Selection of Temporal Aligned Video Frames for Video Stitching Application

T. Luo, R. H. Y. Chung, and K. P. Chow
Department of Computer Science, The University of Hong Kong
Email: {tluo, hychung, chow}@cs.hku.hk

***Abstract*—Multi-view image/video stitching algorithm is an extensive research area in computer vision and image based rendering. Most researches focus on stitching the images from different views with assumption that those images have been already aligned in temporal domain. However it is not the case in real application. If the images from different views are not aligned in temporal domain, or in another words, not time synchronized, the corresponding feature points or regions will not be located correctly among different views, which will result in ghost objects appearing in the final stitching/rendering result. In this paper, we present an epipolar geometry consistency scoring scheme to guide temporal aligned video frame pair selection for multi-view video stitching application. Essentially, the proposed scheme allows us to determine whether a given pair of video frames is temporally aligned well for video stitching. Experimental results confirm that better video stitching results can be obtained with the proposed scheme in place.**

***Index Terms*—video stitching, temporal aligned selection**

I. Introduction

Nowadays, the prevalence of IP video surveillance system is expanding rapidly. Compared with last two generations surveillance systems, namely analog video tape recorder, and digital video recorder, IP system provides more flexibility and extendibility. Under certain scenarios, like public squares, train station, airport, video surveillance system users are more inclined towards multi-view systemdue to the limitation of single camera's field of view (FOV). In other words, some system prefers to use multiple cameras to monitor the same place from different viewpoints. In addition to that, a panorama video may be stitched according to the videos from different viewpoints for more intuitive situation awareness for operators. Many good results have been reported on static image stitching like [1]. However, real time video stitching applications have to address more challenges than static image stitching, such as time synchronization on video frames captured and parallax on moving objects.

Manuscript received October 18, 2013; revised December 5, 2013.

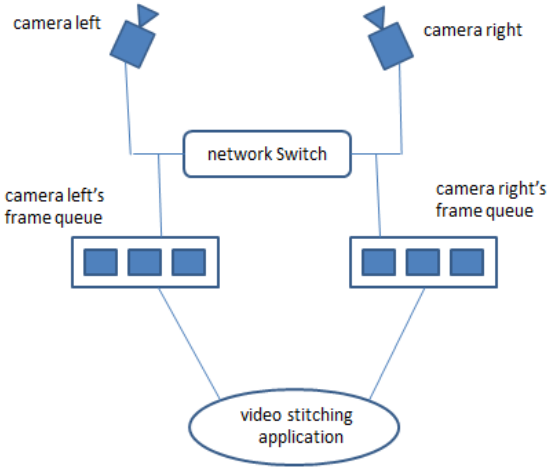


Figure 1. System for multi-view video stitching

Most researches on multi-view video system, as in [2], [3],have assumed that the videos are captured in a time-synchronized fashion, which are usually achieved by specialized hardware or equipment. However it is not always practical in real deployment scenarios. In a general IP video surveillance system, the encoded video frames are streamed from different edge devices like IP cameras or video encoders, transported via IP network, received by backend server and waited in queue for analysis, as illustrated by Fig. 1. Through a time server, the time of video edge devices (IP cameras or video encoders) could be synchronized down to second level, but this is not enough as there are typically 25 frames per second (FPS) in PAL signal, 30 FPS in NTSC signal, a slight difference in time reference could result in substantial temporal mis-alignment of video frames. Much worse, due to the inherent characteristics of network packet transmission, the video frame could be dropped or delayed, which further complicates the temporal frames alignment issue. With poorly temporally aligned video frames, moving objects will result in the ghosting effect on the stitching result. Thus, we are motivated to devise a new method to align video frames from static cameras in temporal domain.

