Gaze and Body Pose Estimation from a Distance

Nils Krahnstoever Ming-Ching Chang Weina Ge
GE Global Research Center, One Research Circle, Niskayuna, NY, USA
nils@krahnstoever.com {changm,gewe}@research.ge.com

Abstract

We present a comprehensive approach to track gaze by estimating location, body pose, and head pose direction of multiple individuals in unconstrained environments. The approach combines person detections from fixed cameras with directional face detections obtained from actively controlled pan tilt zoom (PTZ) cameras. The main contribution of this work is to estimate both body pose and head pose (gaze) direction independently from motion direction, using a combination of sequential Monte Carlo Filtering and MCMC sampling. There are numerous benefits in tracking body pose and gaze in surveillance. It allows to track people's focus of attention, can optimize the control of active cameras for biometric face capture, and can provide better interaction metrics between pairs of people. The availability of gaze and face detection information also improves localization and data association for tracking in crowded environments. The performance of the system will be demonstrated on data captured at a real-time surveillance site.

1. Introduction

Detecting and tracking individuals under unconstrained conditions such as in mass transit stations, sport venues, and schoolyards are important. On top of that, the understanding of their gaze and intention are more challenging due to the general freedom of movements and frequent occlusions. Moreover, face images in standard surveillance videos are usually low-resolution, which limits the detection rate. Unlike previous approaches [15, 17, 12] that at most obtained gaze information, we use multi-view pan tilt zoom (PTZ) cameras and tackle the problem of joint, holistic tracking of both body pose and head orientation in real-time. Following Stiefelhagen *et al.* [16], we assume that the gaze can be reasonably derived by head pose in most cases [15]. Throughtout the paper we refer to head pose as gaze or visual focus of attention and use them interchangably. The coupled person tracker, pose tracker, and gaze tracker are integrated and synchronized, thus robust tracking via mutual update and feedback is possible. The capability to reason over gaze angle provides a strong indication of attention, which benifits a surveillance system on many fronts. In particular as part of interaction models in event recognition, it is important to know if a group of individuals are facing each other (*e.g.*, talking), facing a common direction (*e.g.*, looking at another group before a conflict is about to happen) or facing away from each other (*e.g.*, because they are not related or because they are in a "defense" formation).

Our contribution is three-fold. *(1)* We propose a unified framework to couple multi-view person tracking with asynchronous PTZ gaze tracking to jointly and robustly estimate pose and gaze. The novelty over [13] is a coupled particle filtering tracker that jointly estimates body pose and gaze. While we use person tracking to control PTZ cameras, which allow us to perform face detection and gaze estimation, we in turn utilize the resulting face detection locations to further improve tracking performance. *(2)* We can thus actively leverage track information to control PTZ cameras in maximizing the probability of capturing frontal facial views. Our work significantly improves previous efforts [7] that used the walking direction of individuals as an indication of gaze direction, which breaks down in situations where people are stationary. *(3)* Our framework is general and applicable to many other vision-based applications. For example, we allow optimal face capture for biometrics particularly in environments where people are stationary, because it obtains gaze information directly from face detections.

We use a network of fixed cameras to perform site-wide person tracking. This person tracker drives one or more PTZ cameras to target individuals (details in §3) to obtain close-up views. A centralized tracker operates on the ground plane to fuse together information from person tracks and face tracks. Due to the large computational burden on inferring gaze from face detections, the person tracker and face tracker must operate *asynchronously* to run in real-time. Our system can operate on either a single or multiple cameras. The multi-camera setting improves overall tracking performance in crowded conditions. Gaze tracking in this case is also useful in performing high-level reasoning *e.g.*, to analyze social interactions [18], attention model, and behaviors [3].

