International Journal of Advanced Smart Convergence Vol.5 No4 1-9 (2016)
https://doi.org/10.7236/IJASC.2016.5.4.1

IJASC 16-4-1

Fast key-frame extraction for 3D reconstruction from a handheld video

Jongho Choi[1], Soonchul Kwon[2], Kwangchul Son[2], Jisang Yoo[1*]

[1]*Department of Electronic Engineering, Kwangwoon University, Seoul, Korea*
[2]*Graduate School of Information and Contents, Kwangwoon University, Seoul, Korea*
e-mail : {mrchoi, ksc0226, kcson, *jsyoo}@kw.ac.kr

Abstract

In order to reconstruct a 3D model in video sequences, to select key frames that are easy to estimate a geometric model is essential. This paper proposes a method to easily extract informative frames from a handheld video. The method combines selection criteria based on appropriate-baseline determination between frames, frame jumping for fast searching in the video, geometric robust information criterion (GRIC) scores for the frame-to-frame homography and fundamental matrix, and blurry-frame removal. Through experiments with videos taken in indoor space, the proposed method shows creating a more robust 3D point cloud than existing methods, even in the presence of motion blur and degenerate motions.

Key words: *key-frame, handheld, 3D reconstruction, motion blur, geometric estimation*

1. Introduction

To reconstruct real objects or scenes in three dimensions is one of the most challenging areas of computer vision. Multi-view stereo (MVS) or structure from motion (SfM) methods are typical and progress in various fields [1-4]. With images acquired at various locations, the objects are reconstructed in three dimensions by estimating camera pose. MVS and SfM, also called image-based modelling, are a more progressive approach than LiDAR or structured-light reconstruction, since the objects can be reconstructed without any prior-knowledge or constraints on the scenes. However, this modeling generally utilizes a lot of images acquired at the fixed viewpoint. Each image contains only the partial shape of the objects, so that hidden surfaces occur naturally in the image. Therefore, obtaining multi-view images abundantly is important to the purpose of reconstructing all the surfaces of the objects.

The video can be used to easily obtain the multi-view images. It provides more useful information when compared with still images taken from various locations. Moreover, scanning the scenes with video mode is a lot faster than photographing repeatedly. However, preprocess for extracting key frames in the video must be accompanied. Importantly, a set of the extracted key frames has to satisfy conditions that the frames reflect the

Manuscript Received: Sep. 10, 2016 / Revised: Sep. 25, 2016 / Accepted: Oct. 7, 2016
Corresponding Author: jsyoo @kw.ac.kr
Tel: +82-2-940-5112, Fax: +82-2-912-7305
Graduate School of Information and Contents, Kwangwoon University

whole motion of the video, and the number of the extracted frames should be as small as possible considering the efficiency of 3D reconstruction.

A number of researchers have reported studies on key-frame extraction from the video for 3D reconstruction. Gibson *et al*. [5] proposed a method of selecting key frames according to the ratio of corresponding points between frames and re-projection errors of the homography and the fundamental matrix. Seo *et al*. [7] was similar to [5], but the correspondence distribution of the frames was considered to determine the accuracy of the fundamental-matrix estimation. As the corresponding points are uniformly distributed in the frame, the reliability of the fundamental-matrix estimation increases. Pollefeys *et al*. [8] utilized geometric robust information criterion (GRIC) [9] to determine degenerate cases that are difficult to estimate the fundamental matrix. In the degeneracies, to represent the geometric model between frames with the homography is more suitable than with the fundamental matrix. Ahmed *et al*. [10] defined a function of the key-frame selection as the weighted sum of the GRIC and point to epipolar-line cost (PELC), which is more robust to the noisy frames and further reduces the re-projection error.

