# Tracking with Intent

Rolf H. Baxter
Heriot-Watt University, Edinburgh, UK
Email: R.H.Baxter@hw.ac.uk

Michael Leach
Roke Manor Research, UK

Neil M. Robertson
Heriot-Watt University, Edinburgh, UK

***Abstract*—This paper presents the novel theory for performing behaviour-based tracking using intentional priors. Motivated by our ultimate goal of anomaly detection, our approach is rooted in building better models of target behaviour. Our novel extension of the Kalman filter combines motion information with an intentional prior. We apply our 'Intentional Tracker' to a pedestrian surveillance and tracking problem, using head pose as the intentional prior. We perform a statistical analysis of pedestrian head pose behaviour and demonstrate tracking performance on a set of simulated and real pedestrian observations. We show that by using intentional priors our algorithm outperform a standard Kalman filter across a range of target trajectories.**

## I. Introduction

The grand aim of our work is statistical anomaly detection and good models of 'normal' behaviour are needed to detect subtle anomalies. In this paper we introduce *intentional priors* into a Kalman Filter (KF) to better predict pedestrian behaviour, mediated by gazing patterns. This theory generalises: consider a car approaching a crossroads and the indicator light signals intention to turn. Contextual knowledge enables better predictions. Recent advances in head-pose detection at-a-distance make this possible, i.e. the Benfold model [1] allows heads to be detected and tracked and head-pose to be identified (see Fig. 1). This information can be associated with trajectories of positions and we propose that head pose information should be used as part of person tracking algorithms. The solid black line indicates the person's motion track while the grey sections indicate where the person is looking (see inset). The person moves from top right to bottom left, changing their trajectory at coordinates (-2, 2). In the presence of occlusions a tracker using a constant velocity model could only predict that the target would continue along the same trajectory, while the intuition is that head pose information from before the occlusion could be a more informative prior than previous velocity.

We verify the hypothesis that people do indeed tend to look where they are going, and show that it can be used to make better behaviour predictions. We test our model within a tracking problem where behaviour predictions can be analysed in a well understood and principled way. Specifically, we compare the tracking error between competing models. The contributions of this work are: (a) we show that head-pose is well correlated with a person's direction of travel using 3 benchmarked video datasets; (b) we present a novel method for integrating intentional priors into the KF to improve tracking, and; (c) we validate our approach within a person tracking application, using head pose as the intentional prior, comparing to the KF.

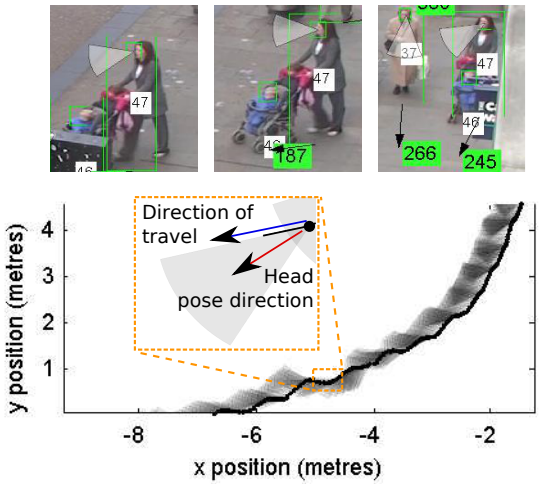


Fig. 1. (*Top*) Sample frames from the Benfold dataset [1] showing the head detections & pose of a pedestrian. (*Bottom*) The extracted ground truth trajectory and head pose behaviour of the person over time.

**Behaviour based tracking.** Target behaviour models may be adjusted and/or switched according to the behaviour of the target. Pelligrini assumes that socially connected pedestrians will have similar trajectories while accounting for social norms regarding collision avoidance and 'personal space' [2]. Their tracker is based on Conditional Random Fields and changes predictions about target movement according to the likelihood that the targets are socially connected. We do not consider social connections when making predictions but in contrast use head pose alone to predict target trajectory.

