

Modeling, Combining, and Rendering Dynamic Real-World Events From Image Sequences

Sundar Vedula, Peter Rander, Hideo Saito, and Takeo Kanade

*The Robotics Institute
Carnegie Mellon University*

Abstract

Virtualized Reality creates a model of time-varying real-world events from image sequences. The model can be used for manipulating and combining these events, and for rendering new virtual images. In this paper, we present two recent enhancements to Virtualized Reality. We present Model Enhanced Stereo (MES) as a method to use widely separated images to iteratively improve the quality of each local stereo output in a multi-camera system. We then show, using an example, how Virtualized Reality models of two different events are integrated with each other, and with a synthetic virtual model. In addition, we also develop a new calibration method that allows simultaneous calibration of a large number of cameras without visibility problems. The method goes from capturing real image sequences, integrating two or more events with a static or time-varying virtual model, to virtual image sequence generation.

1 Introduction

Many methods for obtaining graphics models of real objects have been studied recently. A large amount of work has focused on recovery of three dimensional shape models from range images, which are obtained by direct range-scanning hardware [2][4][18], or image-based shape reconstruction techniques [5][11][14][15][17]. Image-based modeling [1][3][7][8] has also seen significant development, in which a set of real images implicitly represent the object scene.

Most of this work, however, has been on developing algorithms to build static models of relatively small objects. Instead, our goal is to reconstruct dynamic models of larger objects so that real events can be represented in a virtual world. Modeling dynamic events requires a multi-camera video capture system, rather than a typical setup including a turntable with a single camera for modeling small, static objects. In the *Virtualized Reality* system [12][13], we have demonstrated the ability to recover dynamic three-dimensional geometric models of a scene with multiple human-sized objects.

Early versions of our system used stereo with some manual intervention to compute the 3D structure, and used a three-dimensional image-based rendering method to synthesize novel views. We then incorporated a volumetric integration algorithm [10] to create a unified 3D model of the scene. Due to inaccuracies in both calibration and stereo matching, these models often contained errors in the recovered geometry.

This paper presents two recent enhancements to our system: model refinement by enhanced stereo, and ability to integrate multiple events into a virtual reality model, together with an improved calibration method. We propose Model Enhanced Stereo (MES), which iteratively uses the three-dimensional model obtained by merging the range images of all cameras for improving the stereo depth estimates at each camera. We then show using an example, how Virtualized Reality models of two different events are integrated with each other, and with a synthetic virtual model. In addition, to improve the accuracy of the camera calibration,

we developed the calibration system so that densely spaced calibration points spread throughout the volume of interest can be simultaneously viewed by all cameras. We also show how precise silhouette information can be used to refine the shape of a 3D model.

2 System Overview

Our multi-camera 3D system for Virtualized Reality is shown in Figure 1. It is designed for imaging a human-sized dynamic event from 51 omni-directionally distributed video cameras. For every time instant, we run multi-baseline stereo [11] for each camera to obtain a set of range images. These range images are then merged into a volumetric model and an iso-surface extraction is run on the resulting model to produce a polygonal mesh. This mesh is then texture mapped, which lets us synthesize virtual images from arbitrary viewpoints [13]. Figure 2 explains the detailed dataflow in our system.

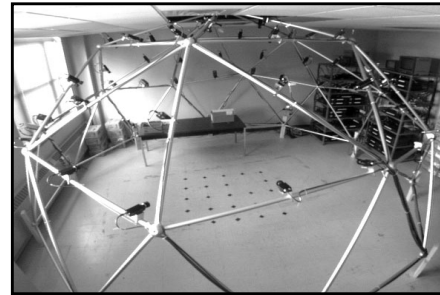


Figure 1: Virtualized Reality 3D Dome camera configuration: 51 cameras are distributed on a hemispherical structure, 5 meters in diameter

Before any processing, the cameras are calibrated to relate the 2D coordinates on the image plane of every camera to 3D coordinates in the scene. Any system, including ours, that produces a full Euclidean model of the scene being imaged needs a strong calibration method, or knowledge of the exact mapping from scene to image coordinates. Calibrating a volume of many cubic meters accurately with respect to cameras in all directions poses many challenges because of visibil-

