A Movable Image-Based Rendering System And its Application to Multiview Audio-Visual Conferencing

S.C.Chan, Z.Y.Zhu, K.T.Ng, C.Wang, S.Zhang, Z.G.Zhang
Electrical and Electronic Engineering Department, The University of Hong Kong, Hong Kong, P. R. China
E-mail: {scchan; zyzhu; ktng; cwang;szhang,zgzhang}@eee.hku.hk

Abstract—**Image-based rendering (IBR) is an emerging technology for rendering photo-realistic views of scenes from a collection of densely sampled images or videos. It provides a framework for developing revolutionary virtual reality and immersive viewing systems. This paper studies the design of a movable image-based rendering system based on a class of dynamic representations called plenoptic videos. It is constructed by mounting a linear array of 8 video cameras on an electrically controllable wheel chair with its motion being controllable manually or remotely through wireless LAN by means of additional hardware circuitry. We also developed a real-time object tracking algorithm and utilize the motion information computed to adjust continuously the azimuth or rotation angle of the movable IBR system in order to cope with a given moving object. Due to the motion of the wheel chair, videos may appear shaky and video stabilization technique is proposed to overcome this problem. The system can be used in a multiview audio-visual conferencing via a multiview TV display. Through this pilot study, we hope to develop a framework for designing movable IBR systems with improved viewing freedom and ability to cope with moving object in large environment.**

I. INTRODUCTION

Image-based rendering/representation (IBR) is an emerging and promising technology for rendering new views of scenes from a collection of densely sampled images or videos. It has potential applications in virtual reality, immersive television and visualization systems. A recent survey of IBR can be found in [15]. Central to IBR is the plenoptic function [1], which describes all the radiant energy that can be perceived by the observer at any point in space and time. The plenoptic function is thus an 7-dimensional function of the viewing position (V_x, V_y, V_z), the azimuth and elevation angle (θ, ϕ), time, and wavelengths. Traditional images and videos are just 2D and 3D special cases of the plenoptic function. In principle, one can reconstruct any views in space and time if sufficient number of samples of the plenoptic function is available. Many IB representations have been proposed and they differ from each other in the amount of geometry information used [book review etc].

While there has been considerable progress recently in the capturing, compression and transmission of image-based representations [15, 16], most multiple camera systems are not designed to be movable so that the view-points are somewhat limited. Moreover, they may not able to cope with moving objects in large environment. Apart from many system design issues, there are also many important problems and difficulties in realizing movable IBR systems such as object tracking, video stabilization and enhancement, etc. This motivates us to study the design of a movable image-based rendering system based on a class of dynamic IBR called plenoptic videos [7-9, 11]. In particular, we developed an automatic real-time object tracking algorithm and utilized the motion information computed to adjust continuously the azimuth or rotation angle of the movable IBR system in order to cope with moving speakers. Due to imperfect tracking, the videos may appear shaky and new video stabilization technique is proposed to overcome this problem.

The paper is organized as follows: The design and development of the prototype movable plenoptic video system is described in Section II. The details of other important processing functions such as object tracking, video stabilization, compression and the application of the proposed system to the multiview audio-visual conference will be given in Section III. Finally conclusions are drawn in Section IV

II. THE PROPOSED MOVABLE IBR SYSTEM

As mentioned earlier, the key issue with the switch-mode approach is the proper selection of the switching threshold T between the two modes and the other related parameters. In this paper, a novel threshold parameter selection scheme is proposed based on the theoretical analysis proposed recently for the NLMS algorithm in Gaussian noise [10] by the authors. The selection of related parameters will also be discussed shortly after the performance analysis in the next section.

As mentioned previously, the proposed movable IBR system consists of a linear array of cameras mounted on an electrically controllable wheel chair so as to provide improved viewing freedom to users and ability to cope with moving objects in large environment. Figure 1 shows the movable IBR system that we have constructed. It consists of a linear array of 8 Sony HDR-TGIE high definition (HD) video cameras which is mounted on a FS122LGC wheel chair.

