3D Ranging and Virtual View Generation using Omni-view Cameras

Kim C. Ng[a], Mohan M. Trivedi[a], and Hiroshi Ishiguro[b]

[a]Computer Vision and Robotics Research Laboratory, University of California-San Diego
[b]Department of Social Informatics, Kyoto University, Japan

ABSTRACT

In this paper we describe elements of an *ubiquitous vision system* - a system that enables viewers to observe a remote dynamic scene from any viewing perspective at any time instant. This integrated framework has the potential to achieve *real-time* generation of many 3D virtual views concurrently with high-dynamic events in the scene, by *not* utilizing 3D model. The paper presents a brief overview of the system including its requirements, components, and preliminary results of acquiring 3-D range estimation and virtual view generation.

Keywords: 3D ranging, view generation, view-on-demand, dynamic scenes, multiple baseline stereo, omni-directional viewing, wide baseline stereo

1. UBIQUITOUS VISION: REQUIREMENTS

Ubiquitous vision system can be realized using multiple cameras. There are a number of important issues which need to be addressed in the development of such a system of practical utility. We are developing such a realization of the system and the important features of this include: 1) Multiple omni-directional vision sensors are exploited to provide wide scene coverage with smallest possible number of cameras, data redundancy, and a way to synthesize view by composing multiple cameras' pixels. 2) Both 3D range and virtual views are generated in *one step*, in which virtual and real worlds are immersed into one, in contrast to the total virtual of Virtual Reality. 3) Distributed computing is utilized to allow many views generated concurrently. 4) Efficient data structure and image caches are exploited to speed up view generation process.

There are three fundamental problems associated with the development of a useful ubiquitous system:

1) How to generate 3D virtual views of a dynamic scene in real-time while simulating the effects of dynamic observers.
2) How to cover the entire scene with smallest possible number of cameras.
3) How to allow many virtual views generated concurrently.

The first problem arises mainly because both objects and viewers are dynamic. Objects in the scene can be moving and the viewers can choose their prefer viewpoints anywhere in the 3D real world. The viewpoints are constantly moving, rotating, and translating away from the cameras' location at every video frame instant. Since the world is dynamic, every video frame is going to be different view. Constructing the entire 3D scene is computational intensive and network bandwidth demanding; most importantly, it is not necessary when the dynamic viewer is interested in only a small field of view for every video frame instant. However, the conventional virtual view creation methods necessitate first have 3D data known for every camera's pixel in order to create a new virtual view for a dynamic viewer. The computational overhead of these conventional geometric approaches increases explosively with increasing number of cameras and increasing coverage of the scene. Therefore, a method that can create a very specific view and compute only the necessary 3D is to be developed. The method has to be relatively computationally insensitive to the number of cameras employed and geometric complexity of the scene as well.

In order to employ minimum possible number of cameras, multiple omni-directional vision sensors (ODVS) are a better choice. Omni-directional sensors have wide field of view, thus large overlapping viewing areas, which is necessary for depth

[a] K.C.N.: Email: kimng@ece.ucsd.edu; WWW: http://swiftlet.ucsd.edu/~kimng
M.M.T: Email: trivedi@ece.ucsd.edu; WWW: http://swiftlet.ucsd.edu/~trivedi
[b] H.I. (Visiting Professor): Email: ishiguro@kuis.kyoto-u.ac.jp; WWW: http://www.lab7.kuis.kyoto-u.ac.jp/~ishiguro

extraction using stereo. They also provide continuous and smooth observation of the environment, which is essential for smooth virtual walk-thorough with coherent spatial-temporal content. Naturally, ODVSs have replaced the need for many cameras; that reduces costs and acquisition time. Additionally, these ODVSs should be placed much further apart than conventional stereo. If the ODVSs are closely together, one ODVS will observe another, or in other words, one ODVS will block another ODVS' view. One of the major attractiveness of ODVS will then be obstructed. Consequently, wider camera baseline has to be configured, and a new method has to be developed to perform wide baseline stereo and to overcome occlusion problems it introduces.

The third problem rules out the possibility of having a centralized computer to generate large number of views. Therefore, the users' computers must be doing the job of view generation, and input data must be traveling through the network. Sending 3D-video stream over the network to client side demands at least three times overhead than the regular 2D-video stream. A cluster of four video streams will have 12 times higher overhead! The best way is to send raw video data across the network. This leaves the clients' computers compute both 3D and views. If implementing the conventional view rendering approaches, 3D must be first extracted for every pixel. This again introduces long latency between viewpoint picking and final view display. Worst yet, when the viewer changes his/her viewpoint for the next video frame to the next video cluster, if a completely new set of video images arrives, the previously computed 3D for every pixel will be wasted. The previously computed 3D for every pixel will be completely wasted if the viewer never comes back to the same video cluster's viewpoint. Between the walk through from one cluster to another, the system will have longer latency. Ultimately, compute 3D only for the small window of necessary virtual pixels will be the most efficient and consistent way to possess ubiquitous vision.