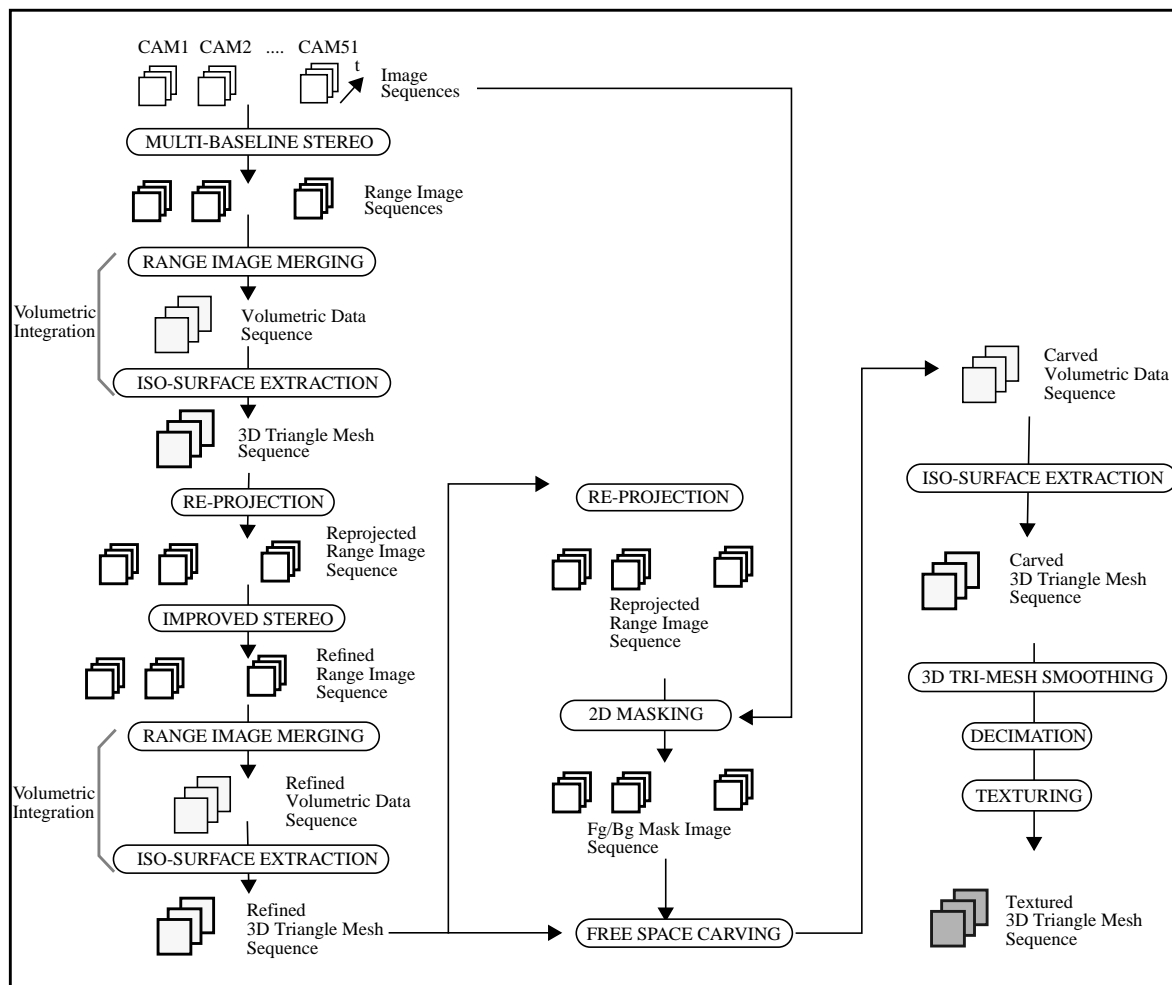


Figure 2: Block diagram of dataflow in the system

ity constraints. Standard calibration patterns and markers are not a viable solution, because of the need for all cameras to simultaneously view all marked points. Also, the volume of the space to be calibrated is around 8 cubic meters; thus any large fixture is bound to create problems of visibility across cameras.

