Exploring Local Feature Descriptors for Construction Site Video Stabilization

Jung Yeol Kim[a] and Carlos H. Caldas[b]

[a] PhD Candidate, Dept. of Civil, Architectural and Environmental Engineering, Univ. of Texas at Austin, USA
[b]Associate Professor, Dept. of Civil, Architectural and Environmental Engineering, Univ. of Texas at Austin, USA
E-mail: jungyeol.kim@utexas.edu, caldas@mail.utexas.edu

Abstract -

Recent studies on automated activity analysis have adopted construction videos as an input data source to recognize and categorize construction workers' actions. To ensure the representativeness of its analysis results, these videos have to be gathered randomly in terms of time and location. In doing so, such videos must be taken with hand-held cameras, a fact that inevitably leads to videos including jittery frames. Such frames can decrease the accuracy of automated activity analysis results. One area of the most recent and effective action recognition methods involves using spatio-temporal action recognition algorithms. The jittery frames, however, are fatal to the recognizing of a human worker's action using such an algorithm. Jitters can be removed from the videos by using video stabilization technologies. The video stabilization is the pre-processing of action recognition for automated activity analysis. Regarding the video stabilization, local feature descriptor plays a major role in the stabilization process, and the correct selection of proper descriptor is critical. Therefore, the purpose of this study is to identify the best local feature descriptor for the video stabilization. This paper describes detail steps of the stabilization and provides performance analysis of various local feature descriptors in terms of stabilization of videos from construction site.

Keywords -

IT applications, video stabilization, video interpretation, activity analysis, work sampling, productivity measurement, action recognition

1 Introduction

Videotaping has long history in gathering on-site data for construction productivity analysis [1][2][3][4][5]. Recent years, due to the advancement of computer vision technologies, many researchers have studied on automated interpretation methodologies of construction site videos[6][7][8][9][10][11][12][13][14].

One of the construction industry's computer vision application areas is activity analysis. Activity analysis is a continuous measurement and improvement process that helps craft workers increase their time spent on actual construction work. It includes the application of work sampling in its measurement process and requires manual observations of workers [39]. The main focus of the computer vision application in the activity analysis is to substitute, in construction videos, the manual observation of construction workers' actions with automatic recognition and categorization [9] [12] [18].

To ensure the representativeness of its analysis results at the site, the construction videos have to be gathered randomly in terms of time and location [39]. The well-planned combination of hand-held and fixed closed-circuit television (CCTV) cameras can be a solution to obtaining those random videos. CCTV camera is a convenient tool to obtain videos at random intervals but has a limitation regarding random locations. It cannot cover all the areas of the construction site. A hand-held camera is useful in gathering those videos at random time intervals and places, but such videos inevitably include jittery frames. Jitters in those videos can decrease the accuracy of automated activity analysis results. One area of the most recent and effective action-recognition methods uses spatio-temporal action-recognition algorithms [22]. The jittery frames, however, are fatal to the recognizing of a human worker's action when such an algorithm is being used; the jitters in the videos can distort the spatio-temporal volumes, trajectories, or features. Those jitters can be removed from the videos by using video stabilization technologies. The video stabilization is the pre-processing of action recognition for automated activity analysis. Regarding the video stabilization, local feature descriptor is one of the most important elements. Therefore, the purpose of this study is to identify the best local feature descriptor for the video stabilization. This paper describes detail steps of the stabilization and provides performance analysis of various local feature descriptors in terms of stabilization of videos from construction site.

2 Related work

2.1 Automated Action Recognitions for Activity Analysis

In the automation of activity analysis, researchers have studied three types of action recognition technologies: (1) sensor-based [15][16]; (2) 2D image/video-based [6][9][11][12][13]; and 3D vision data- (i.e., depth image, point clouds, and human skeleton) based ([17][19][18]) action recognition. All these approaches have contributed to the automation of activity analysis. However, the 2D and 3D vision-based approaches assume that their inputs are static. Most of the studies use vision data from fixed 2D or 3D imaging sensors. In actual situations, some of the data have to be gathered by hand-held imaging sensors, where jitters are unavoidable. Therefore, it is necessary to adopt the stabilization technologies of the data. This paper focuses on the stabilization of 2D video data.

