Tracking Gaze Direction from Far-Field Surveillance Cameras

Karthik Sankaranarayanan[†] Ming-Ching Chang[‡] Nils Krahnstoever[‡]
[†]Dept. of Computer Science and Engineering, Ohio State University
[‡]GE Global Research Center, 1 Research Circle, Niskayuna, NY, USA
sankaran@cse.ohio-state.edu {changm, krahnsto}@research.ge.com

Abstract

We present a real-time approach to estimating the gaze direction of multiple individuals using a network of far-field surveillance cameras. This work is part of a larger surveillance system that utilizes a network of fixed cameras as well as PTZ cameras to perform site-wide tracking of individuals. Based on the tracking information, one or more PTZ cameras are cooperatively controlled to obtain close-up facial images of individuals. Within these close-up shots, face detection and head pose estimation are performed and the results are provided back to the tracking system to track the individual gazes. A new cost metric based on location and gaze orientation is proposed to robustly associate head observations with tracker states. The tracking system can thus leverage the newly obtained gaze information for two purposes: (i) *improve the localization of individuals in crowded settings, and* (ii) *aid high-level surveillance tasks such as understanding gesturing, interactions between individuals, and finding the object-of-interest that people are looking at. In security application, our system can detect if a subject is looking at the security cameras or guard posts.*

1. Introduction

Automatically understanding and recognizing behaviors from surveillance video in urban environments such as mass transit, schools and prison yards is challenging due to a large number of factors. Crowdedness and lack of resolution in a typical surveillance camera makes the accurate localization and tracking of individuals difficult and introduces uncertainty to subsequent reasoning stages. In such environment, one can at most hope to perform the localization of individuals without further knowledge about body pose or orientation. Although body pose estimation has been studied in the context of surveillance [7], it is far from real-time performance. In this work we study part of the problem of estimating body pose, or more specifically head orientation of individuals in real-life videos, in order to: *(i)* reason over the orientation of individuals, in particular as part of pair-wise interactions between people, and *(ii)* to understand what people are looking at. The former feature is important for analyzing *group interactions* for which it is important to know *e.g.* if two people that are physically close, facing each other (mutual gaze), facing the same direction, or looking away from each other. The latter feature is of particular relevance to applications such as *retail security* [15] and *facility protection*, where security operators are interested in whether people are surveying (*i.e.*, looking at) camera locations, guard and clerk movements, or similar things of interest to a person who is about to commit a crime.

Head pose and gaze estimation from standard resolution surveillance views is challenging at best and impossible in many other cases. Hence, we utilize a hybrid approach where we perform multi-camera multi-target person tracking within a network of fixed cameras, and pass the tracking information to drive one or more PTZ cameras to zoom-in on individuals (detailed in §3). Face tracking and head pose estimation is then performed within the close-up views of these PTZ captures, fusing information hand-in-hand with the person tracker. The face location and head pose information is mapped back into an unified coordinate system, where it can be used to improve the tracking performance (under crowded conditions) and perform higher level reasoning over the pose (*e.g.*, to analyze social interactions [21] or behaviors [6]). The proposed system operates in real-time under challenging imaging conditions. Due to the relatively larger computational burden in face detection, the person tracking and face tracking must operate *asynchronously*. We will elaborate in §4 how we integrate the information dynamically and consistently in a flexible framework.

This work is the first (to the best of our knowledge) that investigates the augmentation of multi-camera tracking with multi-PTZ facial gaze tracking in the surveillance domain. Our main contribution is a unified approach to robustly fusing together person tracking information with asynchronous PTZ facial tracking information. Our system is flexible to

operate on either a single or multiple cameras. While the use of multiple cameras is not a hard requirement, it does improve the overall tracking performance, in particular in situations where multiple PTZ cameras view a group of people from a set of different directions.

The paper is organized as follows. We will describe related work in §2, the overall system in §3, and our approach to gaze analysis in §4. We will present real-time experimental results in a variety of settings in §5. We will discuss the results in §6 and conclude the paper in §7.

2. Related Work

Head pose estimation from one or more views has been extensively studied over the past 15 years with applications ranging from robotics, human computer interaction [18], driver assistance, and virtual reality. The recent review article of Murphy-Chutorian and Trivedi [17] provides an excellent summary and comparison between various approaches including using appearance, non-linear regression, non-rigid model fitting, tracking and hybrid methods. Among them we highlight automatic methods that detect and track head pose from single or multi-view videos in an unconstrained environment. Works in this category [3, 2] involve head detection followed by pose estimation and tracking, *e.g.* using Kalman filter [9] or particle filtering [4, 12].

