Walking into Panoramic and Immersive 3D Video

Yingbin Nie and Jianmin Jiang(✉)

Research Institute for Future Media Computing, School of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
nybl992@gmail.com, jianmin.jiang@szu.edu.cn

Abstract. To enable viewers to perceive the video content as if he or she is walking inside the video scenes, we need to have two essential video technologies. One is to present the audience with panoramic videos with 360°, and the other is view-adaptive video playback, i.e. presenting video scenes in accordance with the change of viewing angles. For the first technology, we propose a fast video stitching algorithm via exploiting the audio information for frame synchronization, and for the second, we propose a Fake-3D and True-3D mix method to immerse viewers inside the video scene via its dynamic playback of panoramic videos, adaptive to multi-view changes. Our proposed technologies have great potential in practical applications, such as virtual reality gaming and new concept movie shows etc.

Keywords: Panoramic 3D videos · Free-view video processing · Dynamic video playback · Video stitching · Fake-3D · True-3D · Virtual reality

1 Introduction

At present, video production and consumption is limited to the principle that video content is processed and produced for people to watch. In other words, video capture of live events is limited to camera views, and most of the video entertainment only provides people with one single view, facilitated by one or more cameras [7, 8]. As a result, the viewing experience achieved by such a see-what-camera-captures is not comparable with those on-spot views with our naked eyes. On spot, we can move our head and change the viewing angle to enjoy free-view of the event. We could even walk around to select the best possible viewing position, including the adjustment of viewing angles and distance to achieve the most effective viewing experience. An illustrative scenario is watching a show inside an opera house, where somebody may wish to walk onto the stage and enjoy watching the show from the back. In addition, on-spot view also provides the advantage that the audience can interact with relevant people to experience the event rather than watch. This is not achievable with current video technology, although research and technology innovation is active around areas of immersive video, multi-view video, and virtual reality [9, 10]. In this paper, we describe a new 3D panoramic and immersive video processing tool, which pioneers the concept to allow viewers not only watch the video content, but also walking into the video and experience the scene, the event, and the atmosphere as if he or she is on spot.

L.A. Maglaras et al. (Eds.): INISCOM 2016, LNICST 188, pp. 135–142, 2017.
DOI: 10.1007/978-3-319-52569-3_12

2 Panoramic and Immersive 3D Video Constructions

To pioneer the concept of allowing viewers to walk into videos and perceive the content on real-time basis, we need to develop a number of new technologies and overcome three fundamental hurdles. These include: (i) creating panoramic video via capture of the scene with 360° along both horizontal and vertical directions; (ii) presenting and playback the video content in accordance with the change of viewing angles and body movements; (iii) interacting with video objects and participating in activities with instant and natural responses. To overcome the first hurdle, we configured one row of multi video cameras to ensure sufficient video capture for 360° along the horizontal direction. To overcome the second hurdle, we use VR helmet display unit and exploit its accelerometer sensor and magnetic sensor output to control the video playback corresponding to the view or pose changes.

2.1 Multi-camera Synchronization

Panoramic video need multi-camera to capture multi angle perspective, we can't promise all cameras to start recording in the same time, so that these videos captured by multi cameras need to be synchronized. We use the Chroma feature [1] as a descriptor for the audio content of multi video streams in different views, which gives a 12-dimensional representation of the tonal content of an audio signal derived by combining bands belonging to twelve pitch classes (C, C#, D, D#, E, F, F#, G, G#, A, A#, B), corresponding to the same distinct semitones, and use method [11] to synchronization.

Firstly, decompose the audio signals from the multi-camera captured videos and cut a segment starting from the pth frame to the qth frame for all audios and every audio has overlap in this segment. Then compute Chroma features for each segment, where C_n is used to represent the nth camera's audio, F_n the nth camera's audio segment, and v_n the Chroma feature of F_n.

To compute the time-shifts between pairs of recorded videos, we calculate the Euclidean distance, $d_{ij} = E(v_i, v_j)$, between their Chroma features F_i and F_j [2] to derive a distance matrix. Figure 1(a) shows the distance matrix d_{ij} for two feature vectors obtained from two camera recorded segments, each of them has 2 s duration. The distance matrix d_{ij} contains information about the feature matching of the two audios. In order to interpret this information, the point of minimum distance across each row of the distance matrix d_{ij} is calculated. As seen in Fig. 1(a), the distance matrix d_{ij} is a rectangular matrix, in which the main diagonal corresponds to zero time-shift, and the diagonal above or below the main diagonal correspond to positive and negative time-shifts, respectively. We calculate the matching histogram $v_{ij}(\Delta t)$ for C_i and C_j from the distance matrix d_{ij} for the count of the number of minimum distances along each diagonal as shown in Fig. 1(b). As seen, while Δt = 1.00 s, the count reached the maximum, and hence F_i starts one second earlier than F_j, which mean that C_i starts one second earlier than C_j. This is also shown as the sub-diagonal in the 1 s position in Fig. 1(a).