In this paper, we introduce a new method to evaluate the goodness of temporal alignment between two frames. Essentially, the goodness is expressed in terms of a score, which indicates the level of disparity between epipolar geometries for a given video frame pairs. This score is

doi: 10.12720/joace.2.4.367-370

called EGCS (epipolar geometry consistency score), where a higher score indicates larger disparity between epipolar geometries from a pair of video frames. As a result, the problem of selecting the temporal aligned video frames could then be transformed into the problem of selecting the frame pair with minimum EGCS based on the following principles:

- The ground truth of all the cameras' epipoles are firstly registered to make sure the EGCS is zero when the static scene is monitored.
- Video frame pairs are built from the frames queued from different viewpoints when moving objects are detected. As illustrated in Fig. 2, frame pairs consist of frames from left and right camera. And only the video frame pair with minimum EGCS will be regarded as best temporal aligned pair for subsequent video stitching.

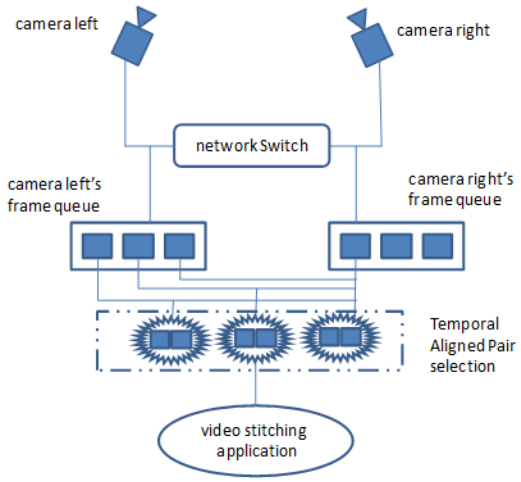


Figure 2. Video frame pair seleted from frame queue with mimimum EGCS

This paper is organized as follows: Section II introduces epipolar geometry and EGCS, Section III describes the proposed video temporal alignment method, followed by Section IV which presents the results on a stereo system. Finally, section V presents conclusions and discusses future research topics.

II. Epipolar Geometry

A. Epipolar Geometry Contraint on Multiple Views

Epipolar geometry is the geometry constraint between multi-view images, which is the foundation for the research on multi-view image.

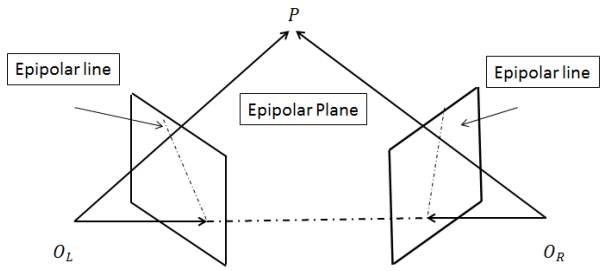


Figure 3. Epipolar geometry

The image in one camera of the projection center of the other camera is called epipole. It could be derived from the corresponding points on the images [4], illustrated in Fig. 3.

The epipole positions are determined only by the camera physical locations. For the static camera, the epipoles will not change with view contents.

B. EGCS (Epipolar Geometry Consistency Score)

Let $eP_{i,k}(x, y)$ be the ground truth epipole positions of *i*-th view projected on *k*-thviewpoint (illustrated in Fig. 4), in which (x, y) is image coordinate on the frame, whereas $CeP_{i,k,q}(x, y)$ be the current epipole positions calculated from *q*-th frame pair, which consists of frames queued from *i*-th viewpoint and *k*-th viewpoint (illustrated in Fig. 8). The epipolar geometry constrain score is define as the normalized sum of the Euclidean distance between the estimated epipole and ground truth on each view, in which $Width$ and $Height$ are resolution of captured frames.

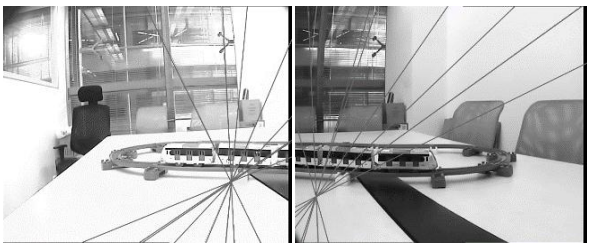
$$EGCS = \frac{\sum_{i,k}|CeP_{i,k,q}(x,y) - eP_{i,k}(x,y)|}{\sqrt{Width \times Height}} \quad (1)$$

When the view is static, there is no moving objects in the scene, $EGCS$ should be zero. Under this setting, it is not necessary to do temporal alignment. On the contrary, when moving objects appear, the image pairs are built from queued video frames. The image pair with minimum $EGCS$ will be regarded as temporal aligned and used for post processing.