Hu *et al.* [8] fit a single Active Appearance Model (AAM) simultaneously to multiple synchronous face images to estimate head pose, with a requirement that the head image quality must be high enough. Voit *et al.* [20] use a neural network classifier to estimate head pose from each view and use Bayesian dynamics to merge estimations, with a strong assumption that the individuals are sitting in fixed seats such that no tracking or camera zooming is required. Lanz and Brunelli [12] track body parts using a Bayesian framework over shape and appearance and estimate head orientation across multiple views using particle filtering. Canton-Ferrer et al. [5] assume head location is known and estimate its orientation by back-projecting the skin appearance patches onto the estimated 3D head model and employ a particle filter in tracking across multiple views. In a recent work, Bäuml et al. [4] assume head location is known and track face pose across a distributed camera network for recognition and re-identification. The face tracker runs separately and independently from the person tracker, and there is no attempt to exploit the advantage of multiple overlapping facial views.

To the best of our knowledge, all existing works make strong assumptions that *(i)* the head locations are (roughly) known and *(ii)* head image quality is (reasonably) good, so as to simplify the problem of simultaneous tracking and pose estimation. A major difference that sets the proposed work apart is that our system operates in a more unconstrained, challenging environment in live, where both person locations and head poses are unknown, in addition to that the close-up PTZ views are dynamically changing as well. The person tracking and PTZ face tracking thus must be performed asynchronously. We try to bring together various observations and fuse them into a consistent, central tracking scheme (see §4).

3. System Description

In this section we will provide a brief outline of our tracking system as well as the type of environment we are addressing in this paper.

3.1. Video Tracking System

The tracking system that we utilize [10, 22, 21] comprises of multiple calibrated static cameras tracking cooperatively in a synchronized fashion. For each view, the position and image dimension of each person at all possible 3D locations in the scene are estimated using calibration. Foreground pixels from online tracking are used to vote for these precomputed image locations to form a set of (foreground) detections [10]. This effectively leverages the calibration information to significantly reduce false positives arising from occlusions and crowdedness.

The set of detections for each view are then projected onto the ground plane in 3D in order to further disambiguate any confusion due to occlusions and crowdedness. These projections are consumed by a centralized tracking system that either *(1)* associates detections with existing tracks based on spatial proximity or *(2)* initiates new tracks. The states of tracks are estimated by a standard Kalman filter, performed in the world reference ground plane. The system is designed to maintain tracks across camera boundaries in order to perform site-wide tracking.

3.2. Pan Tilt Zoom Control

To enable face detection and gaze estimation of uncooperative individuals from a distance, the tracking system controls multiple pan tilt zoom (PTZ) cameras automatically [11]. The control algorithm pursues the goal of optimally scheduling the PTZ cameras in real-time under a variety of performance objectives. The control system provides each PTZ camera with a continuously evolving *schedule* that describes what targets to visit in what order. Schedules are planned several target capture steps into the future based on the current and predicted motion of observed individuals. A given schedule is assigned a probability of achieving the goal of capturing high quality facial shots of all tracked individuals. The quality of facial shots is governed by the distance of individuals from the camera, the angle at which a face is captured, and the accuracy with which a person is being located by the tracking system. A control strategy is

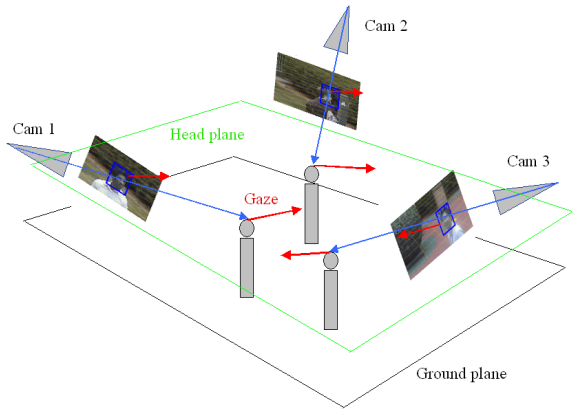

Figure 1. Projecting the detected faces to the head plane and calculating the gaze vectors.

chosen by selecting the schedule with the highest probability from the set of all possible schedules. Details of our PTZ control approach are provided in [11].

Contrary to the fixed cameras, the poses of PTZ cameras change over time. As part of the control, the system's task is to estimate the time varying projection matrices for the PTZ cameras, given as $\mathbf{P}_i^{\mathrm{P}}[t]$, where $i \in \{1, \ldots, N^{\mathrm{P}}\}$ the index over all PTZ cameras and t the time.