2.2 Video stabilization

Video stabilization falls into two types: hardware-based stabilization during recording and software-based, post-processing digital video stabilization [35][21]. Hardware-based stabilizers consist of complex and expensive sensors and lens systems to reduce the movement of cameras. Cheaper cameras also adopt sensors and firmware to offset camera motions. However, these hardware-based systems fail to provide sufficient stabilization function to compensate for complex camera motions and severe jerking. Therefore, to obtain stable videos, post-processing video stabilization is still required [35]. The Post-processing digital video stabilization is defined as “the process of removing the unwanted motion from input video sequence by appropriately warping the images” [37]. It is not a real-time solution but can be applied to the videos taken by any type of cheap hand-held cameras. This paper focuses on the software-based post-processing digital video stabilization.

Software-based video stabilization (hereafter “video stabilization”) can be divided into two types: (1) 2D and (2) 3D video stabilization [20]. A general 2D video stabilization method is composed of the three steps as shown in Figure 1: 1) motion estimation, 2) motion compensation, and 3) image composition [36][23][35]. Motion estimation means the estimation of motion between two sequential frames (i.e., motion between the previous and current frames). Motion compensation provides the computation of global transformation to stabilize the current frame. Based on the transformation, image composition warps the current image. Recently, more innovative approaches have been introduced such as very stable and anti-distortional.

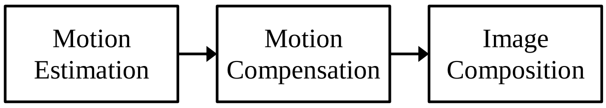


Figure 1. General video stabilization method [36]

3D video stabilization does not mean the stabilization of 3D vision data. It is for 2D videos, though it uses the estimation of 3D model of input camera motion and scene. It also use image-based rendering techniques to render new frames based on the estimated camera motion path. The new rendered frames are frames of video stabilized [24][25][26]. One interesting study that used this method is a content-preserving warping carried out by Liu et al. [20]. It distinguishes itself from other methods by not having a blank area on the stabilized video.

The 2D video stabilization is limited regarding significant depth variations but is still a simple, robust, and efficient solution [35][20]. The 3D video stabilization could overcome the depth variation problem but is more complex and often depends on unreliable depth estimation [35]. Furthermore, the authors assumed that a cameraman does not walk when taking videos, and those videos consequently have less depth variations. Therefore, this paper focuses on the 2D video stabilization methods instead of 3D based methods.

3 Our Approach

3.1 Overview

Our approach, shown in Figure 2, is a variation of the 2D-based general video stabilization method. It consists of five steps with details given in the following paragraphs.

The ***first step*** is to extract local feature descriptors from the first and second frames. Again, these descriptors are one of the most important elements of the video stabilization method. They are used for the estimation of geometrical transform for stabilization. The geometrical transform is estimated by the matched descriptors of sequential frames

In this paper, the authors selected the following four descriptors: (1) Scale-Invariant Feature Transform (SIFT) [28]; (2) Speeded Up Robust Features (SURF) [29]; (3) Fast Retina Keypoint (FREAK) [30]; and (4) Oriented FAST and Rotated BRIEF (ORB) [31]. The authors selected these features because SIFT is well known for its scale and rotation invariant performance [32]; SURF is inspired by SIFT but is known for its higher detection speed and better performance. FREAK is a newer descriptor and shows the faster detection speed and better robustness than SIFT and SURF according to its inventor's experiments [30]. ORB is a combination of FAST (Features from Accelerated Segment Test) corner detector [42][43] and BRIEF (Binary Robust Independent Elementary Features) descriptors. Rublee et al [31], the inventor of ORB, insisted that ORB outperforms SURF and SIFT. The experiment's results regarding the stabilization performance by these feature descriptors will be described in the next section of this paper.

