Browsing Group First-Person Videos with 3D Visualization

Yuki Sugita[1], **Keita Higuchi**[1], **Ryo Yonetani**[1], **Rie Kamikubo**[1], **and Yoichi Sato**[1]

[1]The University of Tokyo

ABSTRACT

This work presents a novel user interface applying 3D visualization to understand complex group activities from multiple first-person videos. The proposed interface is designed to assist video viewers to easily understand the collaborative relationships of group activity based on where the individual worker is located in a workspace and how multiple workers are positioned to one another during the group activity. More specifically, the interface not only shows all recorded first-person videos but also visualizes the 3D position and orientation of each view point (*i.e.*, the 3D position of each worker wearing a head-mounted camera) with a reconstructed 3D model of the workspace. Our user study confirms that the 3D visualization helps video viewers to understand geometric information of a worker and collaborative relationships of group activity easily and accurately.

Author Keywords

First-person videos; Group activity

INTRODUCTION

In this work, we design a user interface for understanding group activities from multiple first-person points-of-view videos recorded by wearable cameras. We consider a scenario where multiple workers in the shared workspace collaborate to accomplish complex group tasks such as in relocation of house, rescue, or construction. To improve the overall efficiency or quality of group performance, it is important to review and assess the task. To this end, we exploit the case where all the workers are equipped with a wearable camera to capture their first-person videos during the task.

The use of wearable cameras is advantageous for recording what workers looked at and how they used their hands. These cues would be helpful for understanding certain group activities but are difficult to observe by using a fixed camera in the workspace. Recent work on HCI has tried to make use of wearable cameras for lifelogging [3, 9, 8, 12] and remote assistance [6, 7, 11].

However, the recorded videos from the wearable cameras cannot always allow observers to grasp geometric relationships among the workers. Each first-person video only captures limited parts of the workspace, and the visible parts can drastically change when the workers move. As a result, video observers face difficulties to track the positional information in the shared workspace, including not only absolute locations of each worker but also positional relationship between multiple workers. It is also difficult to see group performance regarding who is working with whom.

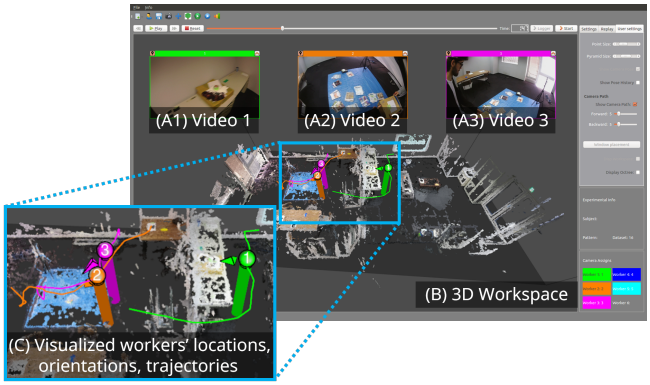

Figure 1. The proposed interface shows (A1-3) tiled multiple first-person videos, (B) the 3D visualization of a reconstructed workspace, and (C) 3D poses and trajectories of workers. Workers are equipped with wearable cameras to record first-person points-of-view videos.

In order to address the aforementioned limitations, we develop a novel interface for browsing multiple first-person videos of group work effectively using computer vision techniques. As shown in Figure 1, in addition to (A1-3) first-person videos of multiple workers displayed in a tiled fashion, our interface is featured by the **workspace-view** widget consisting of (B) reconstructed 3D geometry of the workspace and (C) 3D poses and trajectories of wearable cameras. In our user study, we asked participants to browse first-person videos of several group tasks, and confirmed the effectiveness of the proposed interface to assist the understanding of geometric and collaborative relationships of workers.

RELATED WORK

Many user interfaces have been developed for browsing multiple videos recorded by various types of cameras. For multiple fixed surveillance cameras, prior works have often included a map of the workspace obtained from a fisheye lens [4], a prebuilt 3D model [16], reconstructed 3D models from multiple video streams [19], and crowd-sourcing [17]. Other studies [2, 13] have explored smoothing techniques of video transitions for both moving and fixed cameras. In comparison, we focus especially on browsing multiple first-person videos captured by wearable cameras. Advantages of our setting are to capture

ISS '18, November 25–28, 2018, Tokyo, Japan

© 2018 ACM. ISBN 978-1-4503-5694-7/18/11…15.00

DOI: https://doi.org/10.1145/3279778.3279783

one's field of view up close and to reduce the cost of installing fixed cameras to each new workspace.