In the majority of the experiments shown below we utilize a testbed that is equipped with four fixed cameras for tracking and four PTZ cameras for face capture. It should be noted that the ratio between the number of tracked individuals and available PTZ cameras makes an impact on system performance. A larger number of individuals typically requires a larger number of PTZ cameras or the utilization of high resolution mega-pixel cameras.

4. Gaze Analysis

In this section, we describe in detail our method to obtain gaze tracks for each of the individuals using a Kalman filter based gaze orientation tracking system. We assume that a person's head pose generally aligns with one's gaze direction, even though the head pose is only a coarse estimate of visual gaze (*i.e.* eye ball) direction [17].

As the PTZ cameras locate the individuals, they zoom into the estimated head locations, such that face detection can be performed in each PTZ views. The detected face locations are projected back to the 3D head-plane to obtain an estimate of the person's head position in the 3D world (§4.1). Meanwhile, the head pose is estimated from the face image and converted to a 3D gaze vector using the PTZ camera's rotation matrix (§4.2). This is performed for each PTZ camera that is currently obtaining individual head/gaze location and orientation, see Fig. 1. We develop a Kalman filter based gaze tracker that operates on angular coordinates of the gaze vectors. The tracked gaze orientation augments the person tracker (which is in fact another Kalman filter tracker) that operates on the ground plane. We utilize the Hungarian algorithm [16] to associate the location and orientation of the faces to the individuals by minimizing a cost function (§4.3). Note that our state and observation spaces are both in angular coordinates so there are no non-linearities involved. We track transformed observations rather than raw observations that would be non-linearly tied to the state space. Once the observations corresponding to different trackers are obtained, a Kalman filtering update is performed.

4.1. Face Detection and Projection

We use an off-the-shelf face detector [19, 1] to detect faces in the PTZ views. The algorithm is chosen because it works well with a wide variation in head poses, from frontal to profile views. It is able to detect faces in fairly low-resolution video. In this work, we deal with 640×480 pixel images and the system controls the PTZ to capture facial shots with a rough resolution of 20-30 pixels eye-to-eye. The face detector has a low false-positive rate in our system. As we will demonstrate later, remaining false-positive detections can be handled robustly by the gaze tracker.

Face detections in each image view are used to estimate each individual's head location in the 3D world. This is done by *(i)* projecting a ray from the optical center of the PTZ camera $\mathbf{P}_i^{\mathrm{P}}[t]$ through the center of the face location in the image plane, and then *(ii)* finding the intersection of this ray with the head-plane, which is assumed parallel to the ground plane at a height of 1.8 meters; see Fig. 1. Also, the width and the height of the face are used to estimate a covariance confidence level for the face location. The covariance is projected to the groundplane using an unscented transform (UT) from the image to head plane, followed by downprojection to the groundplane. The above operation is performed for all face detections in all PTZ views to obtain multiple head locations for all individuals, with estimates of (mean, covariance) pair simultaneously. All observation information is organized in a unified 3D world coordinate system, where a central tracker can operate in an integrated manner.

Along with the face detection, face orientation (head pose) can be estimated from either using *(i)* active appearance models (AAM) matching [13, 14], or *(ii)* face feature detection followed by pose estimation (as done in [1]). In theory, one could model the full egocentric parameters of the head: the *yaw* (left/right), *pitch* (chin up/down) and *roll* (around "nose" axis). However the roll direction is not significant for gaze estimation, and the pitch is often unreliable. We thus mainly focus on tracking the yaw orientation. As we will show in §4.2, we can still estimate the *global* head pitch, because its 3D pose is viewed from different

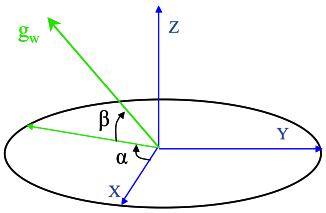

Figure 2. Relation of the gaze egocentric angles (yaw α and pitch β) with the gaze vector g_w.

PTZ cameras mounted at different heights.

4.2. Head Pose to 3D Gaze

In order to track gaze from multiple cameras, we need to transform it from local camera coordinates to the central 3D world space (to be tracked by the central tracker). To do this, the gaze vector (face normal) is first obtained in Cartesian coordinates in the camera space from the head pose angles, and transformed to the world space using the camera rotation matrix. Finally the transformed gaze vector is converted back in terms of egocentric angles (orientations) in the world space.

