SEEING AS IT HAPPENS: REAL TIME 3D VIDEO EVENT VISUALIZATION

Yueming Yang ⋆ *Ming-Ching Chang* † *Peter Tu* † *Siwei Lyu* ⋆

⋆ University at Albany, State University of New York, NY, USA
† GE Global Research, Niskayuna, NY, USA

ABSTRACT

We present a video event visualization system that can render steerable 3D views of tracked targets onto a reconstructed 3D site representation. The framework takes object tracking meta-data generated from a multi-camera event tracking system as input and produces an immersive 3D playback as a representation of the observation. This 3D representation can provide users *seeing as it happens* of the events for surveillance applications. Our system can further "animate" the virtual viewing camera and generate a first-person immersive playback of the event, either from the trajectory of a specified real-world target or from a virtual avatar. Such synthetic view can provide additional insights for event recognition for on-line monitoring, investigation, and forensic applications.

Index Terms— 3D video, surveillance, event summarization, immersive visualization, animation.

1. INTRODUCTION

With the pervasive installations of large camera networks and recent automatic visual detection and tracking algorithms, the presentation of video analytics outcome has become an emerging issue. This is because as events occur in the 3D physical world, human recognition and reasoning of such events are most effective with the 3D spatial temporal context. However, when displayed as 2D videos, such activities are greatly hampered by the limited viewing angles [1], and are further exacerbated by the distractions introduced when multiple views of the same scene are available. To this end, we focus on producing a holistic visualization of video analytics outcome in a 3D video representation. We propose a *"seeing as it happens"* video event summarization system that can integrate the analytical results of multiple video feeds into a 3D visualization framework.

The challenge on building an adequate mapping from 2D video observations to the underlying 3D world has been studied extensively in computer vision (i.e., projective and multi-view geometry [2]). Here we choose an alternative route to go from 2D to 3D for the visualization purpose. We create a 3D visualization based on the fusion of the moving *foreground* objects (targets being tracked) and the static *background* (3D site model) with Structure-from-Motion (SfM) from a collection of images. We use a robust multi-view multi-target tracking system [3], and communicate the tracked location using real-time message passing. Our 3D visualizer then show moving targets together with recognized events to generate a realistic synthetic 3D video summarization. Fig. 1 provides an overview of the system. Our system can "animate" the virtual viewing camera(s) arbitrarily and generate a steerable, first-person immersive replay of the event. The replay can be performed from the trajectory of a real-world target or a virtual avatar with a customizable viewpoint, to better suit the applications such as on-line surveillance, investigation, and forensic operations.

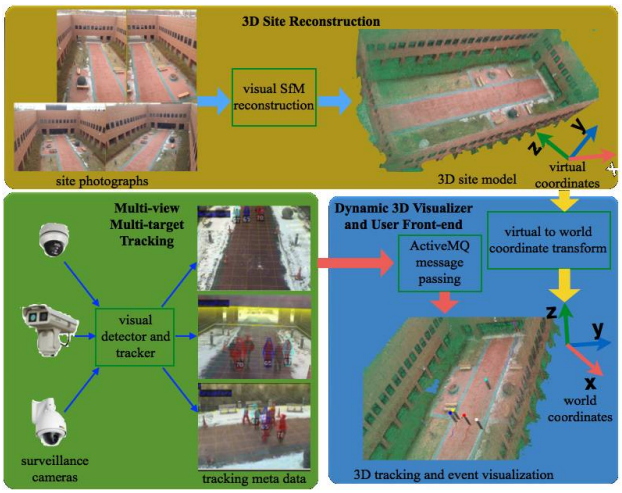

Fig. 1. *The proposed* **3D video event visualization framework** *consists of three components: (1) 3D site construction, (2) target tracking, and (3) 3D tracking visualization.*

The main contributions of our work are summarized as follows

- First, to the best of our knowledge, our work is the first attempt to integrate tracking with visualization using a high-resolution 3D site model obtained from recent SfM methods.
- Secondly, in order to visualize a *3D video*, where at each snapshot the user can steer and visualize the space-time volume from any viewing angle, we visualize the moving targets using a simple *dual textured*

This work is jointly supported by (1) grant award #2013-IJ-CX-K010, National Institute of Justice, US Department of Justice, and (2) Faculty Research Award Program (2012-2014), SUNY Albany.

planar model (see §3.4).