III. Proposed Method

A. Ground Truth Epipoles Registration

Firstly the SIFT features (Scale-invariant feature transform) is detected [5] on static objects. Applying KLT iterative matching procedure, the corresponding features on different views are associated [6]. With RANSAC algorithm [7], the epipoles are determined, illustrated in Fig. 4.



(a) (b)

Figure 4. Epipoles ($eP_{i,k}(x, y)$) registration from different views. The epipole is the intersection point of epipolar lines. (a) epipole of left camera, (b) epipole of right camera

B. Epipoles Estimation

When motion objects are detected, the motion regions are segmented with speedy algorithm suggested in [8], illustrated in Fig. 5.

Similar to the ground truth epipoles registration procedure, the pair's epipole is estimated by

RANSACing the corresponding SIFT points from motion region.

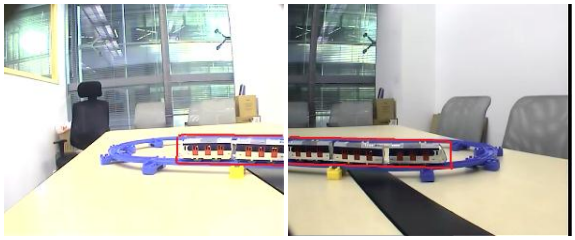


Figure 5. Motion Regions are segmented and highlighted with bounding box.

C. The Pair with Minimum EGCS Selection

Using equation (1), $EGCS$ of three image pairs consists of queued consecutive three frames are calculated. The pair, which has the minimum EGCS whose value is less than a pre-defined threshold, will be regarded as good enough video frame pair for video stitching. Otherwise the frame pair will be discarded.

IV. Experiment and Discussion

To test the method proposed in the paper, a video stitching system monitoring train platform is build up. It consists of the following hardware:

- Two static analog cameras
- Two digital video encoders, the video is compressed in H264 format with 704x576 resolution and 25 FPS
- A network switch to connect the digital video encoders and PC

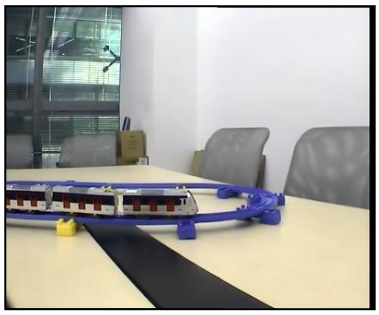
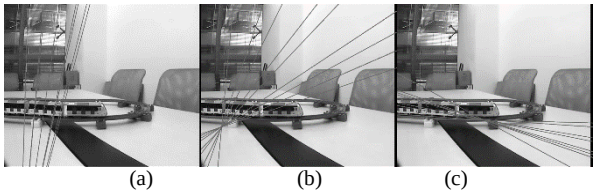


Figure 6. One frame captured by right camera

(a) (b) (c)

Figure 7. Paired with Fig. 6, left camera's three consecutive candidate frames for $EGCS$ evaluation

Fig. 6 is one frame captured from right camera, when the train is moving. Fig. 7 shows consecutive frames captured from left camera with the same timestamp as Fig. 6, in which the timestamp's precision is only down to second level. Fig. 8 shows intermediate results, the estimated epipole position according to different frame pairs between Fig. 6 and each of Fig. 7(a)-(c). Table I summarizes the EGCS score results of individual pairs.