2. Related Work

To the best of our knowledge, we are the first to proactively integrate multi-camera with multi-PTZ pose estimation for gaze tracking in unconstrained environments. Our work involves multi-view person tracking and head pose tracking across one or more views. In practice the face resolution is typically low, so one must either rely on special methods [4, 12] or use PTZ camera [13] to obtain close-up shots (as we did). The method of Robertson *et al.* [12] is data-driven based on Bayesian fusion of priors, thus relies on training videos. Their coupling of face angle and head tracking workes in a limited single field of view. Hoedl *et al.* [5] use a two-camera system (one fixed and one PTZ) to perform pedestrian detection, however no face or gaze analysis is performed.

Head pose and gaze tracking has been studied extensively [10]. However, most existing systems restrict to an indoor room setting and it is often assumed that subjects stay seated (therefore tracking is trivial and no camera control is involved).

The idea of joint person and face tracking is not new, however, existing works do not attempt to fuse both trackers in using one to update the other. Ozturk *et al.* [11] track body and head pose using independent trackers on a single top-view camera. Bäuml *et al.* [1] assume head location is known and track faces across a distributed camera network for recognition and re-identification. Their face tracker and person tracker runs separately, where the overlapping facial views are processed independently. Smith *et al.* [15] estimate multiple individuals' body pose/gaze to track their visual focus of attention. Their work is relevant to us, except that they use a single camera view, while we fuse the tracking across multiple views.

3. Video Tracking and PTZ Control

Our video tracking system is based on [6] consisting of 4 fixed and 4 PTZ cameras. The fixed camera views are utilized by a centralized tracker to estimate the 3D locations of individuals in a common ground plane. The person tracking are then used online to drive the PTZ cameras to capture zoom-in face images. Recognized faces are then associated with person trackers to cooperatively improve the overall tracking.

Our PTZ camera control strategy aims at capturing frontal face views from individuals, even including uncooperative ones at a distance. The system must then determine how (what camera to drive, and where to move) to obtain the best shots automatically. Our control algorithm pursues the goal of scheduling PTZ cameras in in a way optimizing frontal face capture. The control system provides each PTZ camera with a continuously evolving *schedule* [7] 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 continuously capturing high quality facial shots of all tracked individuals. The quality of facial shots is governed by several factors: 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 tracker. A control strategy is chosen by selecting the schedule with the highest probability from the set of all possible schedules. Our PTZ control is pursuing a chicken and egg problem. It leverages gaze information to schedule the PTZ cameras, but (at least for stationary individuals) no gaze information will be acquired until close-up PTZ views have been obtained. The controller hence builds the uncertainty around the gaze estimates into the control strategy.

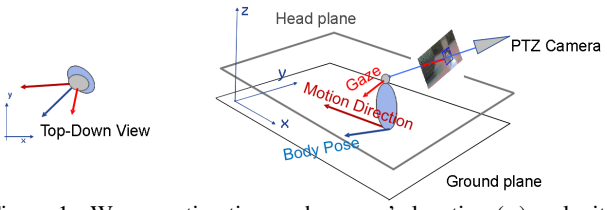

Figure 1. We are estimating each person's location ($\mathbf{x}$), velocity ($\mathbf{v}$), body pose ($\boldsymbol{\alpha}$) and gaze direction ($\boldsymbol{\phi}, \boldsymbol{\theta}$) in world coordinates.