- A third contribution of this work is a general RANSAC+SVD approach for the computation of the transformation between 3D virtual coordinates and the true physical coordinates obtained in the videos.

The rest of the paper is organized as follows. Section 3.1 describes our 3D site model construction process. We use a standard multi-view tracking system to perform target tracking, where the observations are broadcast via message passing (§ 3.3). Target visualization and 3D video replay are described in § 3.4. Fig. 2 summarizes the major components of our system.

2. BACKGROUND

3D reconstruction. 3D site models can be obtained from: *(i)* a CAD model [4], *(ii)* a simple polygon model [5] with less details and fidelity, or *(iii)* using expensive LiDAR or 3D scanning devices which still yield less geometrical details. Currently, advanced COTS 3D sensors such as the Velodyne LiDAR and Carnegie Robotics range sensors can generate 3D point cloud on-the-fly. New mobile 3D sensors such as the Occipital structure sensor for iPAD can perform an instant site survey, however only for in-door settings. In comparison, the SfM based method (e.g., [6]) only requires one-time site photo shooting and yields a high quality 3D model. The process is flexible and applicable in all conditions and achieves high quality reconstruction with a sufficient number of images. Furthermore, the dense 3D point cloud output from the SfM method can be further processed to obtain a textured surface mesh [6].

3D video visualization in surveillance. The idea of constructing an augmented 3D environment for the purpose of visualizing video analytic outcomes is not new [4, 5, 7–9]. Existing works can be categorized using several criteria: (1) how the 3D model is obtained (e.g. using a CAD model, 3D point cloud or mesh generated from a scan), (2) how the target tracking and scene change detection is performed, and (3) how the 3D visualization is updated from video feeds [10]. The early Sarnoff VideoFlashlight system [8] provided immersive visualization by offline fusion of multiple video streams on top of a 3D site model. The USC Augmented Virtual Environments (AVE) system [5] projected oriented bounding boxes of moving objects with textured images (from video frames) onto a 3D scene model obtained from a range LiDAR sensor, where the final 3D scene model has missing details and is unrealistic. Fleck et al. [9] described a surveillance system based on a distributed network of smart cameras, where the XRT protocol is used for communication, and DirectShow is used for visualization. Their 3D site model was obtained using Laser scan and less geometrical details are preserved after post processing.

Large-scale visualization. Video visualization in multimedia and commercial applications is under active exploration.

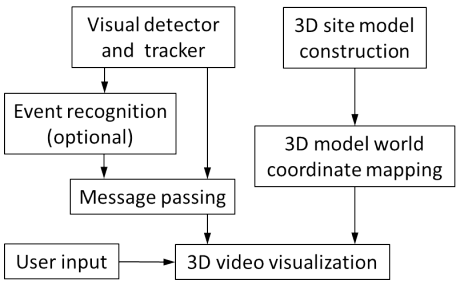

Fig. 2. *Components of the 3D video visualization system.*

The early CMU DARPA VSAM system [7] integrated multi-sensor video surveillance with a terrain database, where several representations were proposed for situational awareness visualization. The MIT HouseFly system [4] created a synthesized world view by de-warping and stitching fisheye camera views in an indoor-environment. CAD model of the house was used as 3D site model, such that the video could be projected and rendered directly onto planar surfaces. Trajectory tracks were shown as line drawings, and (unlike in our system) there was no visualization of the moving targets as 3D models in the scene. Other works in this category focused on large scale city-wide geographical visualization [11] and video visualization with contextual support [12].

3. METHOD

The goal of our system is to produce a holistic 3D event visualization from video analytics outcome. Visualization is created by fusing multi-view target tracking at the foreground with the 3D site model at the background, such that a complete 4D space-time volume can be visualized.