To first obtain the face normal, that is the gaze vector local to the camera coordinate system $\mathbf{g}_{im}$=(x_{im}, y_{im}, z_{im}) from the head pose yaw angle (ϕ_{im}), we use the following equations.

$$x_{im} = \cos(\phi_{im}), \quad y_{im} = \sin(\phi_{im}), \quad z_{im} = 0 \tag{1}$$

The rotation matrix of the PTZ camera is then used to convert the gaze vector from the local image space to the 3D world space. Using the technique of transforming normals, the gaze vector is multiplied by the transpose of inverse of the rotation matrix of a PTZ camera $\mathbf{P}^{\mathrm{p}} = [\mathbf{R}|\mathbf{t}]$ (we will ignore the PTZ index i in the following).

$$\mathbf{g}_w = \mathbf{g}_{im} * (\mathbf{R^{-1}})^{\mathbf{T}} \tag{2}$$

The transformed gaze vector $\mathbf{g}_w = (x_g, y_g, z_g)$ is converted to back to the egocentric angles representation of yaw (α) and pitch (β) in 3D space (see Fig. 2) using the following equations:

$$\alpha = \arctan(\frac{x_g}{y_g}) \tag{3}$$

$$\beta = \arctan(\frac{z_g}{\sqrt{x_g^2 + y_g^2}}) \tag{4}$$

At the end of this step, the head location and gaze orientation is obtained for one or more targets from multiple PTZ cameras projected to a common centralized 3D space.

4.3. Kalman Filtering for Gaze

In order to track the gaze orientations of the individuals, we extend the person tracker state by adding gaze information to it. The gaze state of a target (Θ) is modeled in terms of the two orientation angles (α, β), as well as their first derivatives — the angular velocities $(\dot{\alpha}, \dot{\beta})$. Therefore, $\Theta = \left[\begin{array}{cccc} \alpha & \beta & \dot{\alpha} & \dot{\beta} \end{array} \right]^{\mathbf{T}}$.

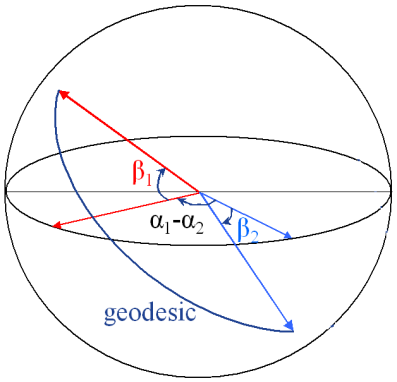

Figure 3. Geodesic distance between orientation vectors measured on the great circle.

The state transition model that relates the gaze state at time $k-1$ to the state at time k is given as

$$\Theta_k = \mathbf{F} * \Theta_{k-1} + \mathbf{w}_{k-1}, \tag{5}$$

where the state transition matrix ($\mathbf{F}$) using a constant velocity model is given as

$$\mathbf{F} = \left[\begin{array}{cccc} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right], \tag{6}$$

and $\mathbf{w}_{k-1}$ is the gaussian distributed process noise. The measurement model which extracts the orientation information from the gaze state (Θ) is given as

$$\left[\begin{array}{c} \alpha \\ \beta \end{array} \right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array} \right] * \Theta_k + \mathbf{v}_k, \tag{7}$$

where $\mathbf{v}_k$ is the gaussian distributed measurement noise.

4.4. Data Association for Faces

Since we are dealing with the tracking of multiple individuals, a necessary step is to associate the different face detections obtained from multiple cameras with their corresponding trackers so that the gaze states of the trackers can be updated appropriately. Note that even though the system knows the identity of an individual when the PTZ camera is allocated to look at, it is still possible that multiple face detections are obtained from a video frame, especially when individuals are close to each other. Therefore, the system may not know exactly which face in the image belongs to which individual. Consequently, this data association step becomes essential to resolve the ambiguities in assigning gaze detections to tracks reliably.

The data association module uses two cues to assign the detected faces to trackers appropriately: *(1)* Head location (from §4.1) and *(2)* 3D gaze orientation (from §4.2).

Let N_d be the number of face detections and N_t be the number of trackers at a given timestep k. In order to perform the data association, we need a distance metric to measure the distance between head observations $\mathbf{h}_i$ (where $1 \leq i \leq N_d$) and tracker states $\mathbf{t}_j$ (where $1 \leq j \leq N_t$). The following cost metric η is proposed to measure the distance between head observation $\mathbf{h}_i$ and a tracker state $\mathbf{t}_j$.