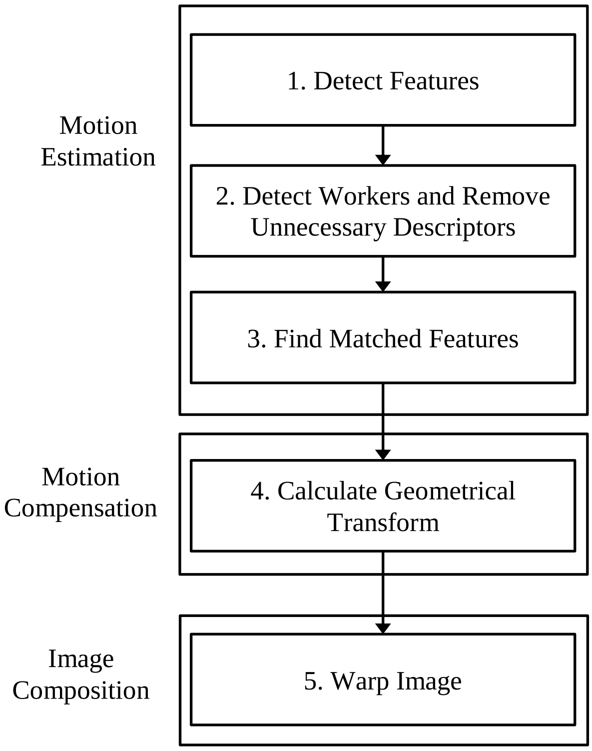


Figure 2. Our construction video stabilization approach

The ***second step*** is to recognize construction workers in each frame using a human-detection algorithm, histograms of oriented gradients (HOG) [33] and to remove unnecessary local descriptors detected within the workers' regions in each frame. Figure 3 shows the example descriptors detected in the worker's regions. Those descriptors can be sources of error during the estimation of geometrical transformation; indeed, the directions of workers' movements (trajectories) can differ from the camera's jittering directions. Figure 4 shows the matched local feature descriptors outside of the worker's regions in the sequential frames. In this case, the estimation of the geometrical transform in the next step will be incorrect [27].

Our approach differs from of Wang and Schmid [27] by eliminating SURF descriptors before identifying matched descriptors. They simply selected SURF descriptor and motion vectors to estimate homography between two consecutive frames and eliminate matched descriptors in the people's region. Our approach is simpler but effective because there are still sufficient amount of descriptors outside of the worker's regions that enable the estimating of the geographical transform. Importance to this step is the accurate recognition of human workers.

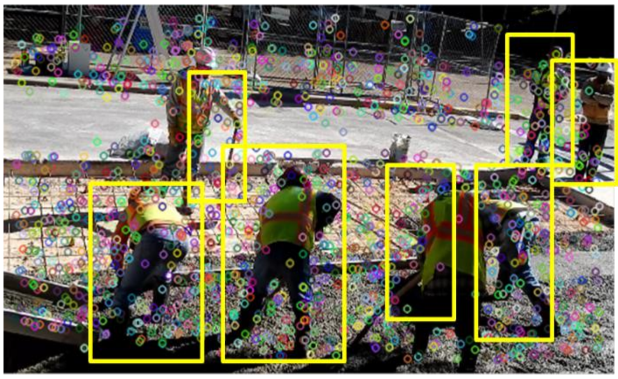


Figure 3. Local feature descriptors detected in the worker's regions (SURF Descriptor Used)

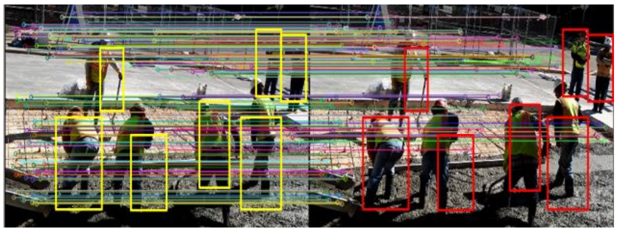


Figure 4. Matched Local Feature Descriptors outside of the Worker's Regions in the Sequent Frames (SURF Descriptor Used, Top 150 Matches Displayed out of around 1,500 Matches)

The authors follow the general process (Figure 1) for the remaining steps [23][34][35]. The ***third step*** is to compare the remaining descriptors of the two frames to discover corresponding points. To match the corresponding descriptors, this study used the Nearest Neighbor Ratio for floating point descriptors and Hamming distances for binary descriptors. The ***fourth step*** is to estimate the geometrical transform with the corresponding points. Affine transform was used in this study and was compensated. The ***last step*** is to warp the second frame with the geometrical transformation and repeat these steps.