In this paper, we propose a method for extracting key frames from a handheld video. When extracting key frames from the handheld video, two practical issues are considered further. One is the blur effect. Motion blur is inevitable in handheld shooting that is vulnerable to hand shake, unlike a professional video that is assisted by motion stabilizers. Blurry frames are unreliable in the process of matching feature points between the frames, and cannot be used as a texture. For that reasons, it is necessary to evaluate the quality of the frames and remove the blurry frames. The other is the size of the video. Usually one minute of video consists of at least 1,000 frames. Eventually, the time required to extract the key frame from the video is proportional to the video size. However, since video has a high similarity between adjacent frames, it is possible to fast search through the video by skipping some frames.

Figure 1 shows the overall structure of key-frame extraction from the video for 3D reconstruction. The composition of the paper is as follows. Chapter 2 explains how to skip the frames by predicting the ratio of corresponding points between frames, and estimating the fundamental matrix, and robustly filtering out the degeneracies. Chapter 3 introduces a technique for refining key-frame set by evaluating the quality of key frames. Through analyzing the relative degree of blur between frames, the blurry frames are removed to obtain final-frame set. Chapter 4 analyzes the results of experiments conducted on handheld videos taken in indoor environment. The proposed method is compared with the existing methods in terms of the number of key frames, the processing time, the ratio of the low-quality frames in the key-frame set, and the number of successful reconstruction. Finally, Chapter 5 gives a brief summary and discuss future research.

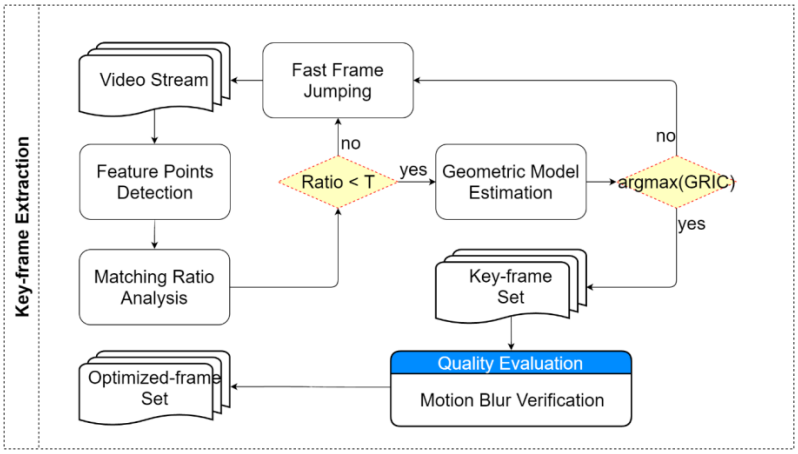

Figure 1. Flowchart of the proposed method.

2. Extraction of key frame to facilitate fundamental matrix estimation

In general, in the video, camera motion has similar characteristics between adjacent frames. If the camera movement is analyzed as an optical flow, the corresponding points between the frames can be detected effectively. Using triangulation formula, depth from the matching-point pairs can be calculated simply. However, in order to reduce the uncertainty of the depth calculation, a baseline distance between cameras should be sufficient. Key frames must also satisfy this condition. We use Equation (1) defined by Seo *et al.* [7] as a measure to determine the baseline distance.

$$R_m = \frac{N_t}{N_f} \tag{1}$$

N_f is the number of feature points of the reference frame, and N_t is the number of feature points tracked in the following frame. The baseline distance and the ratio R_m are inversely proportional. The frame with low R_m must be selected as key frames to obtain the sufficient baseline distance. However, as the ratio decreases, the number of corresponding points required for estimating the fundamental matrix decreases, which affects estimation of camera pose. Therefore, only frames that the ratio R_m satisfies a upper limit threshold T_{upper} and a lower threshold T_{lower} are regarded as candidates for the fundamental matrix estimation.