3.1. 3D Site Model Construction

Our 3D site model is constructed in an one-time site survey process. Starting with taking a sufficient set of images from the site, we follow the following steps proven in previous works [13, 14] to generate a 3D site model:

1. Pre-processing (removal of low-quality, blur, and over-exposure images). It is best to collect site photos in good weather and illumination conditions.
2. SIFT-like keypoint detection and feature extraction for pairwise matching using RANSAC.
3. Sparse correspondence and reconstruction using Bundle Adjustment.
4. Dense correspondence by stereo matching, which is a crucial step in producing a detailed 3D model using information from confirmed, overlapping image pixels in an epipolar search.
5. Geometric modeling and texture mapping, which involves polygonal mesh reconstruction from point clouds and the mapping of pixel colors and textures onto the surface mesh.

Table 1. **Procedure** *EstTransformRANSAC*($\mathbf{X}, \mathbf{Y}, K, \tau$).

$index \leftarrow (1 \text{ to } n)$
$n\_inliers \leftarrow 0$
for $k = 1$ to K **do**
  idx = randomly select 6 numbers from $index$
  $\mathbf{X}_i \leftarrow \mathbf{X}(idx), \mathbf{Y}_i \leftarrow \mathbf{Y}(idx)$
  $\mathbf{R}_i, s_i, \vec{c}_i \leftarrow$ **EstTransformSVD** $(\mathbf{X}_i, \mathbf{Y}_i)$
  $e_i \leftarrow |s_i \mathbf{Y}_i - \mathbf{R}_i \mathbf{X}_i - \vec{c}_i|$
  $n_i \leftarrow$ number of items in e_i which are $\leq \tau$
  if $n_i \geq n\_inliers$ **then**
    $n\_inliers \leftarrow n_i$
    $\mathbf{R}, s, \vec{c}, e \leftarrow \mathbf{R}_i, s_i, \vec{c}_i, e_i$
  end if
end for
return $\mathbf{R}, s, \vec{c}, e, n\_inliers$

Off-the-shelf SfM tools such as the VisualSfM [15] can be used to generate both the sparse and dense point cloud reconstruction in the steps [16]. A recent tool CMPMVS [6] can be used to further generate a textured surface mesh model. Point Cloud Library (PCL) and MeshLab provide additional tools for 3D model processing. In the current prototype system, we only use raw 3D point clouds for the site model, but it is possible to use refined site model such as those based on textured surface mesh [6].

3.2. World Coordinate System Transformation

Since the SfM method produces the model in a virtual coordinate originated from camera projections, the transformation (scaling, translation, and rotation) mapping to world coordinate needs to be calculated. In fact, such coordinate transformation is required for most methods, no matter how the 3D model is obtained (unless it is generated in the world coordinate in the first place). We compute the transformation (scaling, translation, and rotation) between the SfM 3D model in virtual coordinate and the world coordinate by estimating the mapping between a set of 3D landmark points. Rigid transformation up to a scaling change between the two systems can be calculated using singular value decomposition (SVD) [17].

Specifically, given two sets of 3D landmark points, $\mathbf{X} = \{\vec{x_1}, ..., \vec{x_n}\}$ and $\mathbf{Y} = \{\vec{y_1}, ..., \vec{y_n}\}$ as matrices of points in world and virtual coordinates respectively, we denote the procedure **EstTransformSVD** to estimate the scaling s, translation $\vec{c}$, and rotation $\mathbf{R}$ such that $\vec{y_i} = \frac{1}{s}(\mathbf{R}\vec{x_i} + \vec{c})$. The solution is formulated as a least squares optimization problem, $\min_{\mathbf{R},s,\vec{c}} \sum_{i=1}^{n} \|s\vec{y_i} - \mathbf{R}\vec{x_i} - \vec{c}\|^2$, where the rotation can be solved by compensating the translation:

$$\vec{c} = \frac{s}{n}\sum_{i=1}^{n}\vec{y_i} - \frac{1}{n}\mathbf{R}\sum_{i=1}^{n}\vec{x_i}. \quad (1)$$