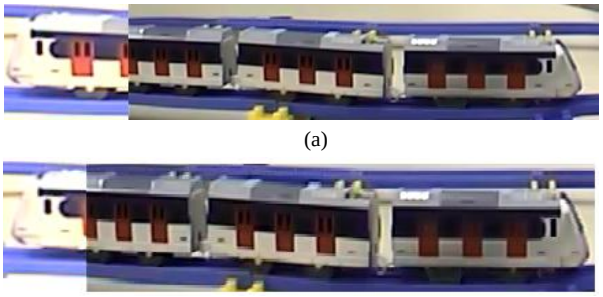


(a) (b) (c)

Figure 8. Estimated epipole positions according frame pairs (a) Fig. 6 & Fig. 7a, (b) Fig. 6 & Fig. 7b, and (c) Fig. 6 & Fig. 7c.

TABLE I. EGCS of Frame Pairs

Reference Frame	Pre-defined minimum EGCS: 0.2		
	Candidate Frames	*EGCS*	*Aligned*
Fig. 6	Fig. 7a (epipole shown Fig. 8a)	0.642	No
	Fig. 7b (epipole shown Fig. 8b)	0.057	Yes
	Fig. 7c (epipole shown Fig. 8c)	0.417	No



(a)

(b)

Figure 9. Video stitching result on the train (a) without temporal alignment. the left three doors carriage is stitched to a 4 door one. And the rail is miss aligned. (b) with temporal alignment. Both the train and rail are stiched well.

With a general PC, Intel i7 2670QM (2.2GHz), RAM 4G, the video temporal alignment and video stitching could be conducted in real time with dual video streaming in 704x576 resolution and 25 FPS. The video is stitched with procedure suggested in [9]. Fig. 9 presents the video stitching results, without temporal aligned frame pairs, the three doors train carriage is missed stitched as a four doors one and rail is miss aligned. However with temporal aligned frame pairs, both train carriage and rail are stitched better.

V. Conclusions

In this paper, a best fitting based on epipolar geometry consistency on multiple view method is proposed to guide video frames alignment in temporal domain before conducting video stitching. According to our experimental results, notable improvements on video stitching can be observed. Future directions will be focused on working out a frame temporal interpolation scheme if no temporal aligned video frame could be found.

References

[1] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems review," in *IEE Proc. Vision, Image and Signal Process*, vol. 152, no. 2, pp. 192-204, April 2005.
[2] C. Hou, H. Ai, and S. Lao, "Multiview pedestrian detection based on vector boosting," in *Proc. 8th Asian Conference on Computer Vision*, vol. 1, 2007, pp. 210-219.
[3] S. Denman, C. Fookes, J. Cook, C. Davoren, A. Mamic, *et al.*, "Multi-view intelligent vehicle surveillance system," in *Proc. IEEE International Conference on Video and Signal Based Surveillance*, 2006.
[4] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2004.
[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[6] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*, 1994.
[7] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, pp. 381-396, 1981.
[8] R. H. Y. Chung, F. Y. L. Chin, K. Y. K. Wong, K. P. Chow, T. LUO, and H. S. K. Fung, "Efficient block-based motion segmentation method using motion vector consistency," in *Proc. IAPR Conference on Machine Vision Application*, Tsukuba Science City, Japan, May 16-18, 2005.
[9] T. Panarungsun, S. Auethavekiat, and D. Gansawat, "Foreground rejection for parallax removal in video sequence stitching," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, 2011.

T. Luo received his MPhil degree from Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology. Currently he is working toward the PhD degree. His research interests include video analytics on video surveillance system.

Ronald H. Y. Chung received his BEng in Computer Engineering, MPhil and PhD in Electrical and Electronics Engineering in 1997, 1999 and 2003, respectively, from The University of Hong Kong.Dr. Chung has been with CS department in HKU since 2002, where he is now serving as an Honorary Assistant Professor. Apart from teaching, Dr. Chung is also involved in a number government and industry supported research projects.

K P. Chow received his PhD degree from UC Santa Barbara. In the recent years, Dr. Chow' research interests have migrated to digital forensics and computer security, and is the leader of the Computer Forensics Research Group (CFRG).