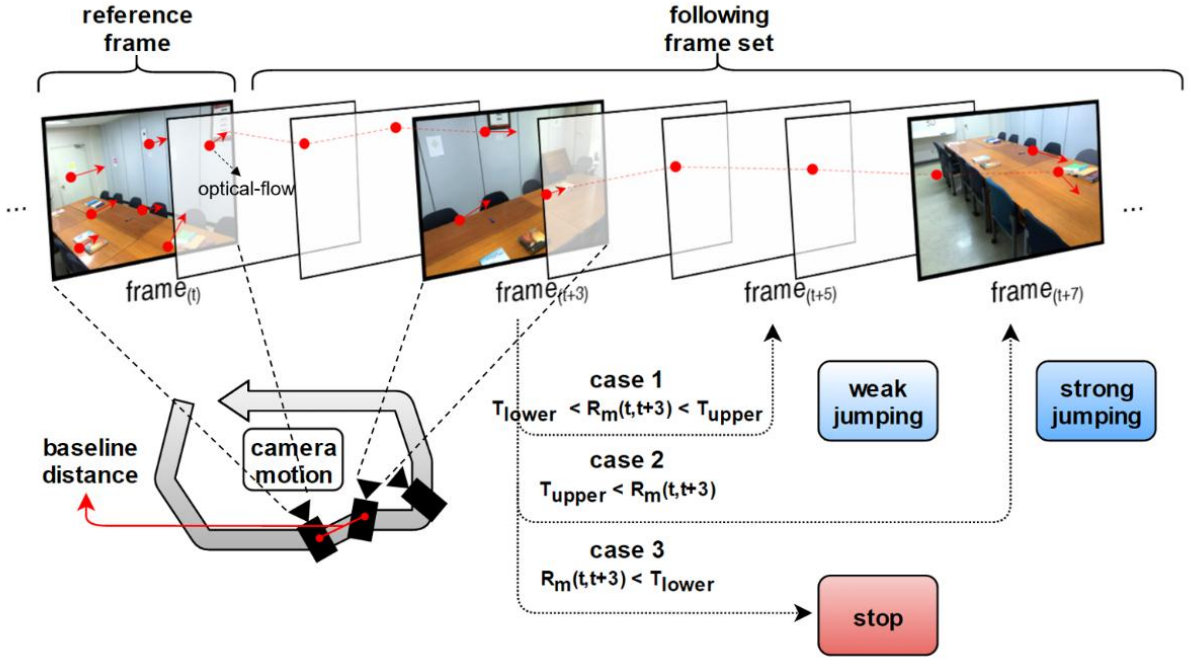

Figure 2. A fast frame jumping in video through analysis of matching points ratio. $R_m(t,t+3)$ represents the ratio R_m between $frame_{(t)}$ and $frame_{(t+3)}$.

At the same time, it is necessary to quickly find frames that satisfy the double threshold in the video. Figure 2 shows a fast frame jumping method. The reference frame represents the most recently extracted key frame. The ratio R_m of the reference frame to the following frame converges to zero when overlapping

scene does not exist, and the ratio is usually decreased as the following frames move away from the reference frame. A prediction for the ratio allows for calculations in some frames, without calculating the ratio in all frames. Since the similarity between adjacent frames in the video is considerably high, loss of camera motion information is not great even if the calculation is done only in specific frames. Through checking $R_m(t,t+3)$, the way to move from $frame_{(t+3)}$ is divided into three cases. In case 1, the double threshold value is satisfied, so $frame_{(t+3)}$ is judged as having a sufficient baseline distance. Thus, $frame_{(t+3)}$ is considered as a candidate for the fundamental matrix estimation, and $R_m(t,t+5)$ is calculated after the weak jump to $frame_{(t+5)}$. In case 2, $R_m(t,t+7)$ is calculated after a strong jump to $frame_{(t+7)}$ because $R_m(t,t+3)$ is too high, i.e., the baseline distance is not sufficient. Case 3 assumes that the corresponding points for the fundamental matrix estimation are insufficient and stops further jumps. Then, among the candidate frames satisfying case 1, the frame that is the easiest to estimate the fundamental matrix with the reference frame is adopted as the next reference frame.

