CAMERA POSE ESTIMATION FOR MIXED AND DIMINISHED REALITY IN FTV

Hideo Saito, Toshihiro Honda, Yusuke Nakayama, Francois de Sorbier

Department of Information and Computer Science, Keio University, Yokohama Japan

ABSTRACT

In this paper, we will present methods for camera pose estimation for mixed and diminished reality visualization in FTV application. We first present Viewpoint Generative Learning (VGL) based on 3D scene model reconstructed using multiple cameras including RGB-D camera. In VGL, a database of feature descriptors is generated for the 3D scene model to make the pose estimation robust to viewpoint change. Then we introduce an application of VGL to diminished reality. We also present our novel line feature descriptor, LEHF, which is also be applied to a line-based SLAM and improving camera pose estimation.

Index Terms **—** camera calibration, augmented reality, free viewpoint image synthesis, feature descriptor, see-through vision

1. INTRODUCTION

Free Viewpoint TV is a framework to provide a functionality of controlling/manipulating observation viewpoints when the video is observed by users. One of the early studies for achieving such functionality of free viewpoint observation of videos was Virtualized Reality [1].

FTV is achieved by integrating a lot of computer vision/computer graphic technologies. One of the important technologies of computer vision for FTV is camera calibration, which is a technique to estimate camera poses, and optical parameters of cameras. In FTV, 3D structure of the object scene should be somehow captured. In the most of typical way to capture such 3D structure of the scene is recovering the 3D structure based using multiple viewpoint videos. For 3D recovery with multiple views, we should get geometry of all cameras that captures the scene, otherwise, it is almost impossible to geometrically merge the different view videos for 3D recovery.

In most of the early works of FTV [2,3], all the cameras for capturing the 3D structure of the scene are assumed to be not moving while the scene is captured, so that camera calibration can be performed before the capturing of the scene. However, such assumption limits applicability of FTV capturing, because we sometimes wish to move the cameras according to the scene change and object moving. Therefore, on-line estimation of camera pose is important for more flexible FTV implementation [4,5].

On-line camera pose estimation is also one of the most important technology for Mixed and Augmented Reality (MAR). One of the pioneer work of on-line camera pose estimation for MAR is Toolkit Markers [6]. It is easy-to-use toolkit for realizing on-line camera pose estimation. For avoiding to use such markers, feature point based camera pose estimation are also extensively studied, but a planar pattern should be placed in the region of interest.

We have proposed a method of camera pose estimation for MAR using 3D shape and texture of the scene [7]. The 3D pose estimation is based on viewpoint generative learning using 3D objects. By having a 3D shape model with surface texture, we virtually generate a number of images of the model from different viewpoints, and then select stable keypoints from those patterns. Our system learns a collection of feature descriptors from the stable keypoints. Finally, we can estimate the pose of a 3D object by using these robust features.

In this paper, the camera pose estimation using the 3D shape of the target scene based on Viewpoint Generative Learning (VGL)[8]. In VGL, a database of feature descriptors is generated for the 3D scene model to make the pose estimation robust to viewpoint change. Then we introduce an application of VGL to diminished reality. Next, we also present our novel line feature descriptor, LEHF [9], which is also be applied to a line-based SLAM and improving camera pose estimation[10].

2. CAMERA POSE ESTIMATION BY VIEW GENERATIVE LEARNING OF 3D MODEL

The camera pose estimation is based on view generative learning (VGL) [8]using 3D model of the object with texture. Basic idea of VGL is that a number of different viewpoint images of a target scene are learned for making the camera pose estimation robust to view point changes. For efficient learning of a lot of different viewpoint images, descriptors of keypoints (feature points) detected from the images are stored in a database of keypoint feature descriptors in advance. By synthesizing the different viewpoint images from a 3D shape model of the target scene, 3D position of each keypoint is also stored in the database. Here, we only store keypoints which are stably detected in different viewpoint images. We call those keypoints as stable keypoints. In the database of feature descriptors, we do not store all descriptors for each keypoint, but store only small number (k) of representative descriptor (for example, K=3) for reducing the data amount of the database and efficient search at the run-time. The representative descriptors are obtained by k-means clustering of all descriptors for each keypoint.