Recent work has also explored visualizing multiple first-person videos. In [10], multiple first-person videos are simply tiled on head-mounted displays to share work in progress among camera wearers for remote collaboration. In [15], 3D poses of multiple wearable cameras are visualized to discover joint attention of camera wearers. Kono *et al.*have developed a view-sharing system that allows camera wearers to share position information among the wearers [14]. To extend the visualization techniques for multiple first-person videos [10, 15, 14], we propose to visualize 3D geometry of a workspace, in addition to poses and trajectories of wearable cameras to assist in understanding of group activities.

INTERFACE DESIGN

This section introduces the design of an interface for assisting an observer to review video-recorded collaborative activities using multiple wearable cameras. Specifically, we consider a scenario where multiple workers share their workspace and collaborate with each other to accomplish a complex task such as relocating house, rescue, and/or construction.

One straightforward approach is to tile and present multiple videos at a time as shown in [10, 16]. In this work, we refer to this design as a baseline interface. The baseline interface, however, becomes problematic in understanding geometric and collaborative relationships of workers from multiple first-person videos. Since such videos are recorded with multiple moving wearable cameras, observers can easily lose their awareness regarding the position of each worker and the geometric and collaborative relationships of multiple workers. Figure 2 shows some examples. (1) is a first-person video of a worker walking forward. Because this video captures only a part of an entire workspace, it is hard to grasp the moving path of the worker. In (2), even though we are able to see three workers being present in the workspace, it is hard to capture their formation as they frequently look down or away. Moreover, the video in (3) shows that five workers are co-drawing on paper while forming some groups around a table. Since they mostly tilted their heads forward, it is difficult to see their group formation accurately.

To address the limitations in capturing spatio-temporal information from multiple first-person videos, we propose to visualize the group activity with geometric information of wearable cameras and the workspace. Our proposed interface specifically offers a *workspace-view* which visualizes the positions, poses, and pathways of wearable cameras in the 3D workspace environment, as shown in Figure 1. The visualization of the workspace view is reconstructed by using computer vision techniques for RGB images of cameras. The proposed interface combines the workspace-view with tiled first-person videos.

The proposed interface is designed to assist observers in understanding the following four components:

- **Workspace geometry from each worker's first-person point-of-view**
  With the 3D visualization of the entire workspace (Figure 1 (B)), we expect that observers can easily and accurately understand the positional relationship of workers and objects captured in each first-person video by seeing the 3D models of the workspace.

Figure 2. Difficulties of understanding geometric and collaborative relationships of workers from multiple first-person videos: (1)location changes of a single worker, (2)formations and (3)groups of multiple workers.

- **Linkage of each camera position to each worker's position**
  By visualizing each camera position with a fine color and an ID label (Figure 1 (C)), we expect that observers can easily link each camera position to the position of each worker. We refer this visualization from [21]. This can be a direct clue of geometric relationships of multiple workers. The height of each camera is also visualized to assist observers to grasp activities of each worker (*e.g.*, standing or sitting).

- **Viewing direction of each first-person video and the head orientation of each worker**
  We assume that the visualization of each camera orientation helps observers to grasp the viewing direction of each first-person video. This indicates head orientation of each worker that can be a beneficial clue to see the workstation (*e.g.*, desk and workbench) of a worker in the 3D environment. Head orientations of workers also can be useful to understand a collaborative group of workers.

- **Each worker's moving trajectory**
  We also expect that the visualization of each camera trajectory assists observers to understand the moving pathway of each worker. This visualization can also give direct clues to know past and future positions of workers.

IMPLEMENTATION

The system implementation can be divided into two stages: reconstructing and visualizing workspace structures and camera movements. For the reconstruction process, the system uses video streams from wearable cameras and pictures of a physical workspace.