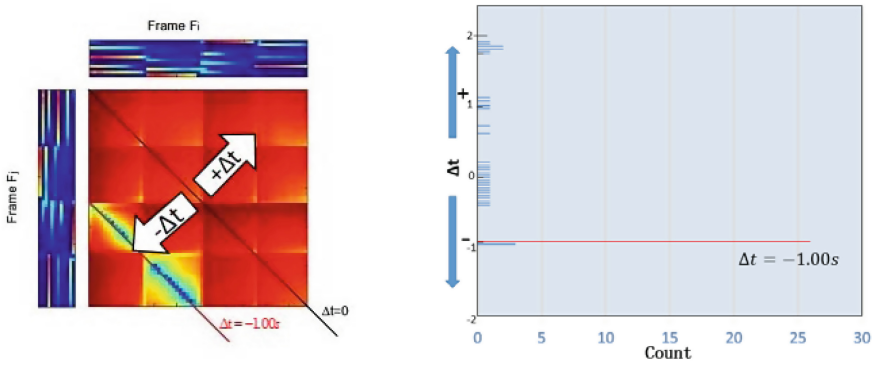


Fig. 1. Illustration of Chroma feature matching for computing the time-shifts between pairs of recorded videos: (a) Using the distance matrix for two camera recorded segments with the duration of 2 s. (b) The minimum across each row is calculated and the count of minimum distances is accumulated across each diagonal to give the histogram, and its peak corresponds to the time-shift.

For each video, this method can find time-shift and adjusting to the same point in time to start stitching video frames.

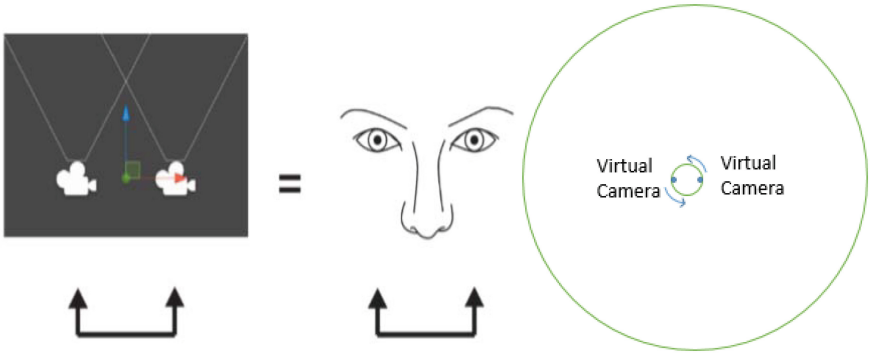
2.2 Panoramic Video Creation

While image stitching has been extensively researched for the past decades [2–6], our investigation reveals that dedicated design of the technique and new elements are still needed for video stitching since: (i) video stitching requires synchronization of frames to guarantee that right frame pairs are stitched together across the entire sequence; (ii) video frame stitching requires low computing cost and high speed, and thus a number of video-unique features need to be exploited. To this end, we introduced two changes into the image stitching method reported by Brown and Lowe [2], which include: (i) considering the calibration of video cameras beforehand, we limit the SIFT extraction and correspondence within the half frame area as shown in Fig. 2; (ii) to improve the stitching effectiveness, we introduce an additional alignment procedure by exploiting the coordinate transformation between the left and right frames to be stitched.

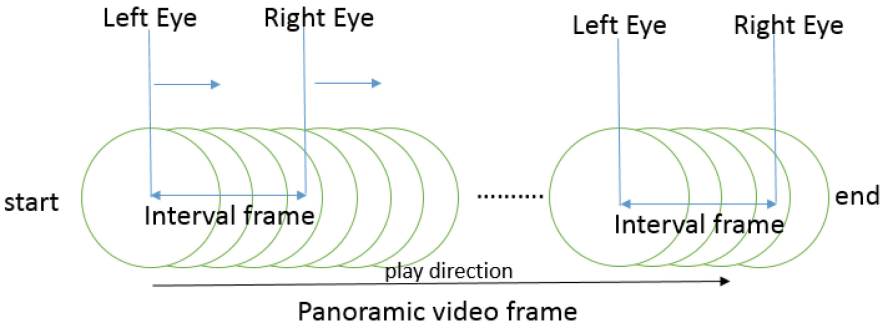
Specifically, given the perspective transformation as follows:

$$[x', y', w'] = [u, v, w] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{1}$$