It is well known that inaccuracies in calibration affect the subsequent stereo matching process, since search is typically limited along the epipolar line. However, the merging of range images into a Euclidean 3-D model would be affected as well, since neighboring cameras would now vote differently for the location of the same point in the scene. The global texture mapping process is also affected by inaccuracies in a similar fashion.

To overcome these challenges, we use a new large-volume calibration method. A calibrated bar with markers at known positions on it is swept through the space, and the grid of points thus defined is used to calibrate all cameras. We built a thin 2 meter horizontal bar with 11 LEDs mounted 20 cm apart. The floor is calibrated with markers 20 cm apart in both dimensions, as is the vertical motion of the tripods that the bar is mounted on. Then, this bar is swept across vertically, and perpendicular to its length so that the precise coordinates of these points are known. These 3-D points are imaged in all the cameras, and their 2-D image locations are determined by simple dot-detection; because the bar is thin, there is no visibility problem. In addition, availability of three dimensional calibration data (unlike our earlier systems [12][13]) allows us to combine the extrinsic and intrinsic calibration steps into one. With this grid of points and their corresponding image locations available, we use a well-known non-planar calibration algorithm by Tsai[16], because it estimates radial distortion in addition to the extrinsic and intrinsic parameters, and the implementation is robust.

3 Model refinement using enhanced stereo and silhouette information

Multi-baseline stereo searches for matches in neighboring images across different levels of disparities, and yields range images. A volumetric model is obtained by sensor fusion of these (somewhat noisy) range images [12]. Most of the inaccuracies in the model are because of incorrect stereo, which results from two problems: first, insufficient texture on an object may result in identical matches for widely separated disparity values. Second, since the algorithm uses a window to match regions across images [11], the best match can result for a window position that partially overlaps the silhouette of an object.

We present a method to refine the volumetric model by limiting these errors in a next enhanced iteration of the stereo process. This model is converted to a polygonal mesh representation, and projected into virtual cameras corresponding to the location of the original cameras. This gives us an approximation to the silhouette of the object, along with a depth value at each pixel.

Based on the projected depth, we iterate the multi-baseline stereo process by imposing tighter bounds on the range of disparity values to be searched for each pixel. Each iteration produces more accurate depth estimates, in addition to eliminating a large number of false matches that initial individual stereo may contain. Also, the contour bounding the projection gives us an estimate of the exact silhouette of the object, which can be used to ensure that the stereo window does not overlap edges of the object during the next iteration of the matching process. We call this method of limiting the search space using depth and window bounds Model Enhanced Stereo (MES). Figure 3(a) shows the result of the initial stereo depth from one viewpoint, while Figure 3(b) shows the depth estimated by MES. We see that the estimates of depth are more accurate and far less noisy. The results of MES for the various viewpoints are again volumetrically merged to yield a significantly more accurate volumetric model. This procedure may be repeated to further improve the accuracy of the model.

Also, MES helps us segment the model into objects and the environment (dome and floor), because volumetric merging of range images from MES gives us Euclidean 3-D coordinates. Knowing the approximate location and bounds of our objects, it is easy to eliminate volumes outside these bounds that are discontinuous from the objects. Methods such as the Lumigraph [3] and voxel coloring [15] are able to have a plane as the background and use a chroma-key technique in image space to separate the foreground objects from the background. This separation is then propagated into the whole model. While chroma-key is usually fairly accurate, it fails for an omnidirectional imaging system since uniform lighting is impossible without the lights being visible in one or more cameras.