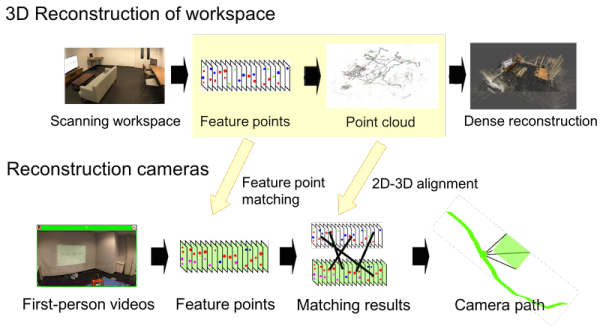

Figure 3. Pipeline of reconstructing workspace and camera paths

Reconstructing geometry of workspace and workers

Figure 3 shows the pipeline of reconstructing a workspace and camera paths from a collection of wearable cameras. In this study, we adopted a computer vision-based 3D reconstruction framework: VisualSFM [23] that offered a pipeline of functions using a set of scene images as an input to compute its 3D point cloud as an output (i.e., structure from motion [22], bundle adjustment [24], and patch-based multi-view stereo [5]). We also used Point Cloud Library [18] to filter out noisy points and reduce data size of the point cloud. In order to estimate the accurate 3D model, images of a workspace were taken separately before recording group activities. Images with heavy motion blur were detected by computing optical flows and omitted from the reconstruction process.

Once the workspace structure is reconstructed, the system runs a next reconstruction process to obtain geometric information of each camera from a video stream. To obtain the geometric information, the system corresponds feature points of each video image to feature points of the workspace using the SIFT feature. In order to reduce a computational cost, this matching was made for every twenty frames. We also partially matched image frames among wearable cameras to obtain more feature correspondences among images. Specifically, given a certain image frame, we evaluated matches against every four frames of the consecutive forty-one frames around that frame for all the cameras including its own camera.

Implementing the proposed interface

As depicted in Figure 1, the 3D structure of the workspace is presented on the 2D plane seen from the high oblique point of view (B). We refer to this projected view of workspace structure as the workspace-view. We set the initial viewing angle of the workspace structure as 45 degrees. We allow observers to resize and rotate the structure around the vertical axis.We then visualize camera positions by a colored circle with an ID (C). We also visualize camera heights with a cylinder. Camera orientations and trajectories are represented by an arrow and a line, respectively. We set the initial viewing length of forward and backward trajectories as 5 seconds.

In addition, also as shown in Figure 1 (A and B), camera locations and corresponding first-person videos are associated with a border color of the video window. For instance, the video with a green border is captured by the green camera in the workspace view. We also implemented some basic playback-control tools such as a seek-bar, a play-and-pause button, fast-forward and fast-backward buttons and a reset button.

USER STUDY

We conducted a series of user study to investigate whether and how the proposed interface for browsing group activities of multiple workers assists in understanding the geometric and collaborative relationships. We implemented a simple video browser as a baseline interface that tiles and shows multiple first-person videos. We assumed that the proposed interface with the workspace view significantly supports understanding geometric information in group activities from first-person videos compared to the baseline interface.

To focus on the investigation of how our proposed workspace-view visualization assists in understanding geometric information, we tried to reduce the confounding factors of the observers' habits and interests to use the interface. In our preliminary study, we observed that interface manipulation behaviors drastically varied depending on the observers and changed the usability outcomes. Therefore, we took screencasts of the interfaces and presented them to the observers, allowing them to play the casts only once. In the study, we asked participants in the role of observers about several questions for understanding group activities. We designed three tasks to see the effectiveness of the proposed workspace-view that we assumed to give important information of group activities.

Dataset

Prior to the user study, our dataset was collected by recording eight sequences of group activity performed by three or five workers equipped with head-mounted wearable cameras. Specifically, the group activities involve: (A) box shipping and transportation, (B) walking around posters, (C) collecting, organizing, and delivering magazines and boxes, (D) drinking around a table, (E) co-drawing on one paper, (F) box transportation, (G) cleaning up connected rooms and (H) block composition. Each dataset was recorded at more than one of the five different locations such as foyers or connected office rooms. The duration of each dataset was 32–632 seconds. In this study, we took a screencast of the proposed interface that played each dataset, and clipped the screencast into multiple experimental videos.