When estimating the fundamental matrix, incorrect matching pairs act as an unstable factor, but there are two situations, called degeneracy, that the estimation even with correct matching pairs is numerically unstable. One is the structure degeneracy where all the corresponding points are located on the coplanar in the three dimensions and the other is the motion degeneracy with only the rotational motion with respect to the focal point without any translation. In this situations, it is more appropriate to express the set of corresponding points by a homography, which is a projection transformation model, rather than by the fundamental matrix. Therefore, it is needed to decide which of two geometric models is more suitable between the reference frame and candidate frames, and candidate frames which are more fit for the homography should be excluded.

We use the geometric robust information criterion (GRIC) [9] as a model selection criterion for filtering the degeneracies. Given corresponding points, the GRIC converts the verification result of each model to a score and determines that the model with the smaller GRIC score is more appropriate for the given data. For that reason, the next reference frame should have the GRIC score of the fundamental matrix less than that of the homography, and is the candidate with the largest difference in score between the two models. Equation (2) represents a function of the key-frame selection.

$$\mathrm{K_{i+1}} = \mathrm{argmax}_{\mathrm{j}\in\forall(\mathrm{K_i})}\left(\frac{\left|\mathrm{GRIC_F(i,j)} - \mathrm{GRIC_H(i,j)}\right|}{\mathrm{GRIC_H(i,j)}}\right) \tag{2}$$

The meaning of the symbols is as follows. i and j are the frame index of the video, K_{i+1} is the next key frame, $\forall(K_i)$ is all the candidate frames for the key frame K_i, and $GRIC_{model}(i,j)$ is the model's GRIC score for the frame i and j. Since the GRIC score is usually proportional to the number of corresponding points used in the estimation step, the GRIC difference between the two models is normalized by dividing by $GRIC_{\mathrm{H}}(i,j)$ to compensate for the number of matching pairs. As a result, the frame with the largest normalization value is selected as the next key frame. Through the fast frame jumping and function of the key-frame selection, a key-frame set is first extracted from the video.

3. Quality evaluation of key frames

The handheld shooting is sensitive to motion speed and hand tremors. The more the irregular motions occur frequently when scanning the object, the more the motion blur becomes noticeable in the video, and then the scene change between the frames are not smooth. Such motion blur can cause a small number of feature points detection within the frame and reduce the reliability of feature points matching, leading to a poor reconstruction. Thereby, the motion blur frame must be detected in the key-frame set and removed properly. The blur measurement method [11-14], which can quantify the degree of motion blur, is important, and most measurements focus on the brightness changes. As the frame is sharper, edge elements in the frame generally increase. Thus, the degree of blur can be defined as the inverse of the edge component in the frame. Equation (3) represents the blur metric of the frame.

$$B_t = 1 \Big/ \sum_{t(x,y)} \{ (\frac{\partial t}{\partial x})^2 + (\frac{\partial t}{\partial y})^2 \} \tag{3}$$

The edge can be expressed by the sum of squares of the partial derivatives in the x and y directions at all the pixel positions (x, y) of t frame, and the inverse of the edge is defined as a blurriness B_t. Figure 3 shows that the edges in some frames are detected. 3-(b) has a larger value B_t because edge components are less detected than that of 3-(d).

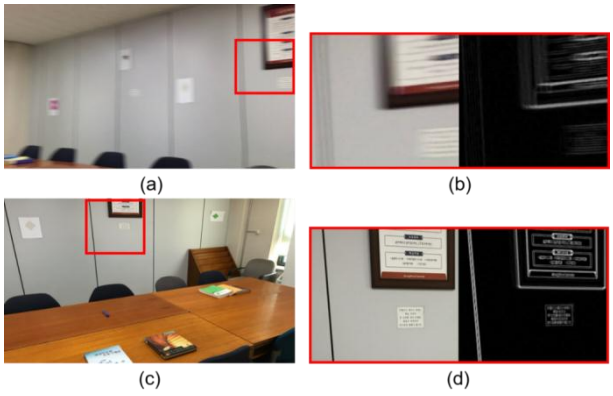

Figure 3. Decrease of edge due to blur. (a) blur frame, (b) edges of a portion of (a), (c) non-blur frame, (d) edges of a portion of (c).