$$\eta(\mathbf{h}_i, \mathbf{t}_j) = \exp(-\frac{d(\mathbf{h}_i^{\mathbf{x}}, \mathbf{t}_j^{\mathbf{x}})}{\sigma_{\mathbf{x}}} - \frac{\lambda(\mathbf{h}_i^{\Theta}, \mathbf{t}_j^{\Theta})}{\sigma_{\Theta}}), \tag{8}$$

where $d(\mathbf{h}_i^{\mathbf{x}}, \mathbf{t}_j^{\mathbf{x}})$ is the Euclidean distance between the head observation's location on the head plane and tracker's location on the ground plane (ignoring the height difference). $\lambda(\mathbf{h}_i^{\Theta}, \mathbf{t}_j^{\Theta})$ is the geodesic distance (see Fig. 3) between the gaze orientation of the head observation and the tracker's current gaze orientation, which is calculated using the spherical law of cosines as

$$\begin{aligned}\mathbf{A}_{ij} = \lambda(\mathbf{h}_i^{\Theta}, \mathbf{t}_j^{\Theta}) = \arccos(&\sin\beta_{\mathbf{h}_i}\sin\beta_{\mathbf{t}_j} \\ &+ \cos\beta_{\mathbf{h}_i}\cos\beta_{\mathbf{t}_j}\cos(\alpha_{\mathbf{h}_i} - \alpha_{\mathbf{t}_j})).\end{aligned} \tag{9}$$

Using the above cost function, the distance between every pair of the i-th head detection and the j-th gaze track is calculated to build a cost matrix $\mathbf{A}_{ij}$ of size $N_d \times N_t$. The task of assigning the heads to the correct trackers is now a combinatorial optimization problem. For this, the Hungarian algorithm [16] is employed to find an optimal assignment of observations to trackers (by minimizing the cost) in polynomial time.

Once the faces have been assigned to their respective trackers, the standard update step of the Kalman filter is performed to update the gaze state of each individual being tracked. In order to visualize the gaze of the target at every point during tracking, the gaze angles (α, β) from the state vector of the Kalman filter are converted into their corresponding Cartesian representation as follows:

$$\hat{x}_{gaze} = \frac{\cos\alpha}{|\cos\alpha|}, \tag{10}$$

$$\hat{y}_{gaze} = \hat{y}_{gaze}\tan\alpha, \tag{11}$$

$$\hat{z}_{gaze} = \sqrt{\hat{x}_{gaze}^2 + \hat{y}_{gaze}^2}\tan\beta. \tag{12}$$

5. Experiments and Results

We first demonstrate results from the first part of the system, which is the face detection and gaze vector calculation from head poses. After that we perform experiments and demonstrate results with tracking the gaze of one or more individuals.

Figure 4. **Sequence A:** Face detections and head pose estimation from three PTZ camera views (top) projected to the three corresponding static camera views (bottom). Observe how qualitatively the gaze vectors are tracked from the visualization of a few trailing frames around the subject. Yellow mesh grids visualize the ground plane in projective view.

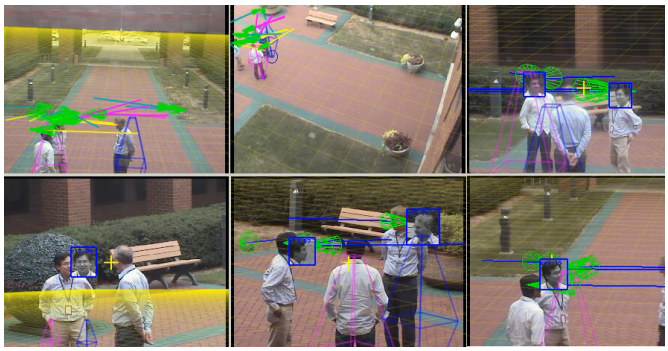

Figure 5. **Sequence B:** Face detections and gaze tracking for multiple individuals in the scene. Observe the asynchronous nature of the mixture of the person tracking (diamond outlines) and face detection updates (grayscale boxes), suggesting the need for a robust data association.

5.1. Gaze Observations from Head Pose

Our system consists of four fixed cameras and four PTZ cameras overlooking portions of a courtyard. As the individuals walk around, the fixed cameras are used to perform tracking and the PTZ cameras zoom into the calculated head location and performed face detections. These detections are then used to estimate the gaze vectors. Results from test sequence A are shown in Fig. 4. Face detections from the PTZ cameras provide different views and consequently different head poses. These poses are transformed to obtain 3D gaze vectors, which are visualized back in the static views. In the bottom row, different colors of the vectors correspond to gaze vector coming in from different cameras. Also shown are gaze vectors from a few trailing frames with reducing intensity.