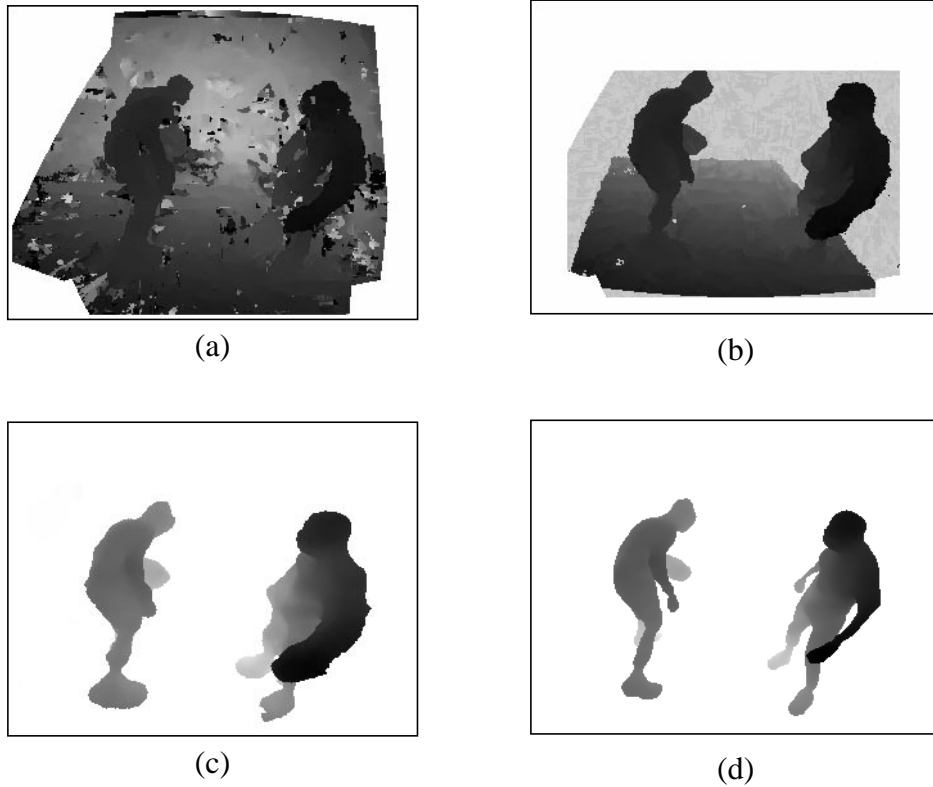


Figure 3: (a) Range image from initial multi-baseline stereo (b) Range image from Model Enhanced Stereo (c) Volumetric model obtained by merging MES results from 51 cameras (d) Volumetric model after carving using 10 exact silhouettes

In our implementation, we use silhouettes to carve out free space from the volumetric model directly, as a post-processing to the stereo algorithm. When the refined volumetric model is projected into each input camera, we get an approximation to the actual silhouette. While a snake [6] algorithm could be used to lock onto the exact contour, we use the approximation as a template for a human operator to modify the silhouette, to visual accuracy. Figure 3(c) shows the volumetric model obtained as a result of merging multiple MES range images. Figure 3(d) shows the same model refined with knowledge of the exact contour. The exact silhouettes need not be specified for all images; in our example, we find that specifying silhouettes for only 10 of the 51 original images suffice.

4 Combining Multiple Dynamic Virtualized Reality Events and Virtual Sets

It is possible to combine the volumetric models of multiple events in a spatial and temporal manner, even when the events are recorded separately and possibly at different locations. Each of these models is transformed into a single, unified co-ordinate system, and their temporal components are matched to ensure that the combined model is correct throughout the length of the sequence. We show such an example, where two events of humans in a laboratory dribbling and passing a ball, are integrated with a virtual basketball court to produce a virtual basketball play sequence.

Before integration, the volumetric models are texture-mapped using the intensity images from various cameras [13]. Since the volumetric models are discretized metric descriptions of every surface, along with texture maps, they can be manipulated in the same way as traditional graphics models and added into any VR system that uses polygon-based rendering. Thus, Virtualized Reality models and textured CAD models can interact with each other in a natural manner.