4 Experiments and Results

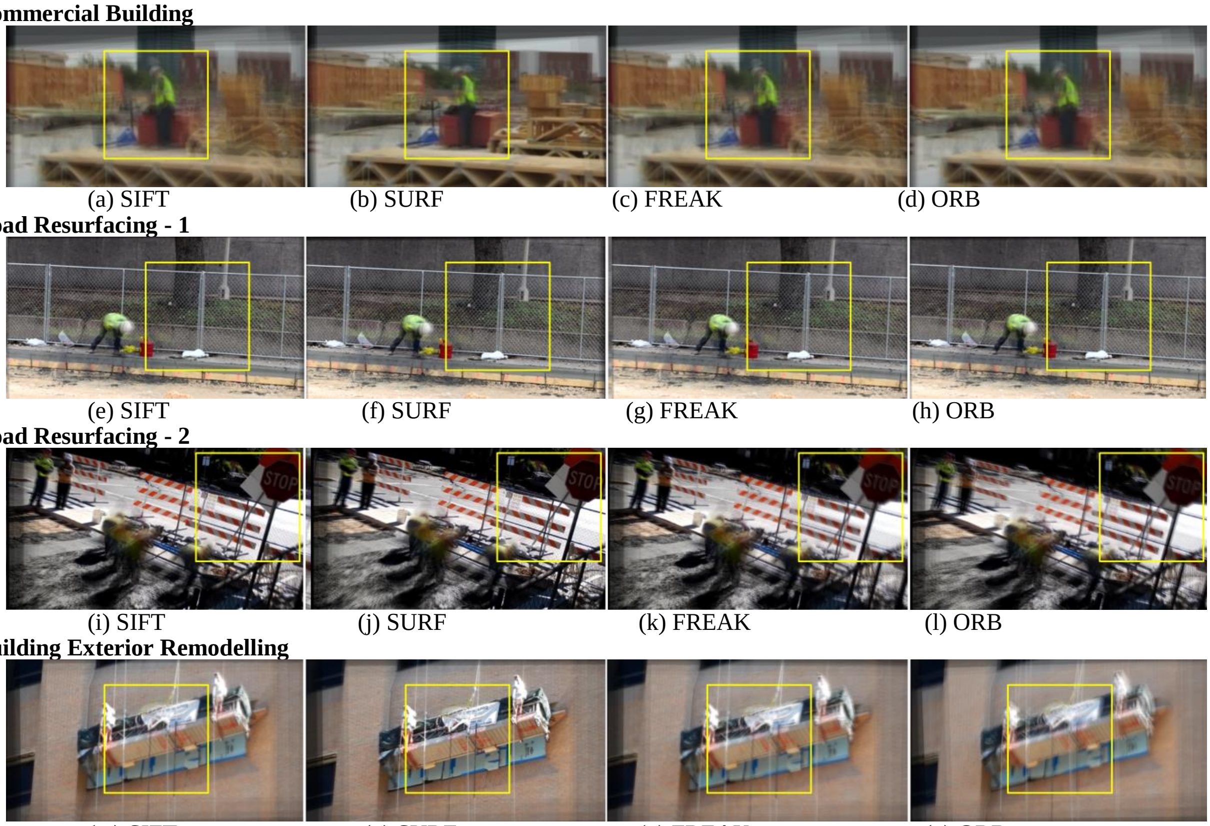
The authors compared the stabilization performance with four feature descriptors. The performance will vary according to the descriptors because they each have a different ability to discover corresponding points with jittery frames. The jittery frames include horizontal movement, vertical movement, rotation, and the combination of all the jitters. The authors used OpenCV, VL-FEAT, C++, and Matlab for the experiments. The videos were gathered from a commercial building, road resurfacing, and building exterior remodelling sites with cheap hand-held camera. The resolution and frame per second (fps) of the video used for this experiment were 573 × 320 and 5 fps. The average computation time per frame is 0.58 seconds.Figure 5 shows the experiment's results. From the first to the fourth rows correspond to the videos from the commercial building, road resurfacing (the second and the third rows), and building exterior remodelling sites.

Each column of the figure corresponds to the video stabilization results with SIFT, SURF, FREAK, and ORB descriptors. Each image is an overlaid image from the first stabilized frame to the last. Therefore, sharpness can be a proper metric to compare their performances. The sharper the image the better the stabilization performance; it means that objects in each stabilized image are at similar locations. The yellow boxes in each image are the sharpest parts. Figure 5 shows that the overlaid images stabilized with SURF descriptor have the highest sharpness. To the naked eye, however, it is hard to distinguish the relative sharpness of the images. Therefore, the authors measured the sharpness of each overlaid image. There are many methods to estimate the sharpness of images. The authors adopted the Brenner gradient based sharpness measurement method because it is more sensitive to the sharpness changes than other methods [44]. The measurement results are shown in Figure 6. The higher number means a higher sharpness.

