An Event Reconstruction Tool for Conflict Monitoring Using Social Media

Junwei Liang, Desai Fan, Han Lu, Poyao Huang, Jia Chen, Lu Jiang, Alexander Hauptmann
Carnegie Mellon University
{junweil, dfan, hlu2, poyaoh, jiac, lujiang, alex}@cs.cmu.edu

Abstract

What happened during the Boston Marathon in 2013? Nowadays, at any major event, lots of people take videos and share them on social media. To fully understand exactly what happened in these major events, researchers and analysts often have to examine thousands of these videos manually. To reduce this manual effort, we present an investigative system that automatically synchronizes these videos to a global timeline and localizes them on a map. In addition to alignment in time and space, our system combines various functions for analysis, including gunshot detection, crowd size estimation, 3D reconstruction and person tracking. To our best knowledge, this is the first time a unified framework has been built for comprehensive event reconstruction for social media videos.

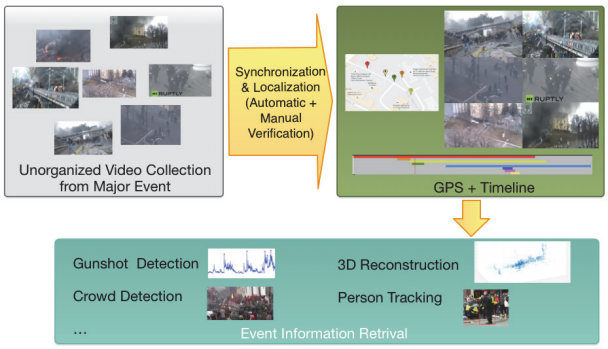

Figure 1: The workflow of Our Event Reconstruction Tool

Introduction

A tremendous amount of videos is being uploaded to social network sites every minute, documenting every aspect of our lives and events all over the world. When an event happens, especially for those involving a large crowd of people, multiple videos would be taken, recording the same event at different moments, from different angles, and at different positions. For example, events like New Year's Eve at NYC, Carnival in Brazil, and Boston Marathon bombing all have hundreds of or even thousands of attendees uploading videos of the event. The collection of these user-generated recordings not only enables new applications such as free-view video (Collet et al. 2015) and 3D-reconstruction (Zhang et al. 2015), but it may also help our society achieve a more unbiased understanding of the event truth (Aronson, Xu, and Hauptmann 2015). Such information would be particularly important for conflict or violence events, in which the truth is critical to the general public and for law enforcement to take action.

Unlike videos captured by fixed, calibrated surveillance cameras, consumer videos are captured "in the wild" (i.e., at various time, location, perspectives, with different devices such as smart phones, watches or camcorders.) These videos are sometimes noisy and have low quality. For example, in an unexpected violent event, people are often scared and the videos taken under that situation may thus be too blurry or shaky to see things clearly. Useful information about the event may spread across different time segments of different videos. To enable accurate common video analysis tasks such as person tracking and scene reconstruction, we first solve the two basic and essential tasks, synchronization and localization of videos.

In contrast to previous work, we consider the video in-the-wild paradigm. Our system synchronizes and localizes videos from a large unorganized collection. Based on the time and location meta data, our system combines advanced analysis tools of gunshot detection, crowd size estimation, 3D scene reconstruction and person tracking.

System Framework

Our system takes a video collection from a major event as input and then puts all videos into a global timeline by synchronization and onto a map by localization, as shown in Figure 1. Give the synchronization and location result, users can utilize our powerful tool for various kinds of event information retrieval, including gunshot detection, crowd size estimation, 3D reconstruction and person tracking.