4. Person, Body Pose and Gaze Tracking

4.1. Problem Definition

We represent each individual with a state vector $\mathbf{s} = [\mathbf{x}, \mathbf{v}, \alpha, \phi, \theta]$, where $\mathbf{x}$ is the location on the (X,Y) groundplane metric world, $\mathbf{v}$ is the velocity on the groundplane, α is the horizontal orientation of the body around the groundplane normal, ϕ is the horizontal gaze angle and θ is the vertical gaze angle (positive above the horizon and negative below it), see Fig.1. There are two types of observations in our system: person detections $(\mathbf{z}, \mathbf{R})$, where $\mathbf{z}$ is a groundplane location measurement and $\mathbf{R}$ the uncertainty of this measurement and face detections $(\mathbf{z}, \mathbf{R}, \gamma, \rho)$ where the additional parameters γ and ρ are the horizontal and vertical gaze angles. Each person's head and foot locations are extracted from image-based person detections [6] and back-projected onto the world head- and ground-plane respectively, using an unscented transform (UT). Next, face positions and poses in PTZ views are obtained using the PittPatt face detector [14]. Their metric world groundplane locations are again obtained through back-projection. Face pose is obtained by matching face features. Individual's gaze angles are obtained by mapping face pan and rotation angles in image space into the world space. Finally, the world gaze angles are obtained by mapping the image local face normal $\mathbf{n}_{\text{img}}$ into world coordinates via $\mathbf{n}_{\text{w}} = \mathbf{n}_{\text{img}}\mathbf{R}^{-T}$, where $\mathbf{R}$ is the rotation matrix of the projection $\mathbf{P} = [\mathbf{R}|\mathbf{t}]$. Observation gaze angles (γ, ρ) are obtained directly from this normal vector. Width and height of the face are used to estimate a covariance confidence level for the face location. The covariance is projected from the image to the groundplane again using the UT from the image to the head plane,

followed by down projection to the groundplane. Fig.3 depicts detected faces and gaze angles in estimation in multiple views.

In contrast to [13], where a person's gaze angle was estimated independently from location and velocity, and body pose was ignored, this work aims at correctly modeling the relationship between motion direction, body pose, and gaze. First, in this work body pose is not strictly tied to motion direction. People can move backwards and sideways especially when people are waiting or standing in groups (*albeit*, with increasing velocity sideways people's motion becomes improbable, and at even greater velocities, only forward motion is assumed). Secondly, head pose is not tied to motion direction, but there are relatively strict limits on what pose the head can assume relative to body pose. Under this model the estimation of body pose is not trivial as it is only *loosely coupled* to gaze angle and velocity (which in turn is only observed indirectly). We perform the entire state estimation using a Sequential Monte Carlo filter, described in the next section.

4.2. Estimation of Location, Pose and Gaze

We assume for now that we have a method for associating measurements with tracks over time. For the sequential Monte Carlo filter, we need to specify *(i)* the dynamical model and *(ii)* the observation model of our system.

Dynamical Model: Following the description in the previous section, our state vector is $\mathbf{s} = [\mathbf{x}, \mathbf{v}, \alpha, \phi, \theta]$ and the state prediction model decomposes as follows:

$$\begin{aligned} p(\mathbf{s}_{t+1}|\mathbf{s}_t) = p(\mathbf{q}_{t+1}|\mathbf{q}_t)p(\alpha_{t+1}|\mathbf{v}_{t+1}, \alpha_t) \\ p(\phi_{t+1}|\phi_t, \alpha_{t+1})p(\theta_{t+1}|\theta_t), \end{aligned} \tag{1}$$

where we used the abbreviation $\mathbf{q} = (\mathbf{x}, \mathbf{v}) = (x, y, v_x, v_y)$. For the location and velocity we assume a standard linear dynamical model

$$p(\mathbf{q}_{t+1}|\mathbf{q}_t) = \mathcal{N}(\mathbf{q}_{t+1} - \mathbf{F}_t\mathbf{q}_t, \mathbf{Q}_t), \tag{2}$$

where $\mathcal{N}$ denotes Normal distribution, $\mathbf{F}_t$ is a standard constant velocity state predictor corresponding to $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{v}_t\Delta t$ and $\mathbf{Q}_t$ the standard system dynamics [2]. The second term in Eq.(1) describes the propagation of the body pose under consideration of the current velocity vector. We assume the following model

$$p(\alpha_{t+1}|\mathbf{v}_{t+1}\alpha_t) = \mathcal{N}(\alpha_{t+1} - \alpha_t, \sigma_\alpha)\cdot \tag{3}$$
$$\begin{cases} (1.0 - P^o)\mathcal{N}(\alpha_{t+1} - \nu_{t+1}, \sigma_{\nu\alpha}) + P^o\frac{1}{2\pi} & \text{if } ||\mathbf{v}|| > 2 \text{ m/s}, \\ \frac{1}{2\pi} & \text{or if } ||\mathbf{v}|| < \frac{1}{2} \text{ m/s}, \\ P^f\mathcal{N}(\alpha_{t+1} - \nu_{t+1}, \sigma_{\nu\alpha}) + & \\ \quad P^b\mathcal{N}(\alpha_{t+1} - \nu_{t+1} - \pi, \sigma_{\nu\alpha}) + P^o\frac{1}{2\pi} & \text{otherwise}, \end{cases}$$