Since the amount of edge or noise detected in the frame is relative, it is not advisable to simply determine the degree of blur by B_t. Although the blurriness of frames cannot be absolutely evaluated, relative blurriness can be measured within a limited neighbor frame if the frame-to-frame scene change is not severe. Matsushita *et al.* [12] defined the relative blurriness RB_t of t frame as B_t / B_{t-1} and judged that t frame is sharper than t-1 frame when its value is less than one. However, the dichotomous judgment considering only the magnitude of edge components has a possibility to incorrectly detect the motion blur. In addition, it is difficult to observe the changes of RB_t in the entire frame because the reference blurriness is applied differently every frame. Therefore, this paper proposes a method to detect the motion blur by applying additional constraints to the relative blurriness verification. The additional constraints are covered in Chapter 4.

4. Experiments

We experimented on the performance of the proposed technique in 7 handheld videos that captured the same indoor space with the Sony HDR camera (30 fps). Each video has a different camera movement speed. Three videos have a slow moving speed and the other four have a fast moving speed. The average number of frames of video is 3,380. A manual calibration was performed using a chess board to obtain the camera internal parameters.

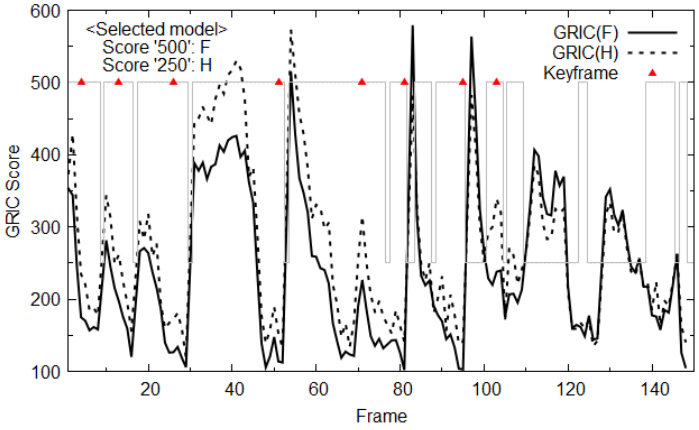

Figure 4. GRIC scores for some sequences.

The solid line and dotted line represent the GRIC for the fundamental matrix and the GRIC for the homography, respectively. The triangle means the selected key frame. To prominently represent the selected model by GRIC, the score is set to 500 if the fundamental matrix is selected and set to 250 otherwise.

In the interior space, flat structures such as wall are common. The degeneracy mostly corresponds to a frame containing these regions. The feature points detected on the wall have almost similar depths and make model estimation difficult. Figure 4 shows the result of selecting key frames through Equation (2) in some sequences. Among the 150 frames, 31 frames mapped with a score of 250 are judged as the degeneracy, which the homography model is more suitable for. The degenerate frames are filtered out, and when a frame having the largest GRIC score difference between the two models is selected, 8 key frames marked with triangle are obtained.

As described in Chapter 3, the additional constraints on relative blurriness are empirical rules derived from experiments. First, RB_t is newly defined as shown in Equation (4) below.

$$\mathrm{RB_t} = \frac{\mathrm{B_{reference}}}{\mathrm{B_t}} \quad \mathrm{t} \in \forall(\mathrm{K}) \tag{4}$$

$$\Delta\mathrm{RB_t} = \mathrm{RB_t} - \mathrm{RB_{t\text{-}1}} \tag{5}$$

New RB_t is a ratio between the blur of reference frame $B_{reference}$ and the blur of t frame B_t. To analyze the total change of RB_t in the key frame set $\forall(K)$, RB_t is calculated by the same $B_{reference}$ to all the frames. Figure 5 is an example showing some RB_t for $\forall(K)$ as a line graph. In the method [12], if B_t / B_{t-1} is less than one, it judges that there is no motion blur at t frame, but not all leads to correct detections.