We denote $\tilde{\vec{x}}_i = \vec{x}_i - \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i$ and $\tilde{\vec{y}}_i = \vec{y}_i - \frac{1}{n}\sum_{i=1}^{n}\vec{y}_i$ as the centered 3D points. The least square problem becomes $\min_{\mathbf{R},s} \sum_{i=1}^{n} \|s\tilde{\vec{y}}_i - \mathbf{R}\tilde{\vec{x}}_i\|^2$. Denote $\tilde{X}$ and $\tilde{Y}$ as the two matrices formed with $\tilde{\vec{x}}_i$ and $\tilde{\vec{y}}_i$ as columns, respectively. Rotation $\mathbf{R}$ can be solved by SVD of $\tilde{X}\tilde{Y}^T = U\Gamma V^T$, i.e.,

$$\mathbf{R} = UV^T. \quad (2)$$

Scaling s can then be obtained using:

$$s = \frac{\sum_{i=1}^{n}\tilde{\vec{y}}_i^T\mathbf{R}\tilde{\vec{x}}_i}{\sum_{i=1}^{n}\tilde{\vec{y}}_i^T\tilde{\vec{y}}_i} = \frac{\text{tr}(\mathbf{R}\tilde{X}\tilde{Y}^T)}{\text{tr}(\tilde{Y}\tilde{Y}^T)}. \quad (3)$$

Manual landmark specification inevitably contains small inaccuracy, and outliers can cause large error due to the mismatch in the least square optimization. We derive a robust RANSAC scheme that can generate a reliable transformation, as shown in the pseudo code in Table 1. K is the number of RANSAC iterations, τ is the threshold between inliers and outliers. The output are the resulting rotation matrix $\mathbf{R}$, scaling s, translation vector $\vec{c}$, and final error norm e.

Degeneracy analysis. The virtual-to-world coordinate transformation can get into degenerate solutions if $\mathbf{X}$ and $\mathbf{Y}$ do not form full rank matrices. This would be the case when points $\{\vec{x}_1, \cdots, \vec{x}_n\}$ or $\{\vec{y}_1, \cdots, \vec{y_n}\}$ are *collinear* or *coplanar*, which leads to matrices $\tilde{X}$ or $\tilde{Y}$ to be rank deficient (with rank one or two, respectively). In the collinear case, only one degree of freedom exists between the two sets of colinear points (i.e. the angle between two intersecting lines). In the coplanar case, two degree of freedom exists out of the 3D rotation. Without adding another pair of corresponding points that are not coplanar, the solution can contain up to a *sign flip* of the corresponding points, when applying the estimated transform. The problem can be avoided by providing additional landmark pairs that leads both $\mathbf{X}$ and $\mathbf{Y}$ to be full rank.

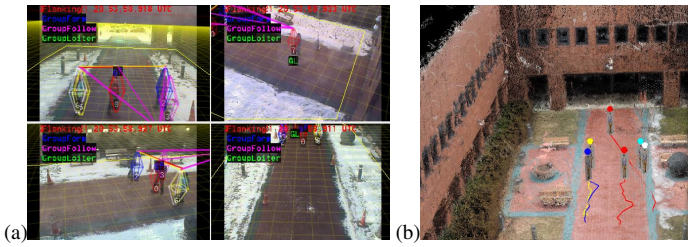
(a) (b)

Fig. 3. *(a)* **Multi-view tracking and event recognition system** *detecting group activities [18]. (b) 3D view showing target identity and track trajectories can be manipulated with free-style zooming, rotating, and panning for better surveillance visualization.*

3.3. Video tracking and message passing

Our video tracking system [3] consists of multiple calibrated cameras cooperatively tracking pedestrians in real-time (see Fig. 3). Foreground pixels are segmented out to vote for precomputed image grid locations to form a set of detection candidates [19]. Such grid based detection voting significantly reduces false positives arising from occlusions and crowdedness. Detections are consumed by a central tracker which runs on either a single-view or multi-view scenario, to *(i)* associate detections with existing tracks based on spatial proximity or *(ii)* initiate new tracks. Tracking states are estimated using