Tasks

In this experiment, we assigned three tasks to investigate whether and how the proposed view would assist observers to understand the movement of each worker in the workspace and group formation of workers easily and accurately.

TASK 1: worker's movement

We asked observers about a given worker's movement during the time when the screencast is in play. As shown in the left of Figure 4 (b), observers draw a line by hand on a canvas, on which a layout of the workspace is also displayed. The duration of each screencast is about twelve seconds.

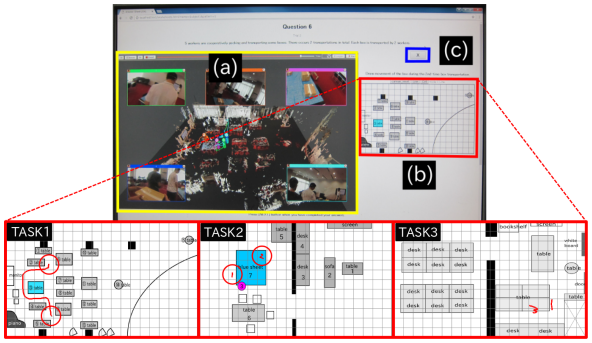

Figure 4. Experimental user interface: (a) video for experiment, (b) canvas for drawing answers with answer examples of three experimental tasks and (c) play-button are placed on 24" monitor

TASK 2: three workers' formation

We asked observers about a formation of given three workers at the end of the screencast. One of the three workers is already displayed on the canvas. Observers answer positions and IDs of the other two workers. Screencasts have different variations for whether the formation change is included during the play. The duration of each screencast is about thirty seconds.

TASK 3: a group of workers

We asked observers to detect two groups from five workers. For instance, when there are five workers forming two groups which are made by two and three workers respectively, observers draw IDs of all the workers who are in the same group as a given worker on the canvas.

Varying screencasts are provided as to whether the group has been formed from the beginning. The duration of each screencast is about thirty seconds.

Experimental Setting

We recruited 12 participants as observers for the user study. To counterbalance, we assigned different order of presenting two visualization methods and two screencasts to each observer; that is, 3 observers were assigned to each combination. Our experimental settings are illustrated in Figure 4. In all the tasks, the observers were asked to draw their answers on a canvas with a floor map (Figure 4(b)). Each observer browsed screencasts with both the baseline and the proposed visualizations alternately for each task. We gave the observers tutorials prior to performing each task. In the tutorial session, we asked observers to perform a practice task to get used to the following main task. Note that datasets used for the tutorial were different from the datasets for the main tasks.

Evaluations

Objective Evaluation

We manually scored accuracy of tasks (0.0 – 1.0) according to the following criteria: starting point, end point and middle path of the movement (Task 1); positions of two workers and the ID order of three workers (Task 2); and group member IDs and their location (Task 3). Note that positional accuracy is scored based on the grid lines overlaid on the canvas. For example, in TASK1, we judged that an end point is correct when the point is inside the correct grid or its adjacent grids.

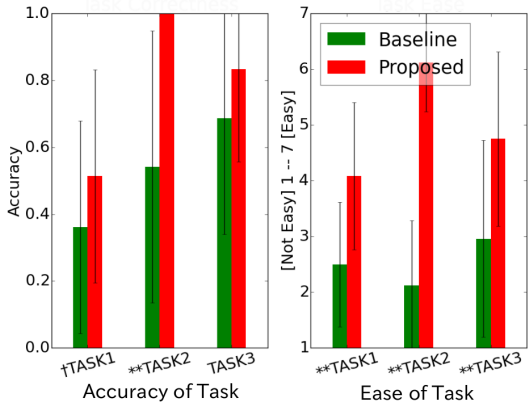

Figure 5. Accuracy of Task and Ease of Task. (†, *, **) under each graph represent each significance level ($p < .1, .05, .01$).

Subjective Evaluation

In each task, we asked observers to score the ease of task according to the seven scales of 1–7 (1: difficult, 7: easy). At the end of the experiment, we also asked observers to score the ease of understanding of the proposed visualizations and the 3D-model of the workspace according to the seven grades respectively. After the experiment, we took interview sessions to see their experience during tasks. We focused on benefits and difficulties in use of the proposed interface.