Sankaranarayanan fuses person tracking information with a Pan-Tilt-Zoom (PTZ) facial tracking system, but do not use head pose for predicting future target location [3]. They focus on active sensor management to obtain close-up images based on predicted location, and tracking the head pose and person position without using head pose as a predictor. Tordoff and Murray use a KF for target tracking in video, but their focus is to automatically control the camera's zoom according to the tracking error [4]. We also use an error signal to adjust the tracker, but that error signal relates to the deviation between the head pose direction and estimated direction of travel.

**Head pose estimation** The objective is to estimate the direction in which a human head is posed. Robertson and

Reid were the first to show that head pose can facilitate behaviour explanation in low/medium resolution images [5]. Benfold and Reid also report good recognition accuracy with their model ($24^\circ$ degrees error). We use their published dataset and extend their analysis of gaze behaviour [1]. Mukherjee and Robertson showed that head pose can be extracted in real time from RGB and depth data [6], and report high degrees of recognition accuracy with their novel Histogram of Azimuth Oriented Depth Normals, even when down sampling the data significantly. Chen and Odobez used surveillance video data and a coupled body and head detection algorithm to account for the fact that head and body location are constrained [7].

We build on this previous work in head pose estimation, and novelly exploit it as an *intentional prior to improve tracking*.

## II. Using head pose to predict behaviour.

Specifically, we consider the application of pedestrian surveillance and tracking. The intuition is that people tend to look where they are going which makes head pose an informative intentional prior for pedestrian targets. Within any tracking paradigm knowing a target's destination can be useful for dealing with occlusions and intermittent detections. This is particularly true of complex targets with irregular movement where past trajectory is not necessarily a good indicator of future location.

To test the intuition that head-pose can be used as an intentional prior we propose and validate the following hypothesis:

***Hypothesis 1:*** *Pedestrian head pose is well correlated with direction of travel.*

We also propose that the head pose behaviour of socially connected persons differs from those that are unconnected, a hypothesis based on the intuition that socially connected persons tend to interact with each other (e.g. looking at each other while talking). As already stated, our approach to anomaly detection it to build better models of normal behaviour, and thus social context is equally as important as more traditional context such as spatial and temporal (e.g. [8]).

***Hypothesis 2:*** *The head-pose behaviour of socially connected pedestrians (to others in the scene) differs from those without social connections.*

### A. Validation of the Hypotheses

We performed a statistical analysis of pedestrian trajectory and head pose behaviour to validate our hypotheses. This analysis was performed on three benchmark video datasets: Benfold [1], Caviar [9] and PETS 2007 [10] using manual annotations of person location (bounding box), head location (bounding box) and head pose direction (angle). The analysis was performed in two steps.

In the first step, pedestrians were manually segmented into two groups: those with and without social connections. In the absence of social ground truth we used visible interactions, proximity and shared trajectory as methods for identifying social connections. Perhaps surprisingly, a large number of high-confidence social connections could be identified in this way. Where high-confidence could not be achieved, no social connection was assumed to exist.

TABLE I
Mean ($\mu$) and standard deviation ($\sigma$) statistics extracted from the Benfold [1], Caviar [9] and PETS 2007 [10] datasets for socially connected (C) and unconnected (U) persons

| **Dataset** | $\mu$ **(C)** | $\sigma$ **(C)** | $\mu$ **(U)** | $\sigma$ **(U)** |
|---|---|---|---|---|
| Benfold | 1.750 | 36.351 | 4.541 | 40.582 |
| Caviar | 1.220 | 60.533 | -10.321 | 40.068 |
| PETS 2007 | 15.233 | 86.800 | 23.922 | 57.550 |

Once segmented, we calculated the deviation between direction of travel and head pose for each pedestrian in each frame of video. For the Caviar and PETS datasets travel direction was calculated using the bounding boxes for each pedestrian to approximate the location of their feet. These locations were projected to the ground plane using a homography transform from which trajectories could be derived for each person. For each point in a trajectory the velocity was calculated and then smoothed by taking the mean of a 24 frame sliding window. For the Benfold dataset body bounding boxes were not available so pedestrian velocity was approximated using the same process but applied to the centroids of the head bounding boxes rather than feet locations.