Figure 6 shows that SURF always outperforms other descriptors. SIFT follows SURF, and FREAK and ORB demonstrate worse performances. Sometimes, the FREAK descriptor fails to find matched descriptors between two sequent frame images. In terms of computation time, stabilization with FREAK generally showed the shortest computation time, while SIFT showed the longest.

Based on the experiments, it can be concluded that SURF and SIFT are robust to stabilize the jittery videos due to their rotation-invariant characteristics. Therefore, the authors selected SURF for the best descriptor for our construction video stabilization.

Commercial Building



(a) SIFT (b) SURF (c) FREAK (d) ORB

Road Resurfacing - 1

(e) SIFT (f) SURF (g) FREAK (h) ORB

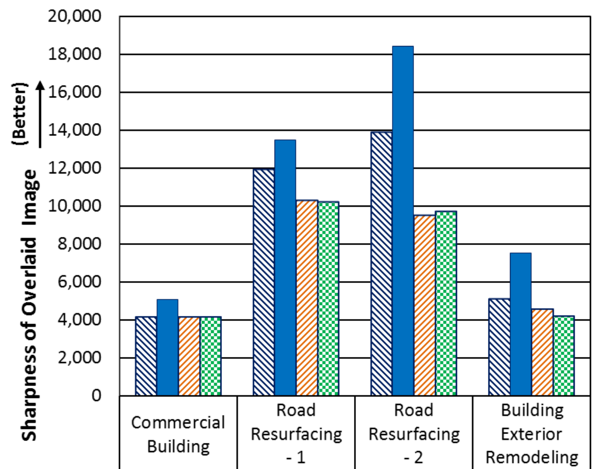
Road Resurfacing - 2

(i) SIFT (j) SURF (k) FREAK (l) ORB

Building Exterior Remodelling

(m) SIFT (n) SURF (o) FREAK (p) ORB

Figure 5. Construction video stabilization results



	Commercial Building	Road Resurfacing - 1	Road Resurfacing - 2	Building Exterior Remodeling
SIFT	4,185.15	11,945.37	13,871.23	5,098.50
SURF	5,077.63	13,474.51	18,426.24	7,540.51
FAST	4,197.66	10,304.22	9,526.38	4,575.10
ORB	4,199.52	10,214.92	9,725.92	4,208.97

Figure 6. Stabilization performance comparisons by different descriptors.

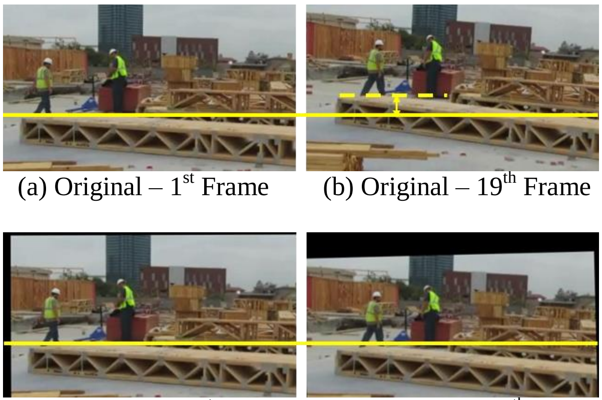
Figure 7 shows an example of original jittery frames and stabilized frames. Figure 7 (a) and (b) are the 1st and 19th frames of the original video. The long solid line is the basis point of the first image and the dotted line is the vertical difference between the two frames. Otherwise, there is a really small amount of vertical difference in the corresponding stabilized frames (Figure 7 (c) and (d)).

5 Conclusions

The authors have explained a modified video stabilization method for the activity analysis while considering construction workers in videos. The authors also performed an experiment to find out the best descriptor with the video stabilization method. The authors used the sharpness of overlaid stabilized images as a metric to measure stabilization performance. In the experiment, SURF descriptor performed best, followed by SIFT descriptor.

This video stabilization method could pave the way for the use, at a construction jobsite, of cheap hand-held cameras and any other mobile video-recording devices, such as Gopro®, Looxcie, and Google Glass. This would mean that it could become easy, and with less expense, to gather random videos for automated activity analysis.