Video Synchronization and Localization

Our video synchronization system operates in two stages. The first stage is pairwise video synchronization. Our system finds the best synchronization for each video to the others. Then in the second stage, our system find the best global synchronization among all the videos based on the pairwise

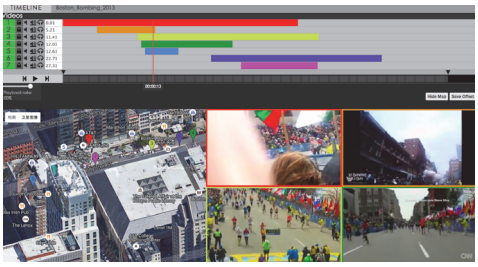

Figure 2: The Interface of our Synchronization and Localization Results

alignment. In an unexpected violent event, videos are often noisy and blurry. Therefore in this system we focus on the audio modality for synchronization. During the first stage, our system first conducts an unsupervised clustering to get an audio signature dictionary, and then assigns each time frame of the video segments to the closest k centers. Then based on the audio signature assignment, we get the pairwise synchronization results between videos. In the second stage, the global alignment is achieved by minimizing the conflict in pairwise synchronization results.

When the videos are synchronized, our system asks the user to provide an area that covers possible locations for the event. Then, our localization system starts downloads images from Google Street View and Flickr near the location. Low-level visual features (SIFT) are extracted from the images and matched to each of the video. The GPS of the closest matched image is assigned to the corresponding video.

Figure 2 shows the interface of our system after video synchronization and localization. The videos are put into a timeline at the top and a map on the left. Users can easily browse through the event timeline and analyze the event.

Gunshot Detection and Crowd Size Estimation

After synchronization and localization of the video collection, users can then utilize the gunshot detection and crowd size estimation tool for further event information retrieval. The gunshot detection tool can show users which segments within the videos contain gunshots. A pre-trained bag-of-audio-words gunshot model is used to detect gunshot on segments of each video and positive segments above a certain threshold are combined.

To estimate the crowd size in a video, we fine-tuned the faster-RCNN model(Ren et al. 2015) and its region proposal network on MS COCO dataset to detect and localize persons with bounding boxes in each video frame. The video segments with a crowd size above a threshold are then presented to the users for further analysis.

Person Tracking and 3D Reconstruction

In order to fully reconstruct the trajectory of a person, it is important that our system can track a person or object in 3D space. To achieve that, we combine person tracking (Andriluka, Roth, and Schiele 2008) with 3D reconstruction (Pollefeys et al. 2008) to allow users to search for a person of interest. In detail, we first reconstruct the whole 3D point cloud model by merging two or more models that were reconstructed from Google Street View images and frames from relavent videos via VisualSFM (Wu 2013), and then use the result of the detection of the person as well as the camera location, which is obtained while we reconstruct the point cloud, to infer where the person locates in 3D space. Finally, the tracking result of a person of interest is transformed into a trajectory in 3D space. Figure 3 shows an example person search results of our system. As we can see, the tracked person of interest in the video is bounded with a red box, and is also located in the 3D space.

Figure 3: Person Tracking and 3D Reconstruction Results

Acknowledgement

This project was conducted in partnership with Carnegie Mellon's Center for Human Rights Science (http://www.cmu.edu/chrs). The authors would like to thank the MacArthur Foundation, Oak Foundation, and Humanity United for their generous support of this collaboration.

References

Andriluka, M.; Roth, S.; and Schiele, B. 2008. People-tracking-by-detection and people-detection-by-tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Aronson, J. D.; Xu, S.; and Hauptmann, A. 2015. Video analytics for conflict monitoring and human rights documentation.

Collet, A.; Chuang, M.; Sweeney, P.; Gillett, D.; Evseev, D.; Calabrese, D.; Hoppe, H.; Kirk, A.; and Sullivan, S. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34(4):69:1–69:13.

Pollefeys, M.; Nistér, D.; Frahm, J.-M.; Akbarzadeh, A.; Mordohai, P.; Clipp, B.; Engels, C.; Gallup, D.; Kim, S.-J.; Merrell, P.; et al. 2008. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision* 78(2-3):143–167.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Wu, C. 2013. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, 127–134. IEEE.

Zhang, Y.; Gibson, G. M.; Hay, R.; Bowman, R. W.; Padgett, M. J.; and Edgar, M. P. 2015. A fast 3d reconstruction system with a low-cost camera accessory. *Scientific reports* 5.