(a)



(b)



(c)

Fig. 2. Illustration of Pseudo-3D method and True-3D method in watching panoramic video: (b) Pseudo-3D method using distance of virtual camera in OpenGL to simulate pupillary distance and change virtual camera location according viewer head orientation; (c) True-3D using the *t*th frame from panoramic video to replace scene of left eye and right eye in fixed angle.

and the transformed coordinates:

$$x = \frac{x'}{w'} = \frac{a_{11}u + a_{21}v + a_{31}}{a_{13}u + a_{23}v + a_{33}} \tag{2}$$

$$y = \frac{y'}{w'} = \frac{a_{12}u + a_{22}v + a_{32}}{a_{13}u + a_{23}v + a_{33}} \tag{3}$$

We select a group of N registered pair of SIFT points, $\{(x_L^1, x_R^1), (y_L^1, y_R^1); \ldots (x_L^N, x_R^N), (y_L^N, y_R^N)\}$, to verify the perspective transformed coordinates via the following:

$$\Delta_x = \frac{\sum_{i=1}^{N} \left(x_L^i - x_R^i\right)}{N} \tag{4}$$

$$\Delta_y = \frac{\sum_{i=1}^{N} \left(y_L^i - y_R^i\right)}{N} \tag{5}$$

and the additional alignment is done by:

$$(x, y) = \begin{cases} \left(x + \Delta_x, y + \Delta_y\right) & if\ |\Delta_x| \geq T\ \cap \left|\Delta_y\right| \geq T \\ \left(x, y + \Delta_y\right) & if\ |\Delta_x| < T\ \cap \left|\Delta_y\right| \geq T \\ \left(x + \Delta_x, y\right) & if\ |\Delta_x| \geq T\ \cap \left|\Delta_y\right| < T \\ (x, y) & if\ |\Delta_x| < T\ \cap \left|\Delta_y\right| < T \end{cases} \tag{6}$$

Where T stands for a threshold, which is determined empirically.

2.3 View-Change Adaptive Playback of Videos

Panoramic video playback needs special treatment in order to simulate the on-spot view changes. Using OpenGL, making panoramic video as a texture, and stick it to a sphere model as shown in Fig. 2(a), and in this way, we are able to create 360° panoramic video effect via the video stitching algorithm described earlier.

It is proved difficult to create 360° 3D panoramic videos due to the fact that (i) the requirement on video frame resolutions leads to processing of huge information quantity; (ii) 3D video stitching requires excessive computing cost. To this end, we propose a pseudo-3D approach to complete the creation of 3D panoramic 360° videos, details of which are described below.

Set two virtual cameras to simulate the viewer's position inside the Sphere as shown in Fig. 2(b). As the viewing point changes, the distance between the virtual left eye and the right virtual right eye will change depending on the value of yaw. Correspondingly, we use this changing distance to select the *t*th frame for the right eye and the current frame (the 1st frame) for the left eye. The specific algorithm is given as follows:

Algorithm:

I. Using d to represent the distance of virtual cameras (eyes), d dynamically changes according to the viewer's head position, and its initial value is $d = d_{\max}$. As the view changes, the distance between the two virtual cameras can be derived by:

$$d = abs(\cos(yaw)) \cdot d_{\max} \quad yaw \in [0, 2\pi] \tag{7}$$

II. In True-3D, let t represents the number of frames controling the selection of the frame for the right eye, we have:

$$t = \text{int}(\sin(yaw) \cdot t_{\max}) \quad yaw \in [0, 2\pi] \tag{8}$$

Where t_{max} is the initial value, which can be determined as zero when yaw = 0. In the proposed algorithm as described above, the value of yaw can be obtained from the orientation sensor in VR helmet such as oculus, or the mouse position etc.