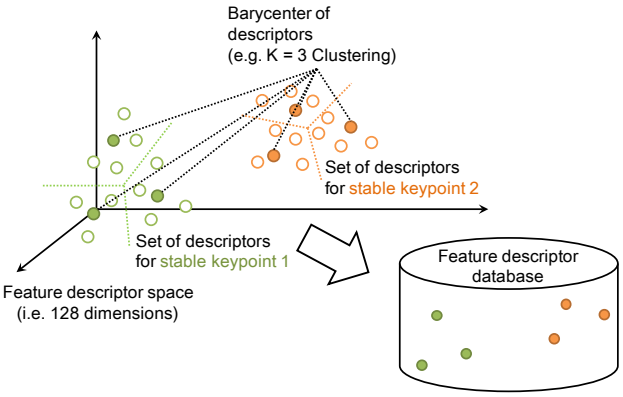


Figure 1. Clustering of Feature Descriptors in VGL

978-1-4799-4758-4/14/$31.00 ©2014 IEEE

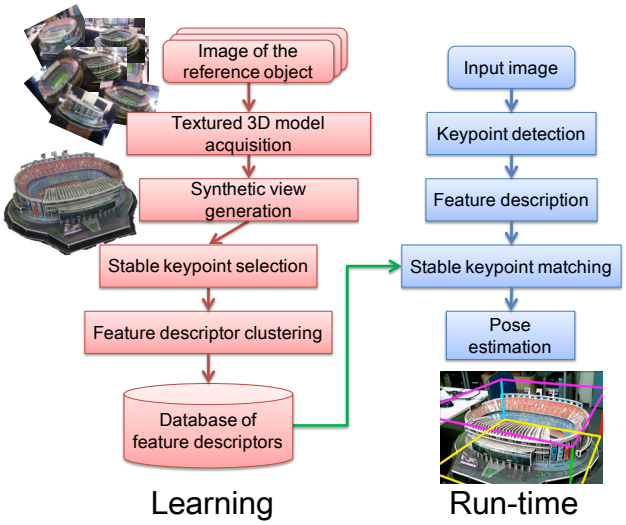
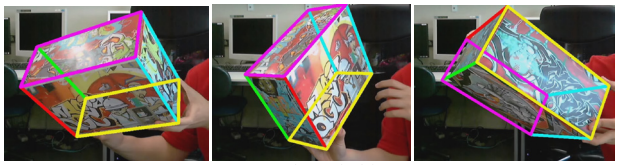
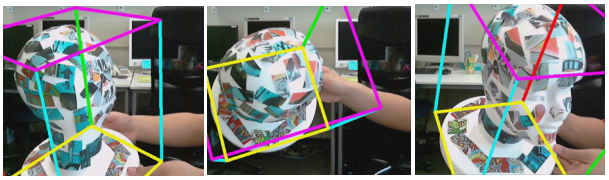


Figure 2. Algorithm of VGL for 3D pose estimation.



(a) Box shaped object

(b) Cup shaped object



(c) Head shaped object

Figure 3. 3D shape model recovered with KinectFusion.

In run-time of camera pose estimation, keypoints are detected for each frame of input video sequence. Each descriptor of each keypoint is matched with the descriptors of the keypoint feature descriptor database for obtaining the 3D position of the keypoint. We can expect that the matching can be robust to viewpoint changes because the different descriptors for each keypoints are stored in the database. The algorithm flow is shown in Figure 2.

Figure 3 shows example results of camera pose estimation using VGL. In those images, the camera is actually fixed, but the objects are moving, so we estimate the relative pose of the camera with the object. The bounding box shape of the object is overlaid for indicating the relative pose of the camera with the object. Those results are demonstrating that VGL can successfully estimate the pose without any limitation of the object pose. We do not use any boundary edges of the object in the image, but the keypoint matching for pose estimation. The results of the cup shaped object and the head shaped object demonstrate that our method can successfully estimate the pose without any edge information which is often used for 3D model based pose etimation [11].