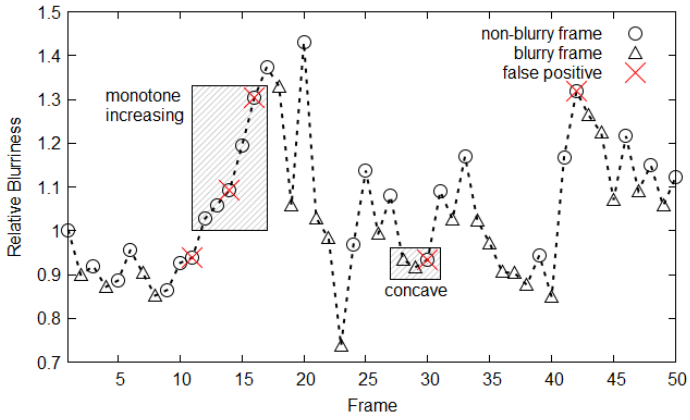

Figure 5. Relative blurriness graph.

The frames marked with circle are judged to be sharp, while the frames marked with triangle are judged to be blurry. False positive indicates a frame that is not actually blurry but is mistakenly determined to be blurry.

There are some false positive frames as shown in the Figure 5. False negative frames should be detected and removed because they have an adverse effect on the reconstruction process. Since the baseline distances are different for each key frame, the rate of change in Equation (5) is not constant. However, near the false positive frames, ΔRB_t is usually constant. This aspect can be seen in the section where RB_t curve is concave (28-30 frame) or monotone increasing (12-16 frame). In these sections, the scene change between the key frames becomes slight because the details of edges are reduced due to the motion blur.

$$0.90\times(\Delta \mathrm{RB}_{t-1}) \le \Delta \mathrm{RB}_t \le 1.10\times(\Delta \mathrm{RB}_{t-1}) \tag{6}$$

$$\mathrm{RB}_t \le 1.05\times(\mathrm{RB}_{t-1}) \tag{7}$$

Equation (6) examines the degree of scene change. If t frame satisfies the condition, it is regarded as false positive and removed from the key frame set. However, if t frame satisfies the Equation (6) and both ΔRB_t and ΔRB_{t-1} are 0.20 or more, it is not considered false positive such as the 24 frame. In addition, in monotone increasing sections, a frame having a very small change in RB_t is regarded as false positive. This corresponds to the 10-11 frame section and can be determined through Equation (7).

Table 1. Result of experiments.

Method	Uniform sampling	Seo et al. [4]	Ahmed et al. [7]	Proposed method
Number of video clips	7	7	7	7
Number of failure in 3D reconstruction	4	1	1	0
Average number of frames per video clip	3,380	3,380	3,380	3,380

Average number of extracted key-frames per video clip	*112*	98	81	52
Average time of extracting key-frames per video clip (min)	*1.8*	6.2	4.5	5.5
Average low-quality frame rate (%)	*29.1*	17.1	12.0	5.5

Through VisualSFM toolbox [15-16], 3D reconstruction was performed on the key frames extracted from the video. Table 1 shows the performance of the proposed method compared with the existing methods. Uniform sampling extracts the key frames in the shortest time, but fails to generate a 3D point cloud in more than half the videos. The methods [7, 10] taken into account geometric relationships of camera movements show better performances at the number of successful reconstruction. However, the ratio of low-quality frames in the key frame set is 17.1% [7] and 12.0% [10]. They do not fully consider the properties of handheld video, where motion blur is frequent, and show that reconstruction fails on a certain video. On the other hand, the proposed method reduces the low quality frame rate to 5.5% on average by removing the low quality frame based on the blur metric. As a result, 3D point cloud is successfully generated in all videos.

One possible limitation of the proposed method is the need to specify some thresholds. We currently set these parameters empirically. However, since all parameters are related to the number of corresponding points acquired or the processing time, in principle, it should be possible to find values that work well for most sequences and adjust them when necessary.