RESULT

Three Tasks

Figure 5 shows average accuracy (i.e., task scores normalized with a range of [0, 1]) and average ease levels of all tasks. To evaluate limited numbers of Likert score data, we used a non-parametric method of Wilcoxon signed-rank test. As a result, in TASK 1, we observed significant increase in both accuracy of task ($p < .1$) and ease of task ($p < .01$) in the proposed compared to the baseline. In TASK 2, we observed significant increase in both accuracy of task ($p < .01$) and ease of task ($p < .01$). We also observed significant increase in ease of task ($p < .01$) for TASK 3.

End Questions

Observers answered two questions at the end of the experiment: (EQ1) how easy was it to understand the workspace-view visualization, and (EQ2) how easy was it to understand the workspace geometry by seeing the 3D model from the presented viewpoint. The average 7 graded scores are 6.2 ± 0.7 and 6.4 ± 0.8 respectively.

Summary of Interviews

We received positive comments about the proposed visualization of camera position, orientation and pathways respectively. Camera positions helped identifying the position of each worker and their geometric relationship/collaboration. Camera orientations also assisted in grasping workers' interests, moving direction, and working space. Visualized camera pathways gave cues to estimate future positions of workers.

In contrast, we also obtained negative comments about the proposed interface. Observers mainly reported estimation errors of reconstructing camera positions and orientations that adversely affected their experience of using the proposed visualization. Few observers said that the workspace-view was not effective when there was no movement of workers.

DISCUSSION

We conclude that the proposed workspace-view effectively assists observers to understand geometric relationships of workers easily and accurately. As shown in Figure 5, we confirmed that the accuracy and ease of two tasks about grasping the movement of each worker (TASK1) and the formation of multiple workers (TASK2) were improved by the proposed interface. In the proposed interface, the observers could use the workspace-view to become aware of positions, directions and pathways of workers directly.

We observed that workspace-view visualization assists observers to easily grasp groups formed by multiple workers. We also confirmed that the subjective ease of task was significantly improved in TASK3. Though solving the task itself was not so difficult enough to yield the difference between the baseline and the proposed, it seems that making use of the proposed visualization instead of comparing multiple videos significantly improved the subjective ease of task. We also observed that the accuracy of task was improved. In the proposed condition, observers referred the workspace-view to see camera positions. It is assumed that observers could easily find the group candidates by seeing the closely positioned visualizations of cameras. Camera pathways were also assumed to be the clue for detecting group candidates of workers, who were apparently moving in collaboration. Camera orientations served as clues for identifying the co-working place and determining whether or not the workers of interest were collaborating.

As a limitation of the work, observers reported estimation errors of camera positions. They said that adversely affected the task performance. In our implementation, we used a structure-from-motion (SfM) technique to reconstruct the cameras and workspace facilities. However, SfM often fails to reconstruct 3d geometry from video frames when there are few point correspondences between the frames. To overcome this limitation, we plan to improve a pipeline of 3D reconstruction by using the state-of-the-art 3D reconstruction methods [1, 20].

CONCLUSION

This work introduced a novel user interface for browsing group first-person videos with 3D visualization of workspace and camera positions. The interface is designed to assist understanding group activity performed by multiple workers in a physical workspace. We developed the prototype interface for evaluating effects of the proposed 3D visualization. Through the user study, we confirmed that the proposed visualizations assisted observers to understand geometric and collaborative relationships of workers.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JPMJCR14E1, Japan.