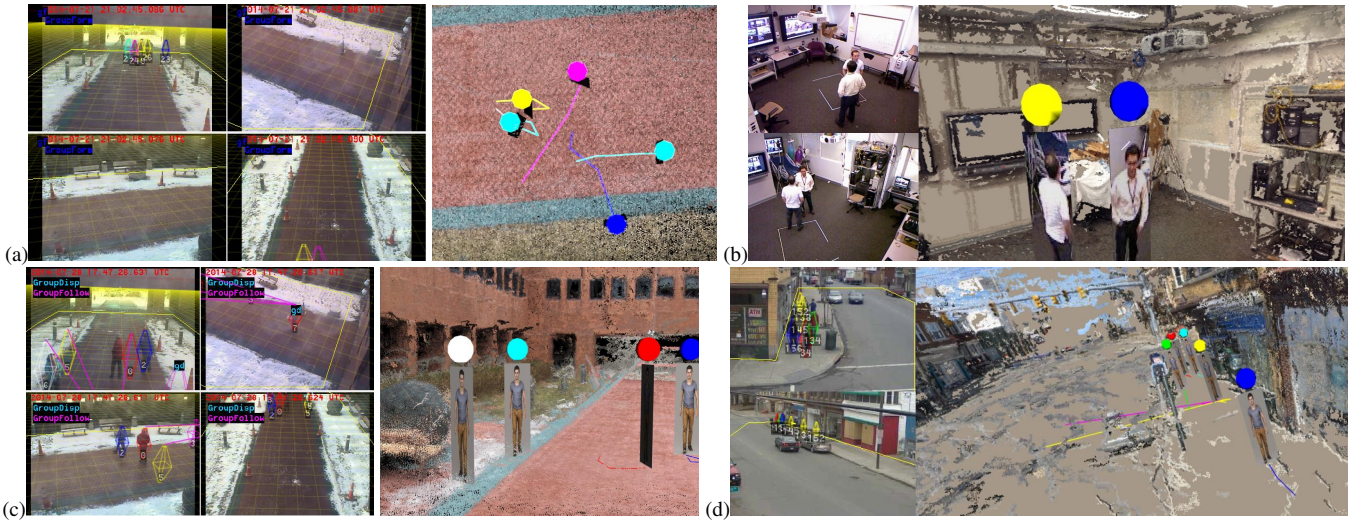

Fig. 4. *(a,c)* **GE Courtyard:** *Events replay from (a) an operator selected view and (c) avatar's view (see text). (b)* **GE Smartroom:** *Visualization of a social interaction session. (d) Visualization of pedestrians on a public street site.*

a Kalman filter performed in the world coordinate ground plane. Tracking meta-data broadcasting is implemented using Apache ActiveMQ message passing to ensure real-time synchronization with the 3D visualizer.

3.4. 3D Visualization and navigation of video events

We use the Visualization Toolkit (VTK) for real-time 3D rendering on a standard PC. We implemented two basic approaches to visualize each moving target: (1) rendering as an dual-faced, oriented planar box with time-synchronized texture mapping that is obtained from cropping the original video, and (2) rendering from fixed given image(s). Other approaches include rendering the target as a cylinder, a box model, or a refined polyhedral model, when sufficient depth information or the geometrical details are available. Target orientation can be computed from facial gaze tracking or pose estimation (if available), or using velocity estimation. In the multi-cam scenario, the image box from the camera with smallest angle between the target orientation and camera orientation (i.e. most frontal view) is used for visualization.

4. RESULTS

We have evaluated the system at two outdoor sites (GE Research courtyard and a public street sidewalk) for surveillance application and an indoor site (GE Smartroom) for social interaction analysis. We collaborated with a local police department to gather feedback from their experience and insight from real-world surveillance needs. Our capability in visualizing events in summary mode and immersive patrol mode are particularly useful from the practitioner's feedback. Fig. 4(a) shows an outdoor group formation event which is hard to recognize in the 2D videos, while the 3D video clearly illustrates the scenario in a top-down view. Fig. 4(b) shows an example of two subjects interacting in an indoor site, where the target is visualized with synchronized video texture mapping. Fig. 4(d) shows an outdoor example for a crowd meeting event, where the 3D video from a customized bird's-eye view can identify the position of each target more clearly.

First-person immersive replay of an event. Our system can visualize the 3D playback video from the angle of a selected target's view, as if it is seen from the viewpoint of that person moving along his trajectory. Such first-person immersive replay can provide on-line security monitoring as well as forensic investigation with valuable insight. Fig. 4(c) shows an example, where the 3D view is automatically created from the subject that is identified as a victim in a group flanking event. In this case, security personnel can get on-line notification with an immersive view of the event.