This study has few limitations. Human detection, a part of the second step to eliminating unnecessary descriptors, needs to be improved in the future. The HOG based human-detection algorithm used in the step has some level of Type-I (False Positive) and Type-II (False Negative) errors. Furthermore, it tends to better detect upright position than other poses. Their effects were not considered because it is not the scope of this study, but the authors believe that using the human detector still could reduce the chance of errors as it stand. Another limitation of this method is that the experiment is performed with only four descriptors and a small number of videos. Experiments involving more descriptors and a greater number of videos need to be performed in the future.



(a) Original – 1st Frame (b) Original – 19th Frame

(c) Stabilized – 1st Frame (d) Stabilized – 19th Frame

Figure 7. Final Results - Comparisons with original and stabilization video frames

References

[1] Fondahl, J. W. Photographic analysis for construction operations. *Journal of the Construction Division*, 86(2), 1960.

[2] Oglesby, C. H. *Productivity improvement in construction*. McGraw-Hill series in construction engineering and project management, McGraw-Hill, New York, 1989.

[3] Everett, J., Halkali, H., and Schlaff, T. Time-Lapse Video Applications for Construction Project Management. *Journal of Construction Engineering and Management*, 124(3), 204–209, 1998.

[4] Abeid, J., and Arditi, D. Time-Lapse Digital Photography Applied to Project Management. *Journal of Construction Engineering and Management*, 128(6), 530–535, 2002.

[5] Gong, J., and Caldas, C. (2010). "Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations." *Journal of Computing in Civil Engineering*, 24(3),

252–263.

[6] Peddi, A., Huan, L., Bai, Y., and Kim, S. Development of Human Pose Analyzing Algorithms for the Determination of Construction Productivity in Real-Time. American Society of Civil Engineers, 11–20, 2009

[7] Gong, J., and Caldas, C. Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations. *Journal of Computing in Civil Engineering*, 24(3), 252–263, 2010.

[8] Gong, J., and Caldas, C. H. An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, 20(8), 1211–1226, 2011.

[9] Gong, J., Caldas, C. H., and Gordon, C. Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced Engineering Informatics*, 25(4), 771–782, 2011.

[10] Chi, S., and Caldas, C. H. Image-Based Safety Assessment: Automated Spatial Safety Risk Identification of Earthmoving and Surface Mining Activities. *Journal of Construction Engineering and Management*, 138(3), 341–351, 2012.

[11] Heydarian, A., Golparvar-Fard, M., and Niebles, J. C. Automated Visual Recognition of Construction Equipment Actions Using Spatio-Temporal Features and Multiple Binary Support Vector Machines. American Society of Civil Engineers, 889–898, 2012.

[12] Rezazadeh Azar, E., and McCabe, B. Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in Construction*, 24, 194–202, 2012.

[13] Memarzadeh, M., Golparvar-Fard, M., and Niebles, J. C. Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32, 24–37, 2013.

[14] Ranaweera, K., Ruwanpura, J., and Fernando, S. Automated Real-Time Monitoring System to Measure Shift Production of Tunnel Construction Projects. *Journal of Computing in Civil Engineering*, 27(1), 68–77, 2013.

[15] Varghese, K., and Joshua, L. Classification of Bricklaying Activities in Work Sampling Categories Using Accelerometers. *Construction Research Congress 2012*, American Society of Civil Engineers, 919–928, 2012.

[16] Cheng, T., Teizer, J., Migliaccio, G. C., and Gatti, U. C. Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data. *Automation in Construction*, 29(0), 24–39, 2013.

[17] Escorcia, V., Dávila, M. A., Golparvar-Fard, M., and Niebles, J. C. Automated Vision-Based Recognition of Construction Worker Actions for Building Interior Construction Operations Using RGBD Cameras. *Construction Research Congress 2012*, American Society of Civil Engineers, 879–888, 2012.

[18] Kim, J. and Caldas, C.H. Vision-Based Action Recognition in the Internal Construction Site Using Interactions between Worker Actions and Construction Objects. Proceedings of the 2013 International Symposium on Automation and Robotics in Construction, Montreal, Canada, August 10-14, 2013.

[19] Weerasinghe, I., Ruwanpura, J., Boyd, J., and Habib, A. Application of Microsoft Kinect Sensor for Tracking Construction Workers. *Construction Research Congress 2012*, American Society of Civil Engineers, 858–867, 2012.