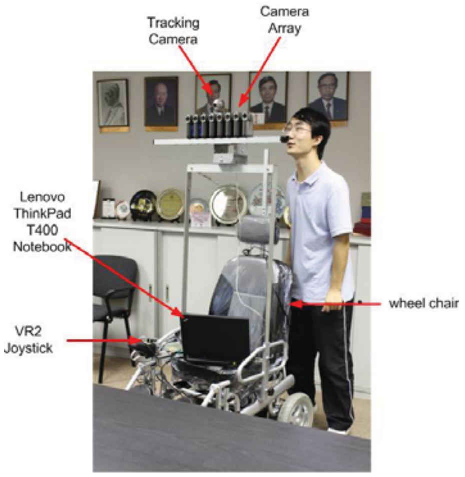


Figure 1. The proposed movable image-based rendering system.

The motion of the wheel chair is originally controlled manually through a VR2 joystick and power controller modules from PG drives technology. To make it electronically controllable, we examined the output of the joystick and generated the (x-,y-) motion control voltages to the power controller using a Devasys USB-I2C/IO micro-controller unit (MCU). Moreover, by using the wireless LAN of a portable notebook mounted on the wheel chair, we can also control its motion remotely.

The HD videos are captured in real-time into the storage cards of the cam-corders. For real-time streaming, it is rather difficult to compress online the HD videos and hence we employ a ThinkSmart IVS-MV02 Intelligent Video surveillance system [18] to compress online a lower resolution (320x240) 30 frames/sec videos. The videos can be retrieved remotely through the wireless LAN for viewing or further processing. The IVS-MV02 system is built from Analog Device DSP and can achieve real-time compression at a bit rate of 400kbps.

Before the cameras can be used for depth estimation, they must be calibrated. This can be accomplished by using a sufficient large checkerboard calibration pattern. We follow the plane-based calibration method [19] to determine the projective matrix of each camera.

Experimental Results

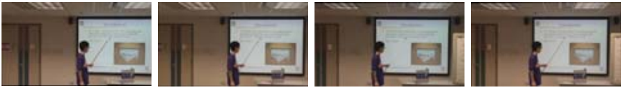
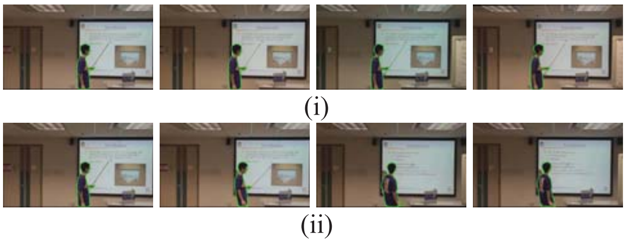


Figure 2(a). Snapshot captured by the movable IBR system at a given time instant.



(i)

(ii)

Figure 2(b). Segmented objects from the scene I) different views at the same time instant, and II) same view at different time instants.

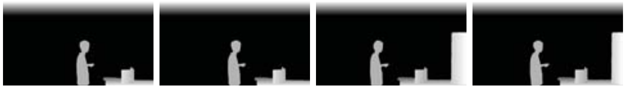


Figure 2(c). Depth maps computed at a time instant.

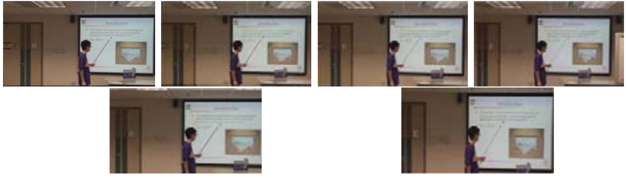


Figure 2(d). Upper: original cameras views 1 to 4. Lower: rendered views between (Left) cameras 1 and 2, and (Right) cameras 3 and 4. Note the rendered views are moved forward and away from the camera array.

Figure 2(a) shows the snapshots of the cameras taken at several time instants. Using an initial segmentation obtained by Lazy snapping [14], the object at other time instants and views are tracked using the level-set method [10]. Some tracking results are shown in Figure 2(b) where the speaker is tracked with the boundary marked in green color. The depth maps of each object are then estimated and are shown in Figure 2(c). Some renderings at other locations are also shown in Figure 2(d).