Fig. 5 shows a few frames of another test sequence, where multiple individuals are tracked and their face detection are associated to form gaze tracks. Even though PTZ cameras are allocated to particular targets, multiple face de-

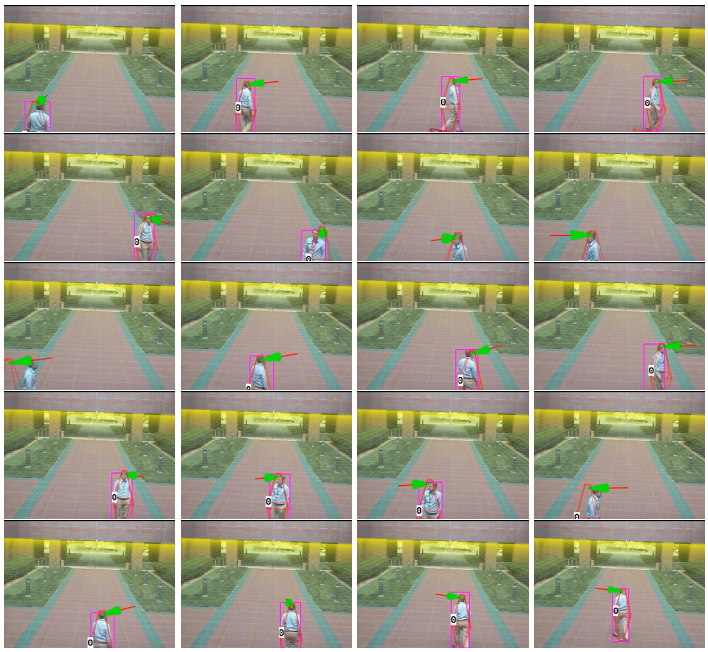

Figure 6. **Sequence A:** Few frames of simultaneous person tracking and gaze tracking at different time.

tections are obtained from each view. Consequently, the association of individual face detections to from gaze tracks becomes necessary (§4.4).

5.2. Gaze Tracking

Single-person gaze tracking: The gaze observations from previous section are used to update the state of the Kalman filters corresponding to each individual's gaze tracker. The gaze orientation vectors are then projected onto each camera views for visualization. Figs. 6 and 8 show such visualization from a few frames of gaze tracking in sequences A and C, respectively. Fig. 7 plots the Kalman filter states α and β for sequence A against time, as well as a scatter plot of the two, demonstrating a smooth tracking of the target's gaze orientation.

Multi-person gaze tracking: We also performed experiments on sequences with multiple individuals in the scene. Fig. 9 (top) shows results of gaze tracking from two interacting individuals. Fig. 9 (bottom) shows the tracking of three individuals.

6. Discussion

6.1. Improving primitive surveillance tasks

Gaze tracking has a lot of potential to improve primitive video surveillance tasks like person detection and tracking. *(i)* For example, the high-res face location can be used to improve person tracking accuracy. As shown in Fig. 10, when the tracker (red diamond) starts to deviate away from

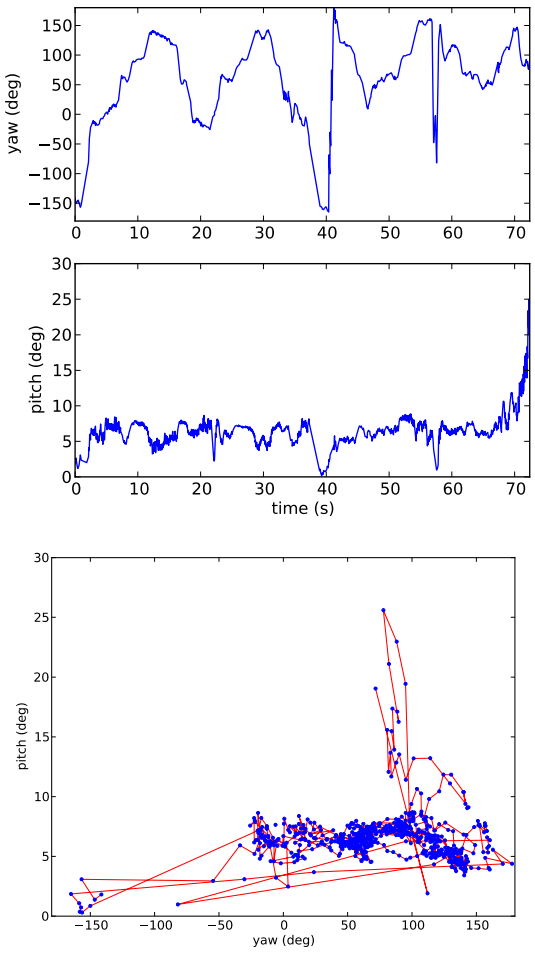

Figure 7. **Sequence A:** (Top two) Gaze trajectory in yaw (α) and pitch (β) plotted against time. (Bottom) Scatter plot of yaw vs. pitch.

the actual target location, at this point the face detection location on the head plane can be treated as a new target observation and correct the tracker location estimate. *(ii)* Similarly, the gaze can also be used to predict the future locations of the person, under the assumption that in most cases a person walks in straight direction one is looking. *(iii)* The gaze states of the trackers can also be helpful in circumventing common issues like trackers getting switched between individuals that are standing close to each other. This can be done by looking at the gaze states of both trackers and using that to resolve the ambiguities. This is especially important in any system that looks to perform group behavior analysis.