3 Experiments

To evaluate the proposed system, we carried out extensive experiments in two phases. While the first phase of experiments is to test the performance of creating the panoramic videos, the second phase of experiments is to evaluate the effectiveness of VR-helmet adaptive display of the panoramic video and the experience of walking into

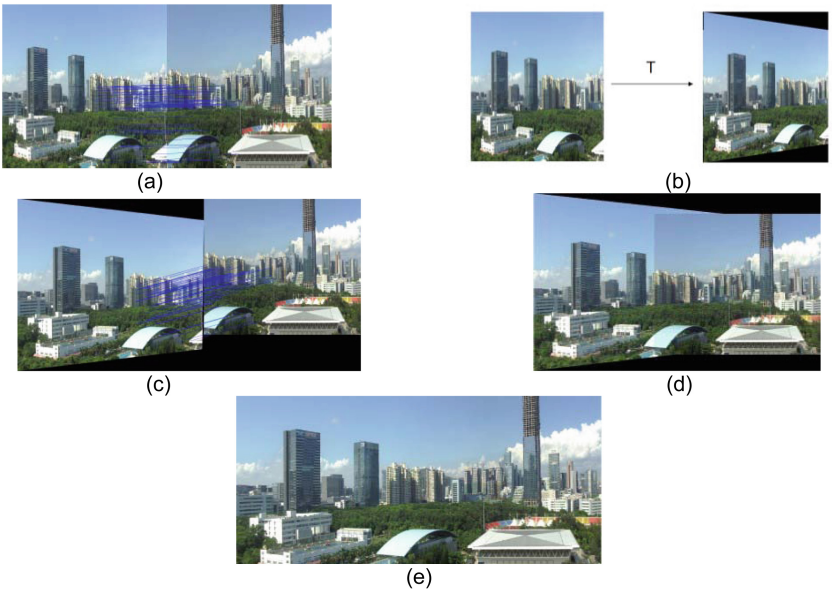


Fig. 3. Illustration of experimental results for creating panoramic videos

the video scene. Figure 3 illustrates the experimental results and the sequence of the panoramic 3D video creation process. As seen, part (a) illustrates the SIFT point extraction and matching process, Part (b) illustrates the effectiveness of perspective transformation, part (c) and (d) illustrate the SIFT point verification and additional alignment, which shows that better stitching quality is achieved by such additional alignments, especially after the fusion as shown in (e). As the value of yaw changes from 0 to 360°, the video frame will be adaptively presented to the viewer (Fig. 4).

Fig. 4. Illustration of adaptive playback of the Pseudo-3D panoramic video: (a) a frame in panoramic video (b) when yaw = 1.4π, the scene mix True-3D and Pseudo-3D method to present video content to the viewers adaptively.

4 Conclusions

In this paper, we introduced a new concept of creating 3D panoramic videos via a fast video stitching algorithm to enable viewers to "walk into video content". We also described our proposed algorithm design for implementing such a new concept. Extensive experiments support that (i) our video stitching is fast and fit the purpose based on the simplifications introduced; (ii) our proposed adaptive pseudo-3D panoramic video playback achieves good 3D and immersive effect, pioneering the concept of immersive, panoramic, multi-view.

The authors wish to acknowledge the financial support for the work from CNSF (Chinese Natural Science Foundation) under the grant number 61373103.

References

1. McVicar, M., Santos-Rodríguez, R., Ni, Y., De Bie, T.: Automatic chord estimation from audio: a review of the state of the art. IEEE Trans. Audio Speech Lang. Process. **22**, 556–573 (2014)
2. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vis. (2007)
3. Tang, W., Wong, T.T., Heng, P.: A system for real-time panorama generation and display in tele-immersive applications. IEEE Trans. Mult. **7**, 280–292 (2005)
4. Burt, P., Adelson, E.: A multiresolution spline with application to image mosaics. ACM Trans. Graphics **2**, 217–236 (1983)
5. Xu, W., Mulligan, J.: Panoramic video stitching from commodity HDTV cameras. Multimedia Syst. **19**, 407–426 (2013)
6. Molina, E., Zhu, Z.: Persistent aerial video registration and fast multi-view mosaicking. IEEE Trans. Image Process. **23**, 2184–2192 (2014)
7. Song, X., Zhang, J., Han, Y., Jiang, J.: Semi-supervised feature selection via hierarchical regression for web image classification. Multimedia Syst. **22**(1), 41–49 (2016)
8. Zhang, J., Han, Y., Jiang, J.: Tensor rank selection for multimedia analysis. J. Vis. Commun. Image Represent. **30**, 376–392 (2015)
9. Li, J.Y., Jiang, J.: Nonrigid structure from motion via sparse representation. IEEE Trans. Cybern. **45**(8), 1401–1413 (2015)
10. Pan, Z.L., Ming, Z., Zhong, H., Wang, X., Xu, C.: Compressed knowledge transfer via factorization machine for heterogeneous collaborative recommendation. Knowl. Based Syst. **85**, 234–244 (2015)
11. Bano, S., Cavallaro, A.: Discovery and organization of multi-camera user-generated videos of the same event. Inf. Sci. **302**, 108–121 (2015)