Next, we shall discuss the object tracking technique for steering the array in order to track a desirable moving object in large environment. Video stabilization technique to compensate for the undesirable shaking effects during tracking motion of the system will also be described.

III. OBJECT TRACKING, VIDEO STABILIZATION AND COMPRESSION

A. *Real-time object tracking*

In principle, the proposed system has two degrees of freedom. For simplicity, we only explore angular domain so that the complicated path planning problem of the movable IBR system can be avoided. Our tracking algorithm is based on the combination of the mean shift algorithm and the Kalman filter. At each frame, the Kalman filter is used to predict the object position, and mean shift algorithm is then used to obtain a more accurate position. The tracking starts by defining the object to be tracked by means of a user specified rectangular window in the screen. In the current implementation, a separate webcam is connected to a Lenovo ThinkPad T400 notebook computer for object tracking. Using the *x*-position of the object in the screen, one can generate a feedback signal to steer the wheel chair and the linear array angularly and position the object as close to the center of the screen as possible. In our current implementation, the tracking can be done completely in real-time in the ThinkPad T400 notebook computer.

Experimental Results

Figure 3 shows example tracking results of a moving object in a video conferencing application. It can be seen that the speaker can be satisfactorily tracked.

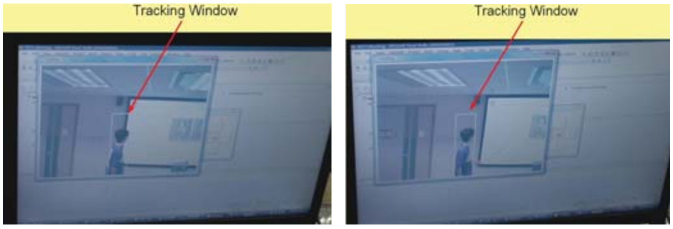


Figure 3. Example tracking results of a moving object at two time instants.

B. Video Stabilization

When the camera array is rotated either manually by the operator or automatically guided by the tracking algorithm, the video captured may be very shaky. To reduce this annoying effect, video stabilization can be employed. The basic idea of video stabilization is to estimate the global motion of the camera and then compensate for the undesirable motion.

In our system, the global motion is estimated by tracking feature points of the scenes. The Kanade-Lucas-Tomasi feature tracker [17] is employed. The histograms of the x- and y- velocities at frame 21 of these feature points are shown in Figure 4. It can be seen that there is a major peak in the histograms which correspond to the global motion. Small isolated peaks usually correspond to the features extracted from the speaker. Since majority of the feature points are coming from the background, all the feature points are used to compute an affine model for global motion. The linear translation computed from the affine model gives the final x- and y- velocities of the camera.

Figure 5(a) shows the extracted global motion over time (in blue color) and the smoothed global motion (in red color). Oscillations are observed, especially when the system is moving and about to settle down. To remove these oscillations and obtain a smooth motion path, we estimate the frequency of the oscillation using Kalman filter-based (KF) frequency tracking algorithm [20]. Using the AIC criterion, it was found that the order M is 3 and hence 2 frequency components should be used in tracking the x- and y- velocities. The tracking results are shown in Figure 5(b). It can be seen that the two high-frequency components (4-5Hz and 8-10Hz) of x and y are similar. They seem to come from the natural fundamental vibration frequency of the system and its 2nd harmonic. These two undesirable components can be removed by applying a time-varying adaptive notch filter to the original x- and y-velocities signals. More precisely, the two frequencies detected in y are used to construct a notch filter to filter out the oscillations in the x- and y-velocity signals. A 2nd-order IIR notch filter is employed twice to remove the fundamental and its harmonic with a Q factor of 1, i.e. the bandwidth of filter is around $f_{\text{notch}}/4$, where f_{notch} is the notch frequency of the filter. After that, the x- and y- velocity signals are further smoothed using a first order IIR filter with pole at 0.9. The smoothed velocity signals are shown in Figure 5(a) in red line. The smoothed velocities are used to modify the translational term of the affine model computed previously. We found that the rotational parameters are quite stable and hence their values are not stabilized. Using this affine model between consecutive images, full-frame warps are performed on the original images to the filtered motion models so as to stabilize the videos.