2. UBIQUITOUS VISION: SYSTEM OVERVIEW

Ubiquitous vision is bridging the gap between computer vision (real world) and computer graphics (virtual world), real scene is being observed from virtual viewpoints. The technique developed here is aiming at solving the fundamental problems of ubiquitous vision system discussed in the previous section, and it is to fuse both computer vision technique and computer graphics technique into a seamless, synergetic one. Put it in other words, the technique cannot be explicitly classified as either computer vision technique or computer graphics technique. One possible technique, which can solve the first three of the four underlying technical problems, is to perform *non-image space search and rendering*. Non-image space search and rendering has the characteristics of *not* needing 3D model and having high parallelism for every virtual pixel. Most importantly, it performed exactly only at the user specified viewing direction. That is what we call "*View-on-Demand*."

The process is different from how a geometrically-valid virtual view has been generated all these years. All these time, 3D model has been first created for the entire scene and then voxels are projected to the small window of new virtual camera's location[2, 3]. In ubiquitous vision system setup, the cameras' coverage is large and the scene is dynamic but the virtual viewing area is small. Following this conventional method of recovering depth for every physical pixel and then re-projecting voxels to the virtual view is too time-consuming. However, when non-image space is introduced, we calculate only the necessary 3D along the virtual viewing rays. These 3D points are much small subset of the entire wide-angle view.

The task of computer vision researchers has always been attempting to recover 3D range by the provided information of image color or object reflectance (object color is assumed known), while the task of computer graphics researchers has always been to recover the view color by the provided information of 3D range (object depth is assumed known). Our technique here is to both estimate depth and recover color simultaneously without the intermediate sequential steps, so that virtual view synthesis with range estimation computation is performed exactly only at the user specified viewing direction.

Despite the additional complexities, our approach is highly promising for real-time view generation of dynamic scene for dynamic observers since 3D global model is not needed and its computing for every virtual pixel can be processed in parallel. Because 3D model is not utilized, the input data to the algorithm are raw omni-directional video images. Raw images are much easier to transmit through the network in comparison with global 3D model. Therefore, virtual views can be generated by viewers' computer in the mean of Distributed Computing.

Additionally, the non-image space technique brings in the capability of performing wide-baseline stereo on omni-directional videos, by that the number of cameras required to cover the entire scene is reduced. When cameras are placed far apart, objects appear much larger in closer camera than in the further ones. This makes it difficult to perform the correspondence in the disparity space (whether it is area-based or feature-based stereo), even if there is no occlusion. However, our developed technique allows us to place the cameras a few feet apart with arbitrary baseline configurations. The template for matching

will be created larger for the closer camera while smaller for the further camera. A pixel in the further camera covers the space larger than a pixel in the closer camera. Therefore, it is not a one-to-one matching process; that is one pixel in the further camera will compare with multiple pixels in the closer camera. A wider baseline besides brings more accurate range estimation, it also reduces the need of dense cameras placement; and thus it reduces the computation cost and network congestion.

3. SYNTHESIZING VIRTUAL VIEW AND RANGE ESTIMATION: EXPERIMENTAL FINDINGS

Our interest is in the development of a *new* depth extraction approach that (1) works with ODVS imagery, and (2) tightly integrates with image rendering. Typical stereo approaches progress from finding disparities first, to then computing depth. Finding disparities requires a solution to the correspondence problem, the classic search problem that, for large numbers of cameras and large fields of view, becomes computationally expensive if not outright impractical. Instead, we propose a *new* approach where we search the *non-image space*.

In the current experimental setup, only four ODVSs are employed. This infrastructure will allow us to fully realize the ideas of interactive walk-through in a remote environment. Discussions on the designs of ODVSs that is more thorough can be found in Ishiguro's paper [1]. These sensors are arranged as shown in Figure 1. The shortest baseline is 2.8 feet, while the longest is 6.2 feet. Figure 2 shows the preliminary results on range estimation. The results seem reasonable, although more concrete analysis will be performed in the future.