[20] Liu, F., Gleicher, M., Wang, J., Jin, H., and Agarwala, A. Subspace Video Stabilization. *ACM Trans. Graph.*, 30(1), 4:1–4:10, 2011.

[21] Chereau, R., and Breckon, T. P. Robust motion filtering as an enabler to video stabilization for a tele-operated mobile robot. *in Proc. SPIE Security and Defence: Electro-Optical Remote Sensing, SPIE*, 2013.

[22] Aggarwal, J. K., and Ryoo, M. S. Human Activity Analysis: A Review. ACM Comput. Surv., 43(3), 16:1–16:43, 2011.

[23] Matsushita, Y., Ofek, E., Ge, W., Tang, X., and Shum, H.-Y. Full-frame video stabilization with motion inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7), 1150–1163, 2006.

[24] Buehler, C., Bosse, M., and McMillan, L. Non-metric image-based rendering for video stabilization. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, II–609–II–614 vol.2, 2001.

[25] Fitzgibbon, A., Wexler, Y., and Zisserman, A. Image-based rendering using image-based priors.

Ninth IEEE International Conference on Computer Vision, 2003. Proceedings, 1176–1183 vol.2, 2003.

[26] Bhat, P., Zitnick, C. L., Snavely, N., Agarwala, A., Agrawala, M., Cohen, M., Curless, B., and Kang, S. B. Using Photographs to Enhance Videos of a Static Scene. *Eurographics Symposium on Rendering (2007)*, Eurographics Association 2007, Grenoble, France, 2007.

[27] Wang, H., and Schmid, C. Action Recognition with Improved Trajectories. *ICCV 2013*, IEEE, Sydney, Australia, 2013.

[28] Lowe, David G. Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision* 60, no. 2: 91–110. doi:10.1023/B:VISI.0000029664.99615.94, 2004.

[29] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3), 346–359, 2008.

[30] Alahi, A., Ortiz, R., and Vandergheynst, P. FREAK: Fast Retina Keypoint. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 510–517, 2012.

[31] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 IEEE International Conference on Computer Vision (ICCV), 2011: pp. 2564–2571, 2011.

[32] Mikolajczyk, K., and Schmid, C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1615–1630, 2005.

[33] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, 886–893 vol. 1, 2005.

[34] Hartley, Richard, and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge, UK; New York: Cambridge University Press, 2003.

[35] Lee, K. Y., Chuang, Y. Y., Chen, B. Y., and Ouhyoung, M. Video stabilization using robust feature trajectories. *2009 IEEE 12th International Conference on Computer Vision*, 1397–1404, 2009.

[36] Morimoto, C., and Chellappa, R. Fast Electronic Digital Image Stabilization. *in: Proceedings of ICPR*, 284–288, 1995.

[37] Morimoto, C., and Chellappa, R. Evaluation of image stabilization algorithms. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, 2789–2792 vol.5, 1998.

[38] K.G. Derpanis, M. Sizintsev, K.J. Cannons, R.P. Wildes, Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence. 35, 527–540, 2013.

[39] Gouett, M., Haas, C., Goodrum, P., and Caldas, C. Activity Analysis for Direct-Work Rate Improvement in Construction. Journal of Construction Engineering and Management, 137(12), 1117–1124, 2011.

[40] Park, D., Zitnick, C. L., Ramanan, D., and Dollar, P. Exploring Weak Stabilization for Motion Feature Extraction. 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2882–2889, 2013.

[41] Jenkins, J. L., and Orth, D. L. Productivity Improvement Through Work Sampling. Cost Engineering, 46(3), 27–32, 2004.

[42] Rosten, E., and Drummond, T. Fusing points and lines for high performance tracking. IEEE, 1508–1515 Vol. 2, 2005.

[43] Rosten, E., and Drummond, T. Machine Learning for High-Speed Corner Detection. Computer Vision – ECCV 2006, Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, eds., Springer Berlin Heidelberg, 430–443, 2006.

[44] Treeby, B. E., Varslot, T. K., Zhang, E. Z., Laufer, J. G., and Beard, P. C. Automatic sound speed selection in photoacoustic image reconstruction using an autofocus approach. *Journal of Biomedical Optics*, 16(9), 090501–090501–3, 2011.