After motion compensation, some part of the compensated images at the boundary may be missing. These missing areas can be filled by image or motion inpainting techniques. For simplicity and avoid different inpainting algorithms from affecting the compression results, we simply reduce slightly the resolution of the video to avoid this problem.

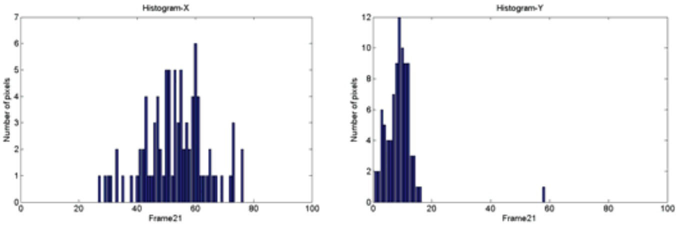
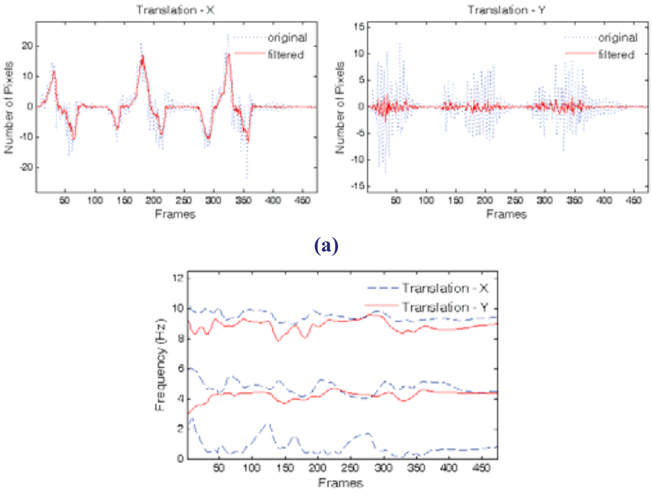


Figure 4. Histograms of the x- and y- velocities at frame 21.



(a)

(b)

Figure 5. Motion estimation and stabilization results: (a) extracted global motion over time (in blue) and smoothed motions after adaptive notch filtering and 1st order recursive smoothing (in red), (b) frequency components extracted from the x- and y- velocity signals by the KF-based frequency tracker.

C. Compression and multiview conferencing

For captured plenoptic videos, the multiple videos can be compressed offline using the object-based coder we have proposed in [12,13]. It is based on the MPEG-4 coder and the picture frame structure is shown in Figure 6. It employs prediction in both temporal and spatial directions. For simplicity, only three video object (VO) streams are shown.

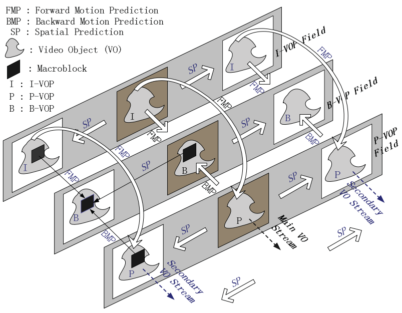


Figure 6. Picture frame structure and basic coding method for the texture coding of an IBR object in a PV.

Experiment results

The performance of the proposed system is now evaluated. The multiview video above is down-sampled to a resolution of 4-CIF. The frame-based coding mode is employed because it is more suitable for video conferencing applications. To explore the spatial redundancy in images from adjacent views, three videos are encoded in a group as showed in Figure 6 and only P-pictures are employed. The reconstructed peak signal-to-noise ratios (PSNR) of the original and stabilized videos versus the averaged bit rate per stream are plotted in Figure 7. It can be seen that due to reduced motion of the stabilized videos, the coding performance is slightly better than the original ones. For simplicity, the audios are not compressed.

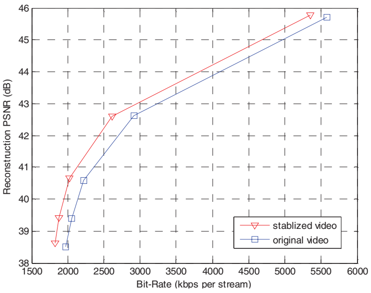


Figure 7. Coding performance of the original and stabilized multiple videos.