Formally, denote a persons velocity direction at frame $t$ as $\theta_t^v$ and their head pose direction as $\theta_t^g$. The head pose/direction deviation can then be calculated as the error $\epsilon_t = \theta_t^v - \theta_t^g$. The extracted deviations were then analysed to expose their statistically properties (mean, variance and distribution) from which probability distributions could be generated and analytically compared. Statistics were extracted for 37 pedestrians from the caviar dataset, 34 pedestrians from the PETS dataset, and 170 pedestrians from the Benfold dataset.

Fig. 2 highlights socially connected and unconnected persons in video frames from the caviar dataset, and shows the extracted behaviour statistics for all three datasets (see also details in Table I). The statistics show clear support for hypothesis 1 although the variance of distributions does differ between the datasets. In two datasets persons have a very high probability of looking in the direction of travel (head pose deviations close to 0). In the PETS dataset variance is larger, but is still peeked at 0. The Benfold dataset shows that there is little difference between the deviation patterns of the two social groups, while the caviar and PETS datasets show more significant differences. Performing the $\chi^2$ variance test shows that the differences between socially connected and unconnected deviations are statistically significant with a p-value of 0.05. This result supports hypothesis 2.

To conclude this section, our analysis has shown that irrespective of social connections, head pose and travel direction is well correlated. Furthermore, head pose deviation behaviour

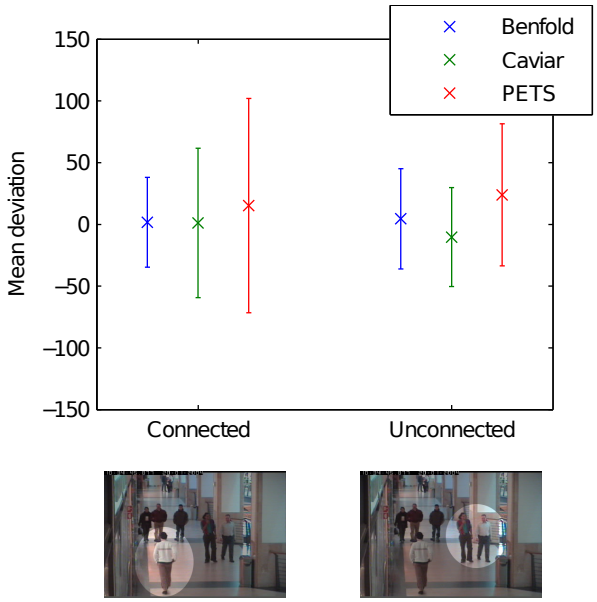


Fig. 2. a) Mean and standard deviations for head pose/velocity deviation (error) statistics extracted from three datasets. b) Example frames of connected and unconnected persons from the Caviar dataset

does differ (statistically significant) between socially connected and unconnected persons. We therefore accept the two hypotheses and use the remainder of this paper to present our approach for integrating intentional priors into tracking.

## III. Tracking with Intent

Having validated head pose as an intentional prior for pedestrian movement, we now focus on integrating an intentional prior into a tracking algorithm. In this instance we will use head pose deviation as the prior, although our algorithm remains generic and independent of the prior being used. As a basis for our tracker we will be using the Kalman filter [11], which is frequently used in computer vision research (e.g. [3]) due to the ease with which the motion model can be manipulated and the effect analysed. The Kalman filter provides an efficient (recursive) way of estimating the state of a system from a set of noisy measurements over time, where the seminal work can be found in [11]. As the basis for our tracker, we give a brief introduction to the relevant parts of the Kalmen filter before introducing the components of our 'Intentional Tracker' (for a complete introduction see [12]).