5. Conclusion

This research focuses on finding suitable frames for 3D reconstruction and discarding unnecessary frames through key frame extraction. When reconstructing the scene from the video, the process of key-frame extraction improves reconstruction performance and minimizes computation time. The result of experiments shows that, especially in the handheld video that is vulnerable to motion blur, properly removing low-quality frames is effective for 3D reconstruction. Future work will focus on enhancing the system for outdoor environments and arbitrary videos. The key frame extraction may need to interact with other process such as moving object segmentation, shot boundary detection, and auto-calibration to achieve this goal.

Acknowledgement

The present research has been conducted by the KLB Foundation Fund in 2016

References

[1] L. Ling, Ian S. Burrent, E. Cheng, "A dense 3D reconstruction approach from uncalibrated video sequences" ICMEW, pp. 587-592, 2012.

[2] J. Frahm, M. Pollefeys, S. Lazebnik, D. Gallup, B. Clipp, R. Raguram, C. Wu, C. Zach, T. Johnson, "Fast robust large-scale mapping from video and internet photo collections" ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 65, No. 6, pp. 538-549, 2010.

[3] N. Snavely, S.M. Seitz, R. Szeliski, “Modeling the world from Internet photo collections”, International Journal of Computer Vision, Vol. 80, No. 2, pp. 189-210, 2008.

[4] M. Pollefeys et al., “Detailed real-time urban 3D reconstruction from video”, International Journal of Computer Vision, Vol. 78, pp. 143-167, 2008.

[5] S. Gibson, J. Cook, T. Howard, R. Hubbold, D. Oram. "Accurate camera calibration for off-line, video-based augmented reality", in: IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR), Darmstadt, Germany, 2002.

[6] J.K. Seo, S.H. Kim, C.W. Jho, H.K. Hong, “3D Estimation and Key-Frame Selection for Match Move”, ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications, pp. 1282-1285, July 2003.

[7] Y.H. Seo, S.H. Kim, K.S. Doo, J.S. Choi, "Optimal keyframe selection algorithm for three-dimensional reconstruction in uncalibrated multiple images", Journal of the Society of Photo-Optical Instrumentation Engineers, Vol. 47, No. 5, pp. 53201- 53400, 2008

[8] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, "Visual modeling with a hand-held camera", International Journal of Computer Vision, Vol. 59, No. 3, pp. 207-232.

[9] P.H.S. Torr, A.W. Fitzgibbon, A. Zisserman, "Maintaining multiple motion model hypotheses over many views to recover matching and structure", in: Proc. The 6th International Conference on Computer Vision, Bombay, India, pp. 485-491, 1998.

[10] M.T. Ahmed, M.N. Dailey, J.L. Landabaso, N. Herrero, "Robust key frame extraction for 3D reconstruction from video streams", in: Proc. The VISAPP, pp. 231-236, 2010.

[11] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, W.T. Freeman, “Removing camera shake from a single photograph”, ACM Transactions on Graphics, Vol. 25, No. 3, pp. 787-794, 2006.

[12] Y. Matsushita, E. Ofek, X. Tang, H.Y. Shum, "Full-frame video stabilization", in: Proc. Computer Vision and Pattern Recognition, pp. 50-57, 2005.

[13] S. Cho, J. Wang, S. Lee, "Video deblurring for hand-held cameras using patch-based synthesis", ACM Transactions on Graphics, Vol. 31, No. 4, pp. 1-9, 2012.

[14] S. Yang, M. Lizhuang, “Detecting and Removing the Motion Blurring from Video Clips”, I.J.Modern Education and Computer Science, Vol. 1, pp. 17-23, January 2010.

[15] Changchang Wu, “Towards Linear-time Incremental Structure from Motion”, in: Proc. International Conference on 3D Vision, pp. 127-134, 2013.

[16] Changchang Wu, "VisualSFM: A Visual Structure from Motion System", http://ccwu.me/vsfm/, 2011