3. DIMINISHED REALITY

Diminished Reality (DR) is a technique for visually removing a real object from an input image for making see-through image by replacing the removed object with a image of occluded area. For generating images of DR, the occluded area should be captured by other cameras placed at different viewpoint from the camera that takes the input image. Since the viewpoints of the other cameras that capture the occluded area are different from the input camera, the images of the occluded area should be transferred to the same viewpoint as the input camera for replacing the removed object in the input image. The viewpoint transfer can be achieved by applying FTV technologies, which synthesize different viewpoint videos from input videos. In this sense, DR can be regarded as an application of FTV.

For synthesizing different viewpoint images of the occluded area, 3D structure of the area should be given. Zokai et al. [12] reconstruct a 3D model of the occluded area by multiple cameras. Enomoto et al. [13] and Honda et al. [14] assume that the occluded area can be approximated as a single planar structure for achieving on-line DR system. Hashimoto et al. [15] approximate the occluded area with a set of planar structures for synthesizing a see-through baseball movie from multi-camera systems. Figure 4 shows an example image of the see-through observation of baseball match. In this case, the pitcher is approximated as a single plane, and the ground is also approximated as another single plane. Based on the structure, the image of the occluded area in the center camera is synthesized from the corresponding areas of left and right camera. A limitation of this method is that the cameras cannot be moved, but need to be placed at fixed positions.

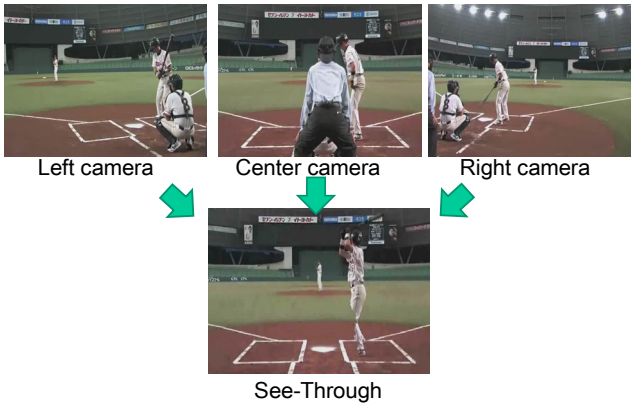


Figure 4. An example of the see-through observation of baseball match.

As indicated in those related works on DR, there are some research issues in DR, which are similar as FTV. The main issues are camera pose estimation for making free-viewpoint observation of the scene, and 3D structure recovery of the scene. For avoiding such difficulties, observing cameras are fixed in [15]. The scene is approximated as planar structures in [13,14,15]. AR marker is used for avoiding on-line camera pose estimation in [13]. Those issues should be solved for expanding DR applications.

4. ON-LINE DIMINISHED REALITY SYSTEM

In this section, on-line diminished reality system with free-viewpoint observation for non planar structure scene using VGL is presented. In this system, we capture 3D structure of occluded area using RGB-D camera. For achieving on-line camera movement for free viewpoint observation of DR with a smartphone, the pose of the camera in the smartphone is estimated by VGL. Figure 5 shows the overview of the DR system.

This system consists of a smartphone (hand-held PC with a camera) and a server PC. Those two devices are connected via wireless networks, so that video data can be transferred between the devices. In the server PC, 3D shape model of a target scene and its keypoint feature database are stored in advance. We apply KinectFusion [16] for recovering 3D shape model of a target scene by taking RGB-D video sequence of the target scene with Kinect. Figure 6 shows an example of 3D shape model recovered by KinectFusion. This model is represented by a set of triangle mesh with texture, so that the viewpoint of synthesizing images can arbitrarily be changed.

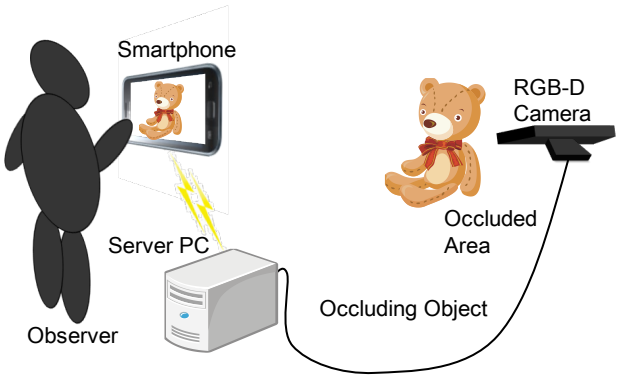


Figure 5. System setup for on-line diminished reality.