REFERENCES

1. Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. ACM Transactions on Graphics 33, 4 (2014), 81.
2. Luca Ballan, Gabriel J Brostow, Jens Puwein, and Marc Pollefeys. 2010. Unstructured video-based rendering: Interactive exploration of casually captured videos. ACM Transactions on Graphics 29, 4 (2010), 8.
3. Yi Chen and Gareth JF Jones. 2010. Augmenting human memory using personal lifelogs. In Proceedings of the 1st Augmented Human International Conference (AH). 24.
4. Philip DeCamp, George Shaw, Rony Kubat, and Deb Roy. 2010. An immersive system for browsing and visualizing surveillance video. In Proceedings of the 18th ACM international conference on Multimedia (ACMMM). 371–380.
5. Yasutaka Furukawa and Jean Ponce. 2010. Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 8 (2010), 1362–1376.
6. Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. 2003. Effects of Head-mounted and Scene-oriented Video Systems on Remote Collaboration on Physical Tasks. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI). 513–520.
7. Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016a. Can Eye Help You?: Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI). 5180–5190.
8. Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016b. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. (2016).
9. Yoshio Ishiguro, Adiyan Mujibiya, Takashi Miyaki, and Jun Rekimoto. 2010. Aided Eyes: Eye Activity Sensing for Daily Life. In Proceedings of the 1st Augmented Human International Conference (AH). 25.
10. Shunichi Kasahara, Mitsuhito Ando, Kiyoshi Suganuma, and Jun Rekimoto. 2016. Parallel Eyes: Exploring Human Capability and Behaviors with Paralleled First Person View Sharing. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI). 1561–1572.
11. Shunichi Kasahara and Jun Rekimoto. 2014. Jackin: Integrating first-person view with out-of-body vision generation for human-human augmentation. In Proceedings of the 5th Augmented Human International Conference (AH). 46.
12. Seita Kayukawa, Keita Higuchi, Ryo Yonetani, Masanori Nakamura, Yoichi Sato, and Shigeo Morishima. 2018. Dynamic Object Scanning: Object-Based Elastic

Timeline for Quickly Browsing First-Person Videos. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18). Article LBW611, 6 pages.

13. Ryohei Komiyama, Takashi Miyaki, and Jun Rekimoto. 2017. JackIn Space: Designing a Seamless Transition Between First and Third Person View for Effective Telepresence Collaborations. In Proceedings of the 8th Augmented Human International Conference (AH). Article 14, 9 pages.
14. Michinari Kono, Takashi Miyaki, and Jun Rekimoto. 2017. JackIn Airsoft: Localization and View Sharing for Strategic Sports. In Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST '17). Article 3, 4 pages.
15. Hyun S Park, Eakta Jain, and Yaser Sheikh. 2012. 3d social saliency from head-mounted cameras. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS). 422–430.
16. Eleanor G Rieffel, Andreas Girgensohn, Don Kimber, Trista Chen, and Qiong Liu. 2007. Geometric tools for multicamera surveillance systems. In First ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC). 132–139.
17. Peter M Roth, Volker Settgast, Peter Widhalm, Marcel Lancelle, Josef Birchbauer, Norbert Brandle, Sven Havemann, and Horst Bischof. 2011. Next-generation 3D visualization for visual surveillance. In Proceedings of the 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS). 343–348.
18. Radu Bogdan Rusu and Steve Cousins. 2011. 3d is here: Point cloud library (pcl). In Robotics and automation (ICRA), 2011 IEEE International Conference on. IEEE, 1–4.
19. James Tompkin, Kwang In Kim, Jan Kautz, and Christian Theobalt. 2012. Videoscapes: Exploring Sparse, Unstructured Video Collections. In ACM Transactions on Graphics, Vol. 31. DOI: http://dx.doi.org/10.1145/2185520.2185564
20. B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. 2017. DeMoN: Depth and Motion Network for Learning Monocular Stereo. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
21. Y. Wang, D. Bowman, D. Krum, E. Coalho, T. Smith-Jackson, D. Bailey, S. Peck, S. Anand, T. Kennedy, and Y. Abdrazakov. 2008. Effects of Video Placement and Spatial Context Presentation on Path Reconstruction Tasks with Contextualized Videos. IEEE Transactions on Visualization and Computer Graphics 14, 6 (Nov 2008), 1755–1762. DOI: http://dx.doi.org/10.1109/TVCG.2008.126
22. Changchang Wu. 2013. Towards linear-time incremental structure from motion. In Proceedings of the IEEE International Conference on 3D Vision (3DV). 127–134.
23. Changchang Wu. Accessed: 2016-05-20. VisualSFM: A visual structure from motion system. http://ccwu.me/vsfm/. (Accessed: 2016-05-20).
24. C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. 2011. Multicore bundle adjustment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3057–3064.