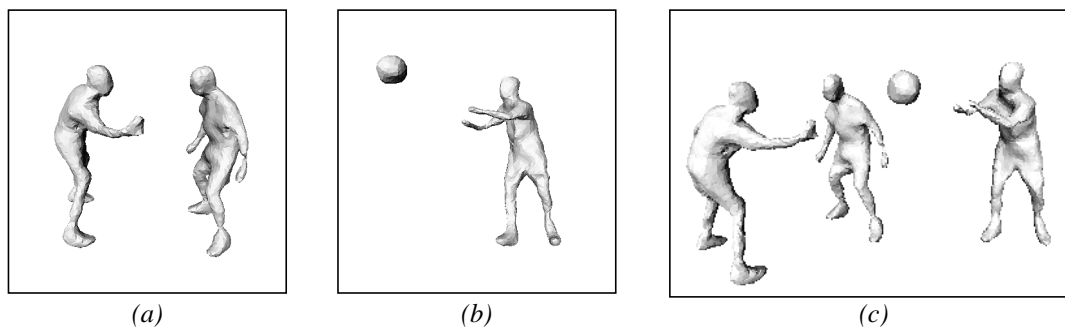


Figure 4: (a) Volumetric model of one frame of the first event with two players (b) Volumetric model of the second event with one player (c) Volumetric model obtained by combining both events spatially to create an event with three players

4.1 Combining Multiple Events into a Single Virtualized Reality Representation

To convert Virtualized Reality models of different events into a single representation, they are first combined and aligned spatially. To achieve this, a rotation and translation are applied to each of the models, so that the origin of each local co-ordinate space is mapped to the location of the world origin with the desired orientation. Each of these models is typically a triangle mesh with a list of vertex coordinates, texture coordinates, and polygon connectivities. The vertex coordinates of each of these models are defined independently with respect to a local coordinate system.

On the other hand, temporal integration of model sequences involves determining how the sequences are related to each other in time. If one or more sequences are not modeled at the frame rate of the desired virtual sequence, those sequences would need to be subsampled or supersampled appropriately. For non-integral multiples, some temporal interpolation between models is called for. Once this is done, each time frame on the motion sequence needs to be mapped to a frame on the global time scale. By this mapping of the component image sequences to the global time scale, the component events are overlapped or concatenated one after the other. In addition, the individual frames of a sequence may be used in reverse.

In our example, two separate events are recorded. The first event, shown in Figure 4(a), involves two players, where one player bounces a basketball and passes it off to the side while the other attempts to block the pass. The second event, shown in Figure 4(b) involves a single player who receives a basketball and dribbles the ball. Since we are free to choose frames in reverse, we actually recorded the second event by having the player dribble and then throw the ball to the side since that was easier. Both these events are recorded separately, so no camera ever sees all three players at once.

The events are combined, so that the final model contains a motion sequence where the first player passes the ball to the third player, as the second player attempts to block this pass. This is done by a spatial transform that places the ball at the end of the last frame of the first event so as to coincide with the position of the ball at the beginning of the second event and

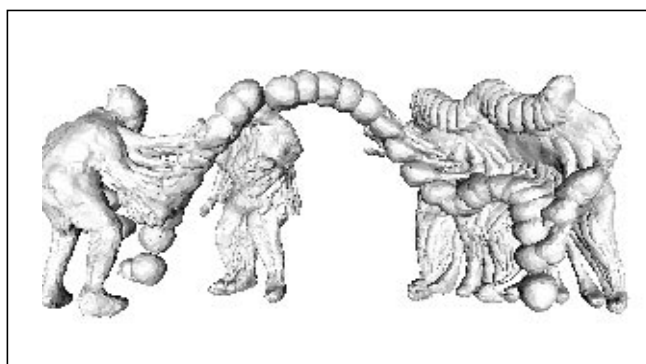


Figure 5: Sequence of combined volumetric models [one of Figure 4(c) for each time instant] superimposed on each other: effect of motion seen similar to that on a time-lapse photograph

a concatenation of the two events in time. Also, the polygonal model of the last frame of the first event is modified so that the ball is removed. This edit prevents two balls from appearing in the scene. The result of combining the two events is shown in Figure 4(c).