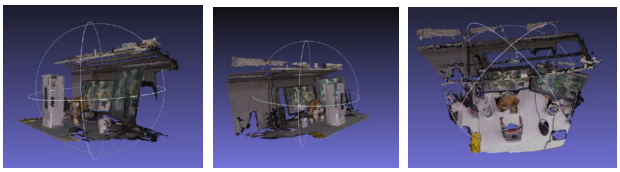


Figure 6. 3D shape model recovered with KinectFusion.

At on-line phase, the observer captures the scene, where the occluding object is hiding the occluded area, using the smartphone. By applying the pose estimation by VGL, the pose of the camera (smartphone) can be estimated on-line. Images of the occluded area that should be observed from the estimated viewpoint is synthesized based on the 3D structure that is also captured by the RGB-D camera. The image of the occluded area is overlaid onto the smartphone image for generating an image without the occluding object.

In diminished reality applications, we also need to detect the area of occlusion. Figure 7 (a) represents an example image of captured image with the smartphone. In this case, the box in the center of the image is hiding the scene behind the box, which can easily be recognized by human, but it is not easy to be detected by the system automatically. In the presented system, we make the user select occluding objects by touching the display of the smartphone. Figure 7(c) shows example of result image of diminished reality, in which the occluding object is removed from the input image, Figure 7(a). Figure 7 (b) shows an image captured from the same viewpoint as Figure 7 (a) without the occluding box-shaped object, which can be regarded as the image that we wish to synthesize by diminished reality techniques. Figure 7 (c) is almost same as Figure 7 (b), which is demonstrating the effectiveness of the presented system.

Figure 8 represents a DR visualization example, in which the occluded area is moving. In this case, the 3D shape of the occluded area is captured by a Kinect, which provides the image of the occluded area from the viewpoint of the observing smartphone.

(a) Input Image (b) Without occluding object (c) Generatedby our system

Figure 7. Diminished reality output image.

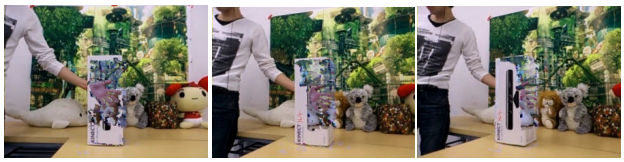
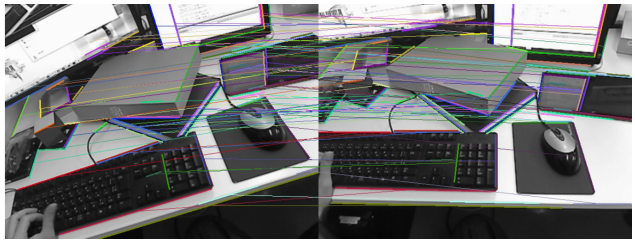


Figure 8. Example of diminished reality.

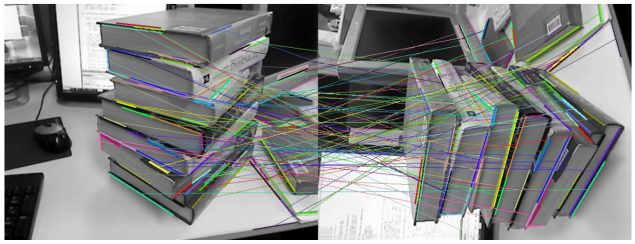
5. LINE FEATURES FOR CAMERA POSE ESTIMATION

In the previous sections, we have presented methods of camera pose estimation based on keypoints (feature points) captured in input image sequences. In input images, we can also consider line features as key-features for camera pose estimation. For taking into account such line features, we have already proposed a line feature descriptor called as Line-based Eight-directional Histogram Feature (LEHF) [9]. The line segments can be detected by using LSD [17], which is an efficient line feature detector. Figure 9 demonstrates the performance of LEHF. Almost all line features are correctly matched even though some lines segments are detected in different positions and orientations with different length.