where $P^f = 0.8$ is the probability (for medium velocities $\frac{1}{2}$ m/s $< \mathbf{v} <$ 2 m/s) of a person walking forwards, $P^b = 0.15$ the probability (for medium velocities) of walking backwards, and $P^o = 0.05$ the background probability allowing arbitrary pose to movement direction relationships, based on experimental heuristics. With ν_{t+1} we denote the direction of the velocity vector $\mathbf{v}_{t+1}$ and with $\sigma_{\nu\alpha}$ the expected distribution of deviations between movement vector and body pose. The front term $N(\alpha_{t+1} - \alpha_t, \sigma_\alpha)$ represents the system noise component, which in turn limits the change in body pose over time. All changes in pose are attributed to deviations from the constant pose model.

The third term in Eq.(1) describes the propagation of the horizontal gaze angle under consideration of the current body pose. We assume the following model

$$\begin{aligned} p(\phi_{t+1}|\phi_t\alpha_{t+1}) = \mathcal{N}(\phi_{t+1} - \phi_t, \sigma_\phi)\cdot \\ \left\{P_g^u\Theta(|\phi_{t+1} - \frac{\pi}{3}|) + P_g\mathcal{N}(\phi_{t+1} - \alpha_{t+1}, \sigma_{\alpha\phi})\right\}, \end{aligned} \tag{4}$$

where the two terms weighted by $P_g^u = 0.4$ and $P_g = 0.6$ define a distribution of the gaze angle (ϕ_{t+1}) with respect to body pose (α_{t+1}) that allows arbitrary values within a range of $\alpha_{t+1} \pm \frac{\pi}{3}$ but favors distribution around body pose.

Finally the fourth term in Eq.(1) describes the propagation of the tilt angle, $p(\theta_{t+1}|\theta_t) = \mathcal{N}(\theta_{t+1}, \sigma_\theta^0)\mathcal{N}(\theta_{t+1} - \theta_t, \sigma_\theta)$, where the first term models that a person tends to favor horizontal directions and the second term represents system noise. Noted that in all above equations, care has to be taken with regard to angular differences.

To propagate the particles forward in time, we need to sample from the state transition density Eq.(1), given a previous set of weighted samples $(\mathbf{s}_t^i, \mathbf{w}_t^i)$. While for the location, velocity and vertical head pose, this is easy to do. The loose coupling between velocity, body pose and horizontal head pose is represented by a non-trivial set of transition densities Eq.(3) and Eq.(4). To generate samples from these transition densities we perform two Markov Chain Monte Carlo (MCMC). Exemplified on Eq.(3), we use a Metropolis sampler [8] to obtain a new sample as follows:

- **Start:** Set $\alpha_{t+1}^i[0]$ to be the α_t^i of particle i.
- **Proposal Step:** Propose a new sample $\alpha_{t+1}^i[k+1]$ by sampling from a *jump-distribution* $G(\alpha|\alpha_{t+1}^i[k])$.
- **Acceptance Step:** Set $r = p(\alpha_{t+1}^i[k+1]|\mathbf{v}_{t+1}\alpha_t^i)/p(\alpha_{t+1}^i[k]|\mathbf{v}_{t+1}\alpha_t^i)$. If $r \geq 1$, accept the new sample. Otherwise accept it with probability r. If it is not accepted, set $\alpha_{t+1}^i[k+1] = \alpha_{t+1}^i[k]$.
- **Repeat:** Until $k = N$ steps have been completed.