Application in Augmented Reality (AR). We also integrated the 3D video system with the Oculus VR Rift head mount to experiment on AR applications. To simulate an AR experience in an off-line mode on our in-door site as shown in Fig. 4(b), our system first tracks the user's location and facial gaze direction using Pan-Tilt-Zoom (PTZ) cameras. We can then playback what the user has seen in our 3D video visualizer. For an on-line experiment using the VR head display, the gyroscope and accelerometers can provide real-time orientation input to update the visualization. We believe such on-line immersive experience of recorded events should be helpful for forensic and investigation usage.

5. CONCLUSION.

We presented a 3D video visualization system that can playback space-time video events in an interactive and immersive way. Future work includes the improvements on more accurate avatar and site model visualization, e.g., using surface mesh with texture mapping and updating the 3D model using the video feed, as well as the integration of advanced video analytic offerings and wearable equipment.

6. REFERENCES

[1] Gerwin de Haan, Huib Piguillet, and Frits Post, "Spatial navigation for context-aware video surveillance," *IEEE CG&A*, vol. 30, no. 5, pp. 20–31, 2009.

[2] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2000.

[3] T. Yu, Y. Wu, N. Krahnstoever, and P. Tu, "Distributed data association and filtering for multiple target tracking," in *CVPR*, June 2008.

[4] Philip DeCamp, George Shaw, Rony Kubat, and Deb Roy, "An immersive system for browsing and visualizing surveillance video," in *MM*. ACM, 2010, pp. 371–380.

[5] Ismail Oner Sebe, Jinhui Hu, Suya You, and Ulrich Neumann, "3D video surveillance with augmented virtual environments," in *First ACM SIGMM international workshop on video surveillance*. ACM, 2003, pp. 107–112.

[6] Michal Jancosek and Tomas Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *CVPR*. IEEE, 2011.

[7] Robert T Collins, Alan Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al., "A system for video surveillance and monitoring," in *tech report CMU-RI-TR-00-12*. Robotics Institute, Carnegie Mellon University, 2000.

[8] HS Sawhney, A Arpa, R Kumar, S Samarasekera, M Aggarwal, S Hsu, D Nister, and K Hanna, "Video flashlights: real time rendering of multiple videos for immersive model visualization," in *ICPS*. ACM, 2002, vol. 28, pp. 157–168.

[9] Sven Fleck, Florian Busch, Peter Biber, and Wolfgang Straber, "3D surveillance a distributed network of smart cameras for real-time tracking and its visualization in 3D," in *CVPR Workshop*. IEEE, 2006, pp. 118–118.

[10] Wenyi Zhao, David Nister, and Steve Hsu, "Alignment of continuous video onto 3D point clouds," *PAMI*, vol. 27, no. 8, pp. 1305–1318, 2005.

[11] Kihwan Kim, Sangmin Oh, Jeonggyu Lee, and Irfan Essa, "Augmenting aerial earth maps with dynamic information," in *ISMAR*. IEEE, 2009, pp. 35–38.

[12] Yi Wang, David M Krum, Enylton Machado Coelho, and Doug A Bowman, "Contextualized videos: Combining videos with environment models to support situational understanding," *TVCG*, vol. 13, no. 6, pp. 1568–1575, 2007.

[13] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch, "Visual modeling with a hand-held camera," *IJCV*, vol. 59, no. 3, pp. 207–232, 2004.

[14] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski, "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, Oct. 2011.

[15] Changchang Wu, "VisualSFM: A visual structure from motion system," http://ccwu.me/vsfm/, 2011.

[16] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multiview stereopsis," *PAMI*, vol. 32, no. 8, pp. 1362–1376, 2010.

[17] David W Eggert, Adele Lorusso, and Robert B Fisher, "Estimating 3-D rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[18] Ming-Ching Chang, Nils Krahnstoever, and Weina Ge, "Probabilistic group-level motion analysis and scenario recognition," in *IEEE ICCV*, 2011, pp. 747–754.

[19] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins, "Multi-view detection and tracking of travelers and luggage in mass transit environments," in *PETS*, 2006.