Figure 5 shows a number of volumetric models of combined events superimposed on each other, as the “pass” happens. The effect produced is similar to that seen in a time-lapse photograph of a scene with moving objects.

4.2 Combining a Virtualized Reality Representation with a Virtual Model

Since a Virtualized Reality representation is a metric description of an event, we can introduce other virtual models into this representation. The Virtualized Reality models combined in Section 4.1 into a sequence of polygonal meshes are textured and introduced into a CAD model of a virtual basketball court, to generate a unified geometry and texture representation. Figure 6 shows a sequence of rendered images of this combined model that simulate a flythrough. The virtual camera spirals around and above the players, as it is pointed at them. We therefore have a sequence of virtual images, that captures spatial motion of the virtual camera, and the dynamic event itself. While this is a case where the virtual model (basketball court) is static, one can imagine a case where a Virtualized Reality model is combined with a time-varying virtual event.

5 Conclusions

In this paper, we presented two enhancements to the Virtualized Reality system. Model Enhanced Stereo iteratively improves the accuracy in stereo correspondence using a 3D model built from initial stereo results. This improvement is achieved by the fact that all images are used in obtaining stereo correspondences for each camera. We also showed how two real events are combined with each other, into a virtual space. Also, to improve the accuracy of the camera calibration, we have developed a new calibration system that allows simultaneous calibration of all cameras, without visibility problems.

Our Virtualized Reality system provides a new capability in creating virtual models of dynamic events involving large free-form objects such as humans. In addition to the modeling, we have the capability to produce synthetic video from a varying virtual viewpoint. The sys-

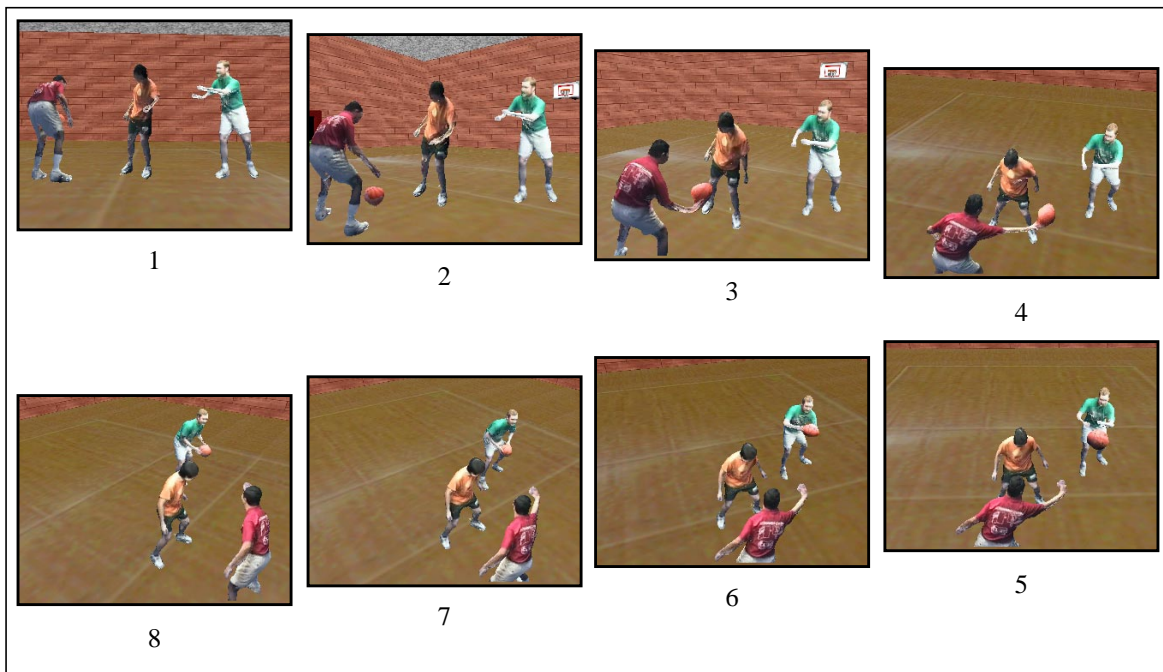


Figure 6: Two separate dynamic events - one with two players and another with a single player are combined into a CAD model of a virtual basketball court. The sequence of images are those seen by a virtual camera that moves along a spiral trajectory around the players, and upwards. Notice that the event is dynamic - the camera motion happens during the course of the play