6.2. Potential Applications

Surveillance: The proposed multi-view multi-target gaze tracking system has application in behavior and social group analysis. Gaze provides valuable information to

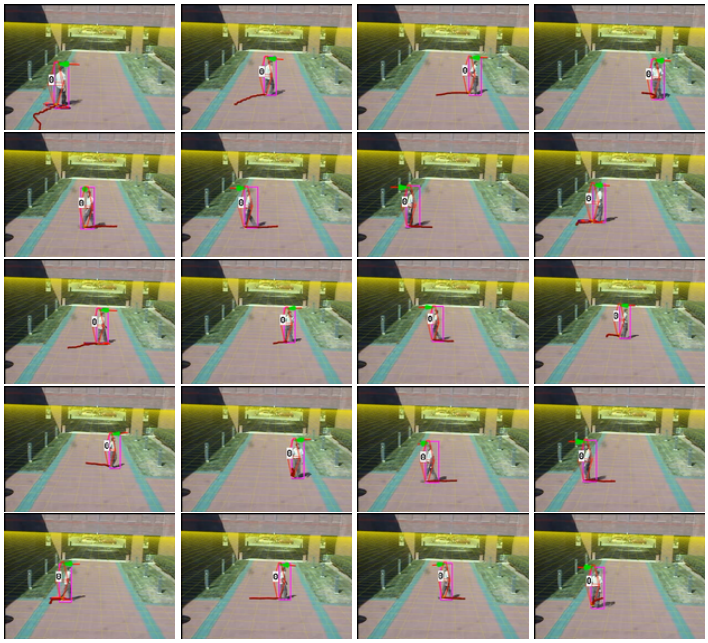

Figure 8. **Sequence C:** Few frames of simultaneous person tracking and gaze tracking at different time.

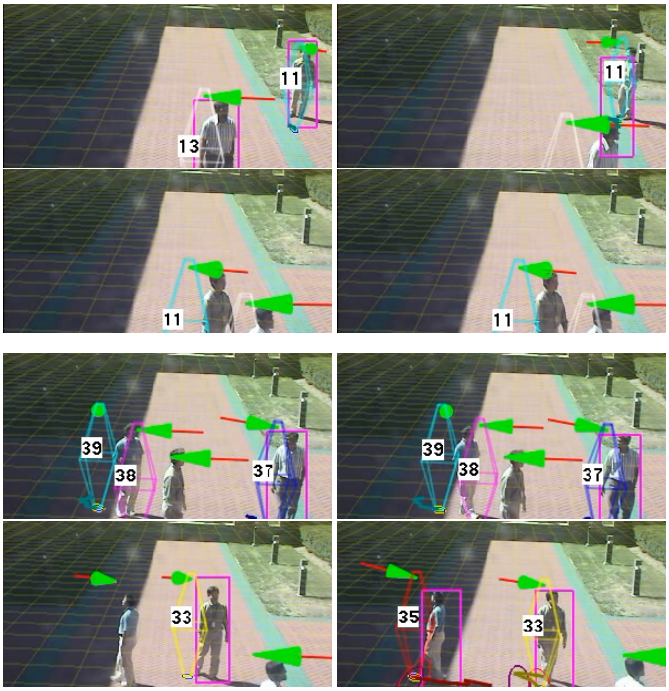

Figure 9. **Sequences D, E:** Gaze tracking with multiple individuals in the scene. Top two rows: two individuals. Bottom two rows: three individuals.

aid detecting events such as grouping formation, aggression [6]. More importantly, it may provide cues towards prediction and prevention of harmful events.

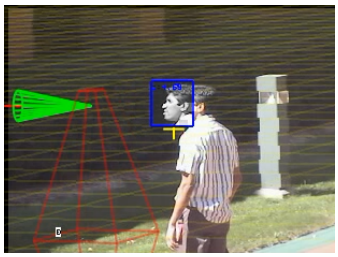

Figure 10. Example of where an individual's ground plane location estimation can be improved by using the observation from face detection.

Retail application: The proposed system has applications in retail that the gazes of customers can be studied to infer product preferences. This system could also be used to study the reaction of customers to advertisements to study attention characteristics [15].