IV. CONCLUSION

The design and construction of a movable image-based rendering system and its associated video processing algorithms were presented. Its effectiveness is demonstrated using a multiview audio-visual conferencing application with a multiview TV display. A real-time object tracking algorithm is implemented and is utilized to adjust continuously the azimuth angle of the system in order to cope with moving objects. A new video stabilization technique based on the estimation of the vibration velocities is developed to overcome the problem of imperfect tracking and mechanical vibration of the system. The system developed serves as a framework for designing movable IBR systems with improved viewing freedom.

REFERENCES

[1] E. H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, pp. 3-20, MIT Press, Cambridge, MA, 1991.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The lumigraph," in *Proc. of the annual conference* on *Computer Graphics (SIGGRAPH'96)*, pp. 43-54, Aug. 1996.

[3] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. of the annual conference on Computer Graphics (SIGGRAPH'96)*, pp. 31-42, Aug. 1996.

[4] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proc. of the annual conference on Computer Graphics (SIGGRAPH'95)*, pp. 39-46, Aug. 1995.

[5] S. Peleg and J. Herman, "Panoramic mosaics by manifold projection," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pp. 338-343, June 1997.

[6] H. Y. Shum and L. W. He, "Rendering with concentric mosaics," in *Proc. of the annual conference on Computer Graphics (SIGGRAPH'99)*, pp. 299 - 306, Aug. 1999.

[7] S. C. Chan, K. T. Ng, Z. F. Gan, K. L. Chan and H. Y. Shum, "The plenoptic videos: capturing, rendering and compression," in *Proc. of IEEE Int'l Symposium on Circuits and Systems* (*ISCAS'04*), vol. 3, pp. 905-908, Vancouver, Canada, May 23-26, 2004.

[8] S. C. Chan, K. T Ng, Z. F. Gan, K. L. Chan and H. Y. Shum, "The plenoptic videos," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1650-2659, Dec., 2005.

[9] Z. F. Gan, S. C. Chan, K. T. Ng and H. Y. Shum, "An object-based approach to plenoptic videos," in *Proc. of IEEE Int'l Symposium on Circuits and Systems* (*ISCAS'05*), pp. 3435-3438, May, 2005.

[10] Z. F. Gan, S. C. Chan, and H. Y. Shum, "Object tracking and matting for a class of dynamic image-based representations," in *Proc. IEEE AVSS'2005.*

[11] S. C. Chan, Z. F. Gan, K. T. Ng and H. Y. Shum, "An Object-Based Approach to Image/Video-Based Synthesis and Processing for 3-D and Multiview Televisions," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 6, pp. 821-831, June 2009.

[12] Q. Wu, K. T. Ng, S. C. Chan and H. Y. Shum, "On object-based compression for a class of dynamic image-based representations", in *Proc. ICIP*'2005.

[13] K. T. Ng, Q. Wu, S. C. Chan and H. Y. Shum, "Object-Based Coding for Plenoptic Videos," to appear in *IEEE Trans. Circuits and Systems for Video Technology*.

[14] Y. Li, J. Sun, C. K. Tang and H. Y. Shum, "Lazy snapping," in *Proc.in SIGGRAPH'04*, pp.303-308, 2004.

[15] H. Y. Shum, S. C. Chan and S. B. Kang. Image-based rendering. Springer, 2007.

[16] S. C. Chan, H. Y. Shum, and K. T. Ng, "Image-based rendering and synthesis: technological advances and challenges," IEEE Signal Processing Magazine: Special Issue on MVI and 3DTV, Nov, 2007.

[17] J. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm", Microprocessor Research Labs, Intel Corporation.2005.

[18] www.ivs-tech.com

[19] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 11, pp. 1330-1334, 2000.

[20] Z. G. Zhang, S. C. Chan and K. M. Tsui, "A Recursive Frequency Estimator Using Linear Prediction and A Kalman Filter-Based Iterative Algorithm," *IEEE Trans. Circuits Syst. II*, vol. 55, issue 6, pp. 576-580, June, 2008.