Typically only a small fixed number of steps ($N = 20$) are performed. The above sampling is repeated for the horizontal head angle in Eq.(4). In both cases the jump distribution is set equal to the system noise distribution, except with a fraction of the variance *i.e.*, $G(\alpha|\alpha_{t+1}^i[k]) =$

$\mathcal{N}(\alpha - \alpha^i_{t+1}[k], \sigma_\alpha/3)$ for body pose; $G(\phi|\phi^i_{t+1}[k])$ and $G(\theta|\theta^i_{t+1}[k])$ are defined similarly. The above MCMC sampling ensures that only particles that adhere both to the expected system noise distribution as well to the loose relative pose constraints are generated. We found 1000 particles are sufficient.

Observation Model: After sampling the particle distribution $(\mathbf{s}^i_t, \mathbf{w}^i_t)$ according to its weights $\{\mathbf{w}^i_t\}$ and forward propagation in time (using MCMC as described above), we obtain a set of new samples $\{\mathbf{s}^i_{t+1}\}$. The samples are weighted according to the observation likelihood models described next.

For the case of person detections, the observations are represented by $(\mathbf{z}_{t+1}, \mathbf{R}_{t+1})$ and the likelihood model is:

$$p(\mathbf{z}_{t+1}|\mathbf{s}_{t+1}) = \mathcal{N}(\mathbf{z}_{t+1} - \mathbf{x}_{t+1}|\mathbf{R}_{t+1}). \tag{5}$$

For the case of face detection $(\mathbf{z}_{t+1}, \mathbf{R}_{t+1}, \gamma_{t+1}, \rho_{t+1})$, the observation likelihood model is

$$\begin{aligned} p(\mathbf{z}_{t+1}, \gamma_{t+1}, \rho_{t+1}|\mathbf{s}_{t+1}) = \mathcal{N}(\mathbf{z}_{t+1} - \mathbf{x}_{t+1}|\mathbf{R}_{t+1}) \\ \mathcal{N}(\lambda((\gamma_{t+1}, \rho_{t+1}), (\phi_{t+1}, \theta_{t+1})), \sigma_\lambda), \end{aligned} \tag{6}$$

where $\lambda(.)$ is the geodesic distance (expressed in angles) between the points on the unit circle represented by the gaze vector $(\phi_{t+1}, \theta_{t+1})$ and the observed face direction $(\gamma_{t+1}, \rho_{t+1})$ respectively.

$$\begin{aligned} \lambda((\gamma_{t+1}, \rho_{t+1}), (\phi_{t+1}, \theta_{t+1})) = \arccos(\sin\rho_{t+1}\sin\theta_{t+1} \\ + \cos\rho_{t+1}\cos\theta_{t+1}\cos(\gamma_{t+1} - \phi_{t+1})). \end{aligned}$$

The value σ_λ is the uncertainty that is attributed to the face direction measurement. Overall the tracking state update process works as summarized in Algorithm 1.

4.3. Data Association

So far we assumed that observations had already been assigned to tracks. In this section we will elaborate how observation to track assignment is performed. To enable the tracking of multiple people, observations have to be assigned to tracks over time. In our system, observations arise *asynchronously* from multiple camera views. The observations are projected into the common world reference frame, under consideration of the (possibly time varying) projection matrices, and are consumed by a centralized tracker in the order that the observations have been acquired. For each time step, a set of (either person or face) detections $\mathbf{Z}^l_t$ have to be assigned to tracks $\mathbf{s}^k_t$. We construct a distance measure $C_{kl} = d(\mathbf{s}^k_t, \mathbf{Z}^j_t)$ to determine the optimal one-to-one assignment of observations l to tracks k using Munkres algorithm [9]. Observations that do not get assigned to tracks might be confirmed as new targets and are used to spawn new candidate tracks. Tracks that do not get detections assigned to them are propagated forward in time and thus do not undergo weight update.

