contribution of this work is to propose a simple and effective way of recovering the positions of the omnidirectional views from image measurements without having to place artificial fiducials in the environment or requiring a separate pose estimation system.

Shum and He [14] describe an innovative approach to generating novel views of an environment based on a set of images acquired while the camera is rotated around a set of concentric circles. This system builds on the plenoptic sampling ideas described by Levoy and Hanrahan [10] and Gortler, Grzeszczuk, Szeliski and Cohen [6]. The presented approach shares the advantage of these image based rendering techniques since the VideoPlus scheme allows you to explore arbitrarily complex environments without having to model the geometric and photometric properties of all of the surfaces in the scene. The rerendered images are essentially resampled versions of the original imagery. However, the scheme presented in this paper dispenses with the need for a specific camera trajectory and it can be used to capture the appearance of extended environments such as office complexes which involve walls and other occluding features which are not accounted for by current plenoptic sampling schemes.

## 4 Future Work

The scheme used to generate views of an environment during a walkthrough is currently quite simple. Given the users desired viewpoint the system selects the omnidirectional image that is closest to that location and generates an image with the appropriate viewing direction. The obvious limitation of this approach is that the viewing position is restricted to locations which were imaged in the original video sequence.

This limitation can be removed by applying image based rendering techniques. One approach to generating novel images is to resample the intensity data from other images depending on the hypothesized structure of the scene as shown in Figure 11. The video plus system has access to the positions of all of the frames in the sequence along with a coarse polyhedral model of the environment which could be used to transfer pixel data from the original views to the virtual view.

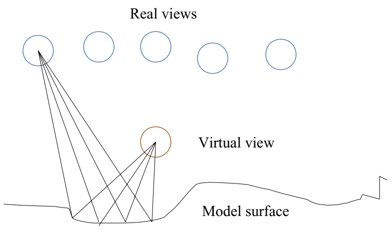

Figure 11: By resampling the pixel data from images in the original sequence it is possible to view the scene from viewpoints not captured in the original data set.

Another approach to generating novel views would be to find correspondences between salient image features in nearby omnidirectional images in the sequence and to use these correspondences to construct a warping function which would map pixels from the original images to the virtual viewpoint [11].

The success of any view generation technique will depend upon having a set of images taken from a sufficiently representative set of viewpoints. A better understanding of how to go about capturing such a data set taking into account the structure of the scene and the viewpoints that are likely to be of most interest is needed. The ultimate goal would be to produce a system where the user could arbitrarily select the desired viewpoint and viewing direction so as to explore the environment in an unconstrained manner.

The largest drawbacks to using omnidirectional video imagery is the reduced image resolution. This effect can be mitigated by employing higher resolution video cameras. One of the tradeoffs that is currently being explored is the possibility of acquiring higher resolution imagery at a lower frame rate. This would allow us to produce sharper images of the scene but would either slow down the data acquisition process or require better interpolation strategies.

## 5 Conclusions

This paper presents a simple approach to capturing the appearance of immersive scenes based on an omnidirectional video sequence. The system proceeds by combining techniques from structure from motion with ideas from image based rendering. An interactive photogrammetric modeling scheme is used to recover the positions of a set of salient features in the scene (points and lines) from a small set of keyframe images. These features are then used as fiducials and tracked through the video sequence in order to estimate the position of the omnidirectional camera at every frame in the video clip.

By augmenting the video sequence with pose information we provide the end user with the capability of indexing the video sequence spatially as opposed to temporally. This means that the user can explore the image sequence in ways that were not envisioned when the sequence was initially collected.

The cost of augmenting the video sequence with pose information is very slight since it only involves storing six numbers per frame. The hardware requirements of the proposed scheme are also quite modest since the reconstruction is performed entirely from the image data. It does not involve a specific camera trajectory or a separate sensor for measuring the camera position.

Future work will address the problem of generating imagery from novel viewpoints and improving the resolution of the imagery generated by the system.

## References

[1] Terrance E. Boult. Remote reality via omnidirectional imaging. In Scott Grisson, Janet McAndless, Omar Ahmad, Christopher Stapleton, Adele Newton, Celia Pearce, Ryan Ulyate, and Rick Parent, editors, *Conference abstracts and applications: SIGGRAPH 98, July 14–21, 1998, Orlando, FL*, Computer Graphics, pages 253–253, New York, NY 10036, USA, 1998. ACM Press.

[2] S. E. Chen. Quicktime vr - an image-based approach to virtual environment navigation. In *SIGGRAPH*, pages 29–38, August 1995.

[3] Satyan Coorg and Seth Teller. Automatic extraction of textured vertical facades from pose imagery. Technical report, MIT Computer Graphics Group, January 1998.

[4] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings of SIGGRAPH 96. In Computer Graphics Proceedings, Annual Conference Series*, pages 11–21, New Orleans, LA, August 4-9 1996. ACM SIGGRAPH.

[5] C. Geyer and K. Daniilidis. Catadioptric camera calibration. In *International Conference on Computer Vision*, pages 398–404, 1999.

[6] 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.

[7] Hiroshi Ishiguro, Takeshi Maeda, Takahiro Miyashita, and Saburo Tsuji. A strategy for acquiring an environmental model with panoramic sensing by a mobile robot. In *IEEE Int. Conf. on Robotics and Automation*, pages 724–729, 1994.

[8] Hiroshi Ishiguro, Kenji Ueda, and Saburo Tsuji. Omnidirectional visual information for navigating a mobile robot. In *IEEE Int. Conf. on Robotics and Automation*, pages 799–804, 1993.

[9] Hiroshi Ishiguro, Masashi Yamamoto, and Saburo Tsuji. Omni-directional stereo. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(2):257–262, February 1992.

[10] 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.

[11] Maxime Lhuillier and Long Quan. Image interpolation by joint view triangulation. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, volume 2, pages 139–145, June 1999.

[12] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. 7th International Joint Conference on Artificial Intelligence*, 1981.

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

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

[15] Tomas Svoboda, Tomas Pajdla, and Vaclav Hlavac. Epipolar geometry for panoramic cameras. In *European Conference on Computer Vision*, pages 218–232. Springer, 1998.

[16] R. Szeliski and H. Y. Shum. Creating full ciew panoramic image mosaics and texture-mapped models. In *SIGGRAPH*, pages 251–258, August 1997.

[17] Camillo J. Taylor and David J. Kriegman. Structure and motion from line segments in multiple images. *IEEE Trans. Pattern Anal. Machine Intell.*, 17(11), November 1995.

[18] Yasushi Yagi, Shinjiro Kawato, and Saburo Tsuji. Real-time omnidirectional image sensor (copis) for vision-guided navigation. *IEEE Journal of Robotics and Automation*, 10(1):11–21, February 1994.