6.3. Challenges

A major challenge with tracking gazes of multiple targets in any environment is the limitation of number of gaze observations that are obtained. This is a hardware issue and is a function of the number of cameras installed. Nowadays, with the proliferation of cameras and their reducing costs, this issue can be expected to of lesser importance if not disappear in the future. Another challenge in real-time performance is the difference in running speed of the face detection module as compared to the ground plane tracking, which could possibly result in a few seconds lag between the face detections and the tracking. We consider this to be a factor of hardware and expect this concern to be alleviated with faster systems in the future. The head pose estimation obtained in our system is relatively coarse grained and can be improved with better pose estimation algorithms (*e.g.*, [13] with a tradeoff in speed), with the requirements decided by the application.

7. Conclusions

We have presented a multi-PTZ, multi-target gaze tracking system that operates in real-time in unconstrained environments. A gaze vector is estimated for each individual based on face detection and a head-plane back-projection. A centralize Kalman filter tracking system is implemented to model the gaze tracks. A new cost metric based on location and gaze orientation is also proposed to robustly associate head observations with tracker states. Experimental results with multiple sequences demonstrate potential in surveillance, security, retail, and social group analysis applications.

Future work includes a quantitative validation of gaze tracking accuracy and reliability using standard datasets summarized in [17].

Acknowledgement. This project was supported by grants #2007-RG-CX-K015 and #2009-SQ-B9-K013 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

References

[1] Pittpatt: Pittsburgh pattern recognition. *http://www.pittpatt.com*. 3

[2] Proceedings of CLEAR'06 workshop: Classification of events, activities and relationships (http://www.clear-evaluation.org). In *Springer LNCS Series*, Southampton, UK, 2006. 2

[3] Proceedings of CLEAR'07 workshop: Classification of events, activities and relationships (http://www.clear-evaluation.org). In *Springer LNCS Series*, Washington DC, USA, 2007. 2

[4] M. Bäuml, K. Bernardin, M. Fischer, and H. K. Ekenel. Multi-pose face recognition for person retrieval in camera networks. In *Advanced Video and Signal Based Surveillance*, 2010. 2

[5] C. Canton-Ferrer, J. Casas, and M. Pardàs. Head orientation estimation using particle filtering in multiview scenarios. In *Proc. Int'l Workshop Classification of Events, Activities, and Relationships*, pages 305–310, 2007. 2

[6] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *In Proc. Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (AMMCSS)*, 2010. 1, 7

[7] P. Fihl and T. B. Moeslund. Pose estimation of interacting people using pictorial structures. In *In Proc. Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010. 1

[8] C. Hu, J. Xiao, I. Matthews, S. Baker, J. Cohn, and T. Kanade. Fitting a single active appearance model simultaneously to multiple images. In *Proceedings of the British Machine Vision Conference*, September 2004. 2

[9] K. S. Huang and M. M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *Proc. Int'l Conference Pattern Recognition*, pages 965–968, Washington, DC, USA, 2004. IEEE Computer Society. 2

[10] 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 *Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), New York*, 2006. 2

[11] N. Krahnstoever, T. Yu, S.-N. Lim, K. Patwardhan, and P. Tu. Collaborative real-time control of active cameras in large scale surveillance systems. In *Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, October 2008. 2, 3

[12] O. Lanz and R. Brunelli. Joint bayesian tracking of head location and pose from low-resolution video. In *Advanced Video and Signal Based Surveillance*, 2010. 2

[13] X. Liu. Discriminative face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(11):1941–1954, November 2009. 3, 7

[14] X. Liu. Video-based face model fitting using adaptive active appearance model. *Image and Vision Computing*, 28(7):1162–1172, July 2010. 3

[15] X. Liu, N. Krahnstoever, T. Yu, and P. Tu. What are customers looking at? In *Proc. IEEE International Conference On Advanced Video and Signal Based Surveillance*, 2007. 1, 7

[16] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of SIAM*, 5:32–38, 1957. 3, 5

[17] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, 2009. 2, 3, 7

[18] R. Ptucha and A. Savakis. Facial pose estimation using a symmetrical feature model. In *Proc.Int'l Conference on Multimedia and Expo.*, pages 1664–1667, 2009. 2

[19] H. Schneiderman. Learning a restricted Bayesian network for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 639–646, 2004. 3

[20] M. Voit, K. Nickel, and R. Stiefelhagen. Head pose estimation in single- and multi-view environments – results on the clear'07 benchmarks. In *Proc. Int'l Workshop Classification of Events, Activities, and Relationships*, 2007. 2

[21] T. Yu, S. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 2

[22] T. Yu, Y. Wu, N. O. Krahnstoever, and P. H. Tu. Distributed data association and filtering for multiple target tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, June 2008. 2