Data : Sample set $S_t = (w^i_t, \mathbf{s}^i_t)$
Result : Sample set $S_{t+1} = (w^i_{t+1}, \mathbf{s}^i_{t+1})$
begin
  for $i = 1, \ldots, M$ *(number of particles)* **do**
    Randomly select sample $\mathbf{s}^i_t = (\mathbf{x}^i_t, \mathbf{v}^i_t, \alpha^i_t, \phi^i_t, \theta^i_t)$ from S_t according to weights w^i_t
    Obtain forward propagated locations $\mathbf{x}^i_{t+1}$ and $\mathbf{v}^i_{t+1}$ by sampling from distribution Eq.(2).
    Perform MCMC to sample a new body pose α^i_{t+1} from Eq.(3).
    Perform MCMC to sample a new horisontal gaze vector ϕ^i_{t+1} from Eq.(4).
    Sample new vertical face angle θ^i_{t+1} from distribution $p(\theta_{t+1}|\theta_t)$.
    Evaluate new state ${w^i_{t+1}}' = p(\mathbf{z}_{t+1}|\mathbf{s}^i_{t+1})$ with Eq.(5) if the observation is a person detection, or Eq.(6) if it is a directional face detection. Renormalize particle set to obtain final update distribution $S_{t+1} = (w^i_{t+1}, \mathbf{s}^i_{t+1})$.
end

Algorithm 1: Location, pose and gaze angle tracking.

The use of face detections lead to an additional source of location information. We explicitly use this to improve tracking. Results show that this is particularly useful in crowded environments, where face detectors are less susceptible to person-person occlusion. Our other advantage is that the gaze information introduces an additional component into the detection-to-track assignment distance measure, which works effectively to assign oriented faces to person tracks.

For person detections, the metric is computed from the target gate as follows:

$$\boldsymbol{\mu}^k_t = \frac{1}{N}\sum_i \mathbf{x}^{ki}_t, \; \boldsymbol{\Sigma}^{kl}_t = \frac{1}{N-1}\sum_i (\mathbf{x}^{ki}_t - \boldsymbol{\mu}^k_t)(\mathbf{x}^{ki}_t - \boldsymbol{\mu}^k_t)^T + \mathbf{R}^l_t,$$

where $\mathbf{R}^l_t$ is the location covariance of observation l and $\mathbf{x}^{ki}_t$ is the location of the i^{th} particle of track k at time t. The distance measure is then given as:

$$C^l_{kl} = (\boldsymbol{\mu}^k_t - \mathbf{z}^l_t)^T(\boldsymbol{\Sigma}^{kl}_t)^{-1}(\boldsymbol{\mu}^k_t - \mathbf{z}^l_t) + \log|\boldsymbol{\Sigma}^{kl}_t|$$

For face detections, the above is augmented by an additional term for the angle distance:

$$C_{kl} = C^l_{kl} + \frac{\lambda((\gamma^l_t, \rho^l_t), (\mu^k_{\phi_t}, \mu^k_{\theta_t}))^2}{\sigma^2_\lambda} + \log\sigma^2_\lambda,$$

where the $\mu^k_{\phi_t}$ and $\mu^k_{\theta_t}$ are computed from the first order spherical moment of all particle gaze angles (angular mean); σ_λ is the standard deviation from this moment; (γ^l_t, ρ^l_t) are the horizontal and vertical gaze observation angles in observation l.

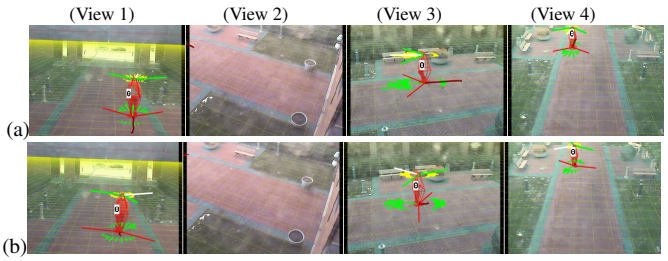

Figure 2. Pose and gaze estimation in absence of face detections: (a) forward motion, (b) backward motion. See text for explanation.

Since only PTZ cameras provide face detections and only fixed cameras provide person detections, data association is performed with either all person detections or all face detections; the gaze of mixed associations does not arise.