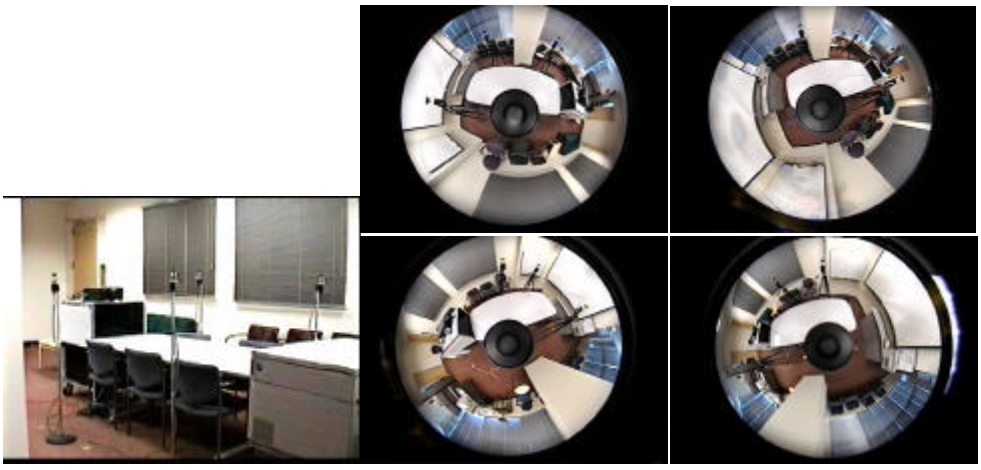
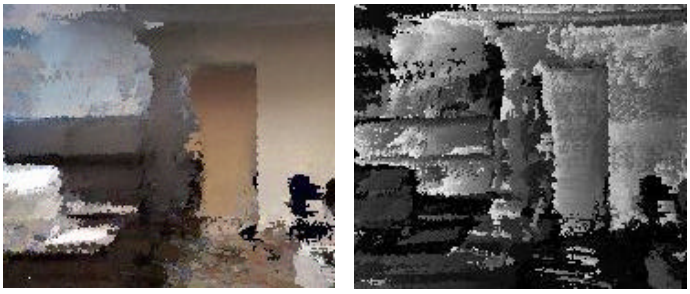


Figure 1 *Experimental setup showing placement of 4 omniview cameras. Examples of images acquired by these cameras are also shown above.*



(a) (b)

Figure 2 *Preliminary range estimation results. (a) A virtual image of our conference room. (b) The depth map of the corresponding image in (a). Closest object is represented in black, while the furthest is represented in white.*

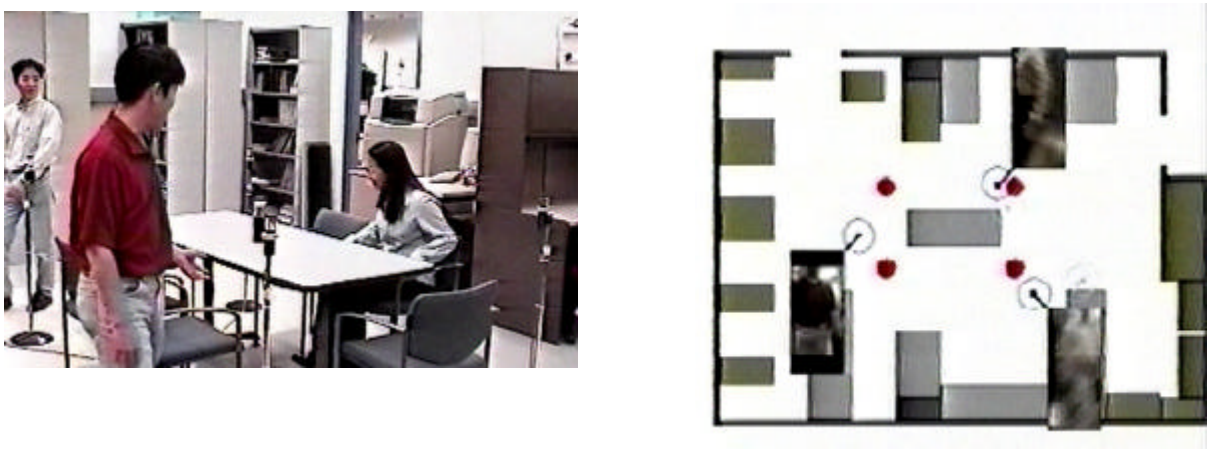


Figure 3 *Experimental setup for person tracking. Results of a real time tracker using 4 omniview sensors are also shown.*

For realizing more sophisticated systems, we can combine the view generation with the real-time human tracking system[1] that we have developed so far. Figure 3 shows the real-time human tracking system using omnivew sensors. The system detects walking people, robustly locates them by a multiple camera stereo, and clips the best views for observing the walking people from the omni-directional images taken by the omnivew sensors. By combining the view generation with this human tracking system, we can generate virtual views taken from the view points of the walking people.

4. CONCLUDING REMARKS

In this paper we have presented an overview of a vision system allows extraction of 3-D range information as well as generation of a virtual view in an integrated manner. The system utilizes a set of omniview sensors and a novel multiple baseline stereo approach. Preliminary results show a definite promise for this approach. Such a system should have a number of applications where remote sites need to be observed and activities need to be monitored.

REFERENCES

1. H. Ishiguro, "Development of low-cost compact omnidirectional vision sensors and their applications," In proceedings of Int. Conference on Information Systems, Analysis and Synthesis, pp. 433-439, 1998.
2. R. Jain and K. Wakimoto, "Multiple perspective interactive video," Proc. Int. Conf. Multipmedia Computing and Systems, 1995.
3. M. Okutomi and T. Kanade, "A multi-baseline stereo", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 353-363, 1993.