Using the line feature descriptor LEHF, we have proposed a novel method for SLAM (Simultaneously Localization and Mapping). In this method, the line segments are detected by LSD from each frame of input image sequence, then matched with the line segments detected in the different frames according to the similarity of LEHF descriptors. Then 3D poses and positions of the detected line segments are estimated, while the pose of camera at each frame is also estimated simultaneously. Figure 10 shows an example of mixed reality visualization in which a 3D CG model is overlaid at the fixed position onto the desk.

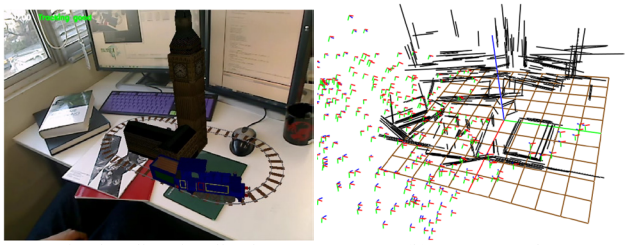


(a) Correct matching 73, wrong matching 0.



(b) Correct matching 80, wrong matching 3.

Figure 9. Example of line segments matching using LEHF.



(a) Mixed reality visualization (b) SLAM result

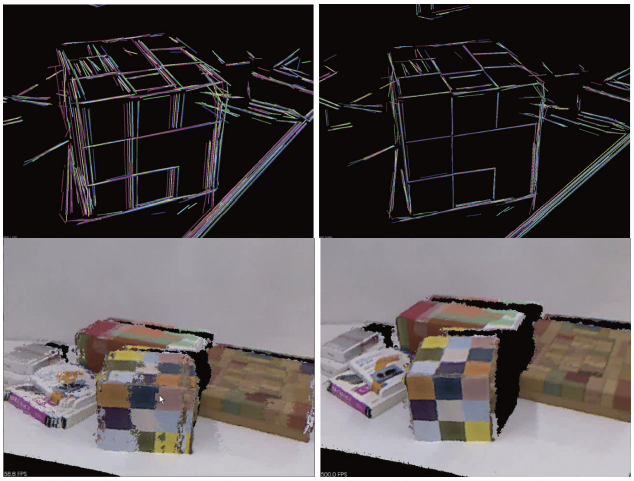
Figure 10. Example of mixed reality visualization by camera pose estimated by SLAM using line segment features.

We have recently proposed a novel method for improving accuracy of camera pose estimation based on the line features [10]. This method is designed for improving the accuracy of camera pose estimation of KinectFusion, which can provide 3D model of a target scene by capturing the scene using a RGB-D camera. In KinectFusion, the camera pose estimation is mainly performed by ICP algorithm, which aligns the camera pose between the different frames based on 3D point matching. However, 3D points captured by a RGB-D camera often affected by errors of the depth measurement. To solve this problem about point based alignment, we propose a method for alignment by using line segments. For the alignment based on line segments, we represent a 3D model of target scene with 3D line segments. Then 2D line segments in RGB images are matched with the 3D line segments in the 3D model using LEHF descriptor to obtain 2D-3D line correspondence.

6. CONCLUSION

Mixed and augmented reality, and diminished reality visualization are applications of FTV technology with a promising future. In those applications on-line camera pose estimation is a significant basic technology, in which 3D computer vision techniques can play a great roles. In this article, we introduced our recent challenges using VGL and line segment features, which make the on-line camera pose estimation stable and robust. Future work will be a hybrid use of line features and point features in VGL framework.

ACKNOWLEDGEMENTS: This work was partially supported by MEXT/JSPS Grant-in-Aid for Scientific Research(S) 24220004.



(a) Without improvement by LEHF (b) With improvement by LEHF

Figure 11. KinectFusion recovers the 3D model of the scene. (a) Original 3D model represented by line segments (upper row) and colored point cloud (lower row). (b) Improved 3D model recovered with accurately estimated camera poses using LEHF matching.