5. Experiments and Results

We first demonstrate the concept of our system with a single person, without the use of face detection information. In this mode the estimation reduces to the prior of the body pose and gaze, conditioned on the motion vector of the person. As one can see in Fig.2, for moderate walking speeds the system correctly estimates the direction to be either forward or backward facing, relative to the motion vector (representing the two possibilities of the person walking forward or backward), and the gaze vector is estimated to fall within a range of these forward and backward directions. Figs. 2 to 5 visualize the gaze and body pose estimates via circular histograms around each person, where the fan of circular bars radiate away from each target. The direction of the bars indicate different pose and gaze angles; the length of the bars indicates the probability associated with this direction. In Fig.2 the yellow histogram around the head represent gaze direction and the green histogram around the feet represent bode pose direction. The two green lines radiating from the head are the one sigma standard deviations from the average gaze direction and the red lines radiating from the feet are the one sigma deviations of the body pose. Also shown are the trajectory in the ground plane (dark red). Fig.2(b) shows how for slow backward motion the ambiguity between forward or backward pose increases, represented by the two almost equally distributed direction possibilities.

We repeated the single person tracking experiment with the face detector enabled in Fig.3. We can clearly see how the introduction of gaze measurements have significantly improved the accuracy of the gaze and pose estimates. We also see how in particular in the case of backward motion, the correct body pose and gaze angles have been maintained. Fig.3(c) also shows how the system can correctly track sideway glances where the motion and gaze angles differ significantly during tracking.

We next demonstrate the system with multiple individuals. Fig.4 shows several examples of two people standing

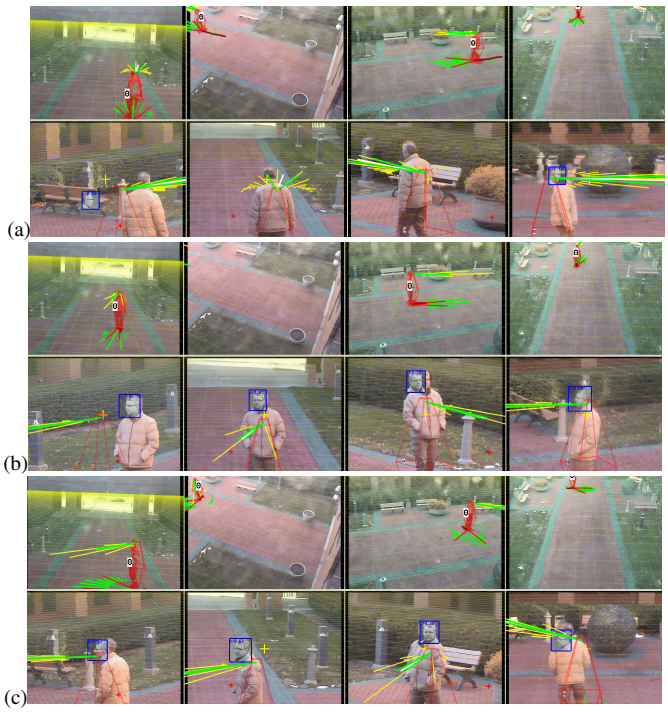

Figure 3. Pose and gaze estimation with face detections from PTZ cameras: (a) forward and (b) backward motion, (c) sideways glance with large difference between body pose and gaze. Overall tracking improves significantly when compared to Fig.2.

Table 1. Average angle difference in degrees for our method, the gaze tracker, and the baseline tracker.

	Our method		Gaze tracker [13]		Baseline	
	pose	gaze	pose	gaze	pose	gaze
one person	**19.47**	**23.82**	57.58	33.01	39.84	33.10
two people	**57.12**	**35.65**	73.40	45.33	73.40	88.32
three people	**42.40**	38.30	73.55	**37.31**	73.55	90.67

closely, where the system correctly estimates the body pose and gaze angles for different situations. Fig.5 shows an experiment with three people moving and turning freely.