tem goes from capturing real image sequences, creating Virtualized Reality models from these observed sequences, integrating two or more events with a static or time-varying VR model, and to virtual image sequence generation. Future work involves development of abstractions to represent recovered 3D geometry in other image based modeling representations. These representations, such as projective shape models or depth/correspondence maps, can be local to the virtual viewpoint to facilitate rendering of virtual views without explicitly recovering the full volumetric model.

6 References

- [1] E. Chen and L. Williams. "View interpolation for image synthesis". *Proc. SIGGRAPH'93*, pp.279-288, 1993.
- [2] B. Curless and M. Levoy. "A volumetric method for building complex models from range images". *Proc. SIGGRAPH '96*, pp.303-312, 1996.
- [3] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen. "The lumigraph". *Proc. SIGGRAPH'96*, pp.43-54, 1996.
- [4] A. Hilton, J. Stoddart, J. Illingworth, and T. Windeatt. "Reliable surface reconstruction from multiple range images". *Proc. ECCV'96*, 117-126, 1996.
- [5] S. B. Kang and R. Szeliski. "3-D scene data recovery using omnidirectional multibaseline stereo". *International Journal of Computer Vision*, 25(2), PP.167-183, 1997.
- [6] M.Kass, A.Witkin, and D.Terzopoulos, "Snakes: Active contour models", *International Journal of Computer Vision*, 1(4), pp.321-331, 1988.
- [7] S. Laveau and O. Faugeras. "3-D Scene representation as a collection of images". *Proc. ICPR'94*, pp.689-691, 1994.
- [8] M. Levoy and P. Hanrahan. "Light field rendering". *Proc. SIGGRAPH'96*, pp.31-42, 1996.
- [9] C.E.Liedtke, H.Busch, and R.Koch, "Shape adaptation for modeling of 3D objects in natural scenes", *Proc. CVPR '91*, pp.704-705, 1991.
- [10] W. Lorensen and H. Cline. "Marching cubes: a high resolution 3D surface construction algorithm". *Proc. SIGGRAPH'87*, pp. 163-170, 1987.
- [11] M. Okutomi and T. Kanade. "A multiple-baseline stereo", *IEEE Trans. Pattern Analysis and Machine Intelligence*. PAMI-15(4), pp.353-363, 1993.
- [12] P. Rander, P.J. Narayanan, and T. Kanade, "Recovery of dynamic scene structure from multiple image sequences". *Proc. IEEE Multisensor Fusion and Integration for Intelligent Systems*, pp. 305-312, 1996.
- [13] P. Rander, P.J. Narayanan, and T. Kanade, "Virtualized Reality: Constructing Time-Varying Virtual Worlds from Real World Events". *Proc. IEEE Visualization '97*, pp. 277-283, 1997
- [14] S. M. Seitz and C. R. Dyer, View Morphing, *Proc. Proc. SIGGRAPH '96*, pp. 21-30. 1996
- [15] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring", *Proc. IEEE CVPR'97*, pp. 1067-1073, 1997
- [16] R. Tsai. "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses". *IEEE Trans. Robotics and Automation*, RA-3(4), pp. 323-344, 1987.
- [17] T. Werner, R. D. Hersch and V. Hlavac. Rendering Real-World Objects Using View Interpolation. *IEEE International Conference on Computer Vision*, Boston, pp.957-962, 1995.
- [18] M.D. Wheeler, Y. Sato, and K. Ikeuchi, "Consensus surfaces for modeling 3D objects from multiple range images", *DARPA Image Understanding Workshop*, pp. 911-920, 1997.