REFERENCES

[1] H. Saito, S. Baba, M. Kimura, S. Vedula, T. Kanade, "Appearance-Based Virtual View Generation of Temporally-Varying Events from Multi-Camera Images in the 3D Room," *Second International Conference on 3-D Digital Imaging and Modeling (3DIM99)*, pp. 516 - 525, 1999.

[2] J. Starck, A. Maki, S. Nobuhara, A. Hilton, T. Matsuyama, "The Multiple-Camera 3-D Production Studio," *IEEE Transactions on Circuits and Systems for Video Technology,* 19, 6, pp.856-869, 2009.

[3] M. Tanimoto, "FTV: Free-viewpoint Television," *Image Communication*, 27, 6, pp.555-570, 2012.

[4] S. Jarusirisawad, H. Saito, "3DTV View Generation Using Uncalibrated Cameras," *International Conference on 3DTV(3DTV-CON08),* pp.57-60, 2008.

[5] N. Hasler, B.Rosenhahn, T. Thormahlen, M. Wand, J. Gall, H.Seidel, "Markerless Motion Capture with Unsynchronized Moving Cameras," *CVPR2009*, pp.224-231, 2009.

[6] H. Kato, M. Billinghurst, "Marker Tracking and HMD Calibration for a video-based Augmented Reality Conferencing System," *International Workshop on Augmented Reality (IWAR 99),* pp. 85-94 , 1999.

[7] D. Thachasongtham, T. Yoshida, F. de Sorbier, H. Saito, "3D Object Pose Estimation using Viewpoint Generative Learning," *Scandinavian Conference on Image Analysis (SCIA 2013)*, LNCS 7944, pp. 512-521, 2013.

[8] T. Yoshida, H. Saito, M. Shimizu, A. Taguchi, "Stable Keypoint Recognition Using Viewpoint Generative Learning," *International Conference on Computer Vision Theory and Applications (VISAPP2013)*,2, pp.310 -315, 2013.

[9] K. Hirose, H. Saito, "Fast Line Description for Line-based SLAM," *British Machine Vision Conference 2012 (BMVC2012)*, pp. 83.1-83.11., 2012

[10] Y. Nakayama, T. Honda, H. Saito, M. Shimizu, N. Yamaguchi, "Accurate camera pose estimation for KinectFusion Based on Line Segment Matching by LEHF," *International Conference on Pattern Recognition (ICPR2014)*, accepted for presentation , Aug. 2014.

[11] T. Drummond, R. Cipolla, "Real-time tracking of complex structures with on-line camera calibration," *Image and Vision Computing,* 20, 5–6, pp. 427–433, 2002.

[12] S. Zokai, J. Esteve, Y. Genc and N. Navab, "Multiview Paraperspective Projection Model for Diminished Reality," *International Symposium on Mixed and Augmented Reality (ISMAR 2003)*, pp. 217 - 226, 2003.

[13] A. Enomoto, H. Saito, "Diminished Reality using Multiple Handheld Cameras," *ACCV'07 Workshop on Multi-dimensional and Multiview Image Processing,* pp.130-135, 2007.

[14] T. Honda, T. Inoue, H. Saito, "Real-Time Diminished Reality using Multiple Smartphones," *International Conference on Artificial Reality and Telexistence (ICAT2011),* ISSN: 1345-1278, pp.143, 2011

[15] T. Hashimoto, Y. Uematsu, H. Saito, "Generation of See-Through Baseball Movie from Multi-Camera Views," *International Workshop on Multimedia Signal Processing (MMSP2010)* , pp.432-437, 2010.

[16] R.Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J.Shotton, S. Hodges, A. Fitzgibbon, "KinectFusion: Real-Time Dense Surface Mapping and Tracking," *International Symposium on Mixed and Augmented Reality (ISMAR 2011)*,pp. 127-136, 2011.

[17] R. Gioi, J. Jakubowicz, J. Morel, G. Randall, "LSD: A Fast Line Segment Detector with a False Detection Control," *IEEE Tran. on PAMI*, 32, pp.722-732, 2010.