We performed a quantatively evaluation based on ground truth on all three sequences in Table 1. Pose and gaze groundtruth was annotated in 3D using a customized user interface. We compare our method against two methods: *(1)* a baseline method which uses a simple groundplane person traker similar to Fig.2, and *(2)* a gaze tracking method reported in [13] which uses a separate person and gaze tracker. For the baseline method, gaze and pose was assumed to be equivalent to motion direction. The method in [13] provided gaze information but not the body pose, so we again equate the pose to motion direction. In five out of six cases, our method has the smallest estimation error. The advantage of correctly modeling the relationship between motion direction, body pose, and gaze is clear.

The above experiments demonstrate the efficacy of estimating gaze and body pose for a group of people in close proximity. We foresee that the system should extend well

Figure 4. Pose and gaze estimation of two people (a) facing each other, (b) standing next to each other, and (c) facing in opposite directions. The top second view in each case depicts an artificial top-down re-rendering, which provides a better visualization of the relative pose and gaze.

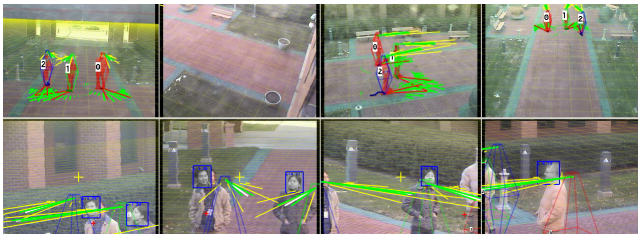

Figure 5. Pose and gaze estimation for three people moving and turning freely.

to work in a crowded condition, provided that a sufficient number of PTZ views are added.

6. Discussion and Conclusions

We have presented a comprehensive system for tracking location, body pose and gaze direction in unconstrained environment using surveillance and PTZ cameras. We have shown through qualitative experiments that the system performs well under a variety of experimental conditions. The algorithm has important applications in security, surveillance, and behavior recognition, as well as the emerging areas of gaming, interactive entertainment and advertising.

Acknowledgement. This work was supported by grant #2009-SQ-B9-K013 awarded by the National Institute 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] M. Bäuml, K. Bernardin, M. Fischer, and H. K. Ekenel. Multi-pose face recognition for person retrieval in camera networks. In *AVSS*, 2010. 2

[2] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, 1999. 3

[3] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *AMMCSS*, pages 56–63, 2010. 1

[4] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley. Head pose estimation on low resolution images. In *R. Stiefelhagen and J.S. Garofolo (eds.) CLEAR 2006 LNCS*, volume 4122, pages 270–280. Springer-Verlag, 2007. 2

[5] T. Hoedl, D. Br, U. Soergel, and M. Wiggenhagen. Real-time orientation of a PTZ-camera based on pedestrian detection in video data of wide and complex scenes. In *ISPRS*, pages 663–668, 2008. 2

[6] 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*, pages 67–74, 2006. 2

[7] 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 *ECCV M2SFA2*, 2008. 1, 2

[8] N. Metropolis, A. W. Rosenbluth, N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *JCP*, 21:1087–1092, 1953. 3

[9] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of SIAM*, 5:32–38, 1957. 4

[10] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: a survey. *PAMI*, 31(4):607–626, 2009. 2

[11] P. Ozturk, T. Yamasaki, and K. Aizawa. Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus and attention analysis. In *ICCV Computer Vision Workshop*, pages 1020–1027, 2009. 2

[12] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV LNCS*, volume 3952, pages 402–415, 2006. 1, 2

[13] K. Sankaranarayana, M.-C. Chang, and N. Krahnstoever. Tracking gaze direction from far-field surveillance cameras. In *WACV*, pages 519–526, 2011. 1, 2, 3, 5

[14] H. Schneiderman. Learning a restricted Bayesian network for object detection. In *CVPR*, pages 639–646, 2004. 2

[15] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *PAMI*, 30:1212–1229, July 2008. 1, 2

[16] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *Visual information and information systems*, pages 761–768, 1999. 1

[17] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *ETRA*, pages 245–250, 2008. 1

[18] T. Yu, S. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, pages 1462–1469, 2009. 1