**Kalman filter basics:** Fundamentally, the Kalman filter attempts to estimate the state $x \in \Re^n$ of a discrete-time controlled process governed by the linear equation $x_t = F_t x_{t-1} + Bu_{t-1} + w_{t-1}$ with measurements $z_t = Hx_t + v_t$ (where $t$ indicate time). $w_t$ and $v_t$ are the process and measurement noise (respectively) and are assumed to be independent and normally distributed with zero mean and covariance $Q_t$ and $R_t$ (respectively). $B$ is the process control input model and $u_{t-1}$ is the control vector [12]. We assume that $B$ is the zero matrix so will not discuss it further and it will be omitted from later equations.

Matrix $F_t$ is often referred to as the motion or transition model and relates the state of the process at $t-1$ to $t$. Matrix H is the observation matrix which we assume to be constant.

The Kalman filter consists of prediction and update steps. The prediction step estimates the state of the system at time $t$ ($\hat{x}_t^-$) given all of the evidence prior to $t$ ($\hat{x}_{t-1}$), and predicts the error covariance matrix $P_t^-$ (details omitted for brevity):

$$\hat{x}_t^- = F_{t-1}\hat{x}_{t-1} + Bu_{t-1} \tag{1}$$

The predictions are then updated given the measurement $z_t$ to give the posteriori state estimate $\hat{x}_t^= \hat{x}_t^- + K_t(z_t - H\hat{x}_t^-)$, where $z_t - H\hat{x}_t^-$ is known as the measurement innovation and is a measure of the discrepancy between the predicted measurement $H\hat{x}_t^-$ and the actual measurement $z_k$. Matrix $K_t$ is the optimal Kalman gain which minimises the posteriori error covariance $P_t$ (see [12] for further details). In essence the Kalman gain gives more weight to the measurement as the measurement error covariance $R_t$ approaches zero, while more weight is given to the predicted measurement $H\hat{x}_t^-$ as the a priori estimate error covariance $P_t^-$ approaches zero. Finally, the error covariance is updated: $P_t = (I - K_t H_t)P_t^-$.

**Integrating intentional priors:** To integrate intentional priors into the Kalman filter we dynamically adjust the transition model $F_t$ according to the intentional prior. Denote $F_0$ as the initial motion model. During the prediction step at time $t$ we now generate a motion model $I_t$ based on the intentional prior, and combine this with the initial motion model $F_0$ using a weighting component $\alpha$. (In future work we will combine $I_t$ with $F_{t-1}$ rather than the constant $F_0$.)

We will first present the generation of $I_t$ for a head pose based prior which assumes zero acceleration and has the general form:

$$I_t = \begin{bmatrix} 1 & 0 & d_t cos(\theta_p) & 0 \\ 0 & 1 & 0 & d_t sin(\theta_p) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Where $d_t$ is the geometric distance travelled by the target between $t-1$ and $t$ and $\theta_p$ is the predicted direction of travel based on the velocity $v_t = [v_x, v_y]$ and estimated head pose deviation ($\theta_d$): $\theta_p = atan2(v_y, v_x) + \theta_d$, where $\theta_d$ is is assumed to be normally distributed: $\theta_d \sim N(\mu, \sigma)$ with parameters learnt from the scene and $atan2$ is the 4-quadrant arctangent function.

Having derived $I_t$ we use weighting component $\alpha_t$ to combine $I_t$ and $F_0$ as follows:

$$F_t = (1 - \alpha_t)F_0 + \alpha_t I_t \tag{3}$$

Intuitively $\alpha$ should increase in line with the strength of the intentional prior $\hat{s}_t$, where $\hat{s}_t$ combines magnitude and persistence. This can be achieved using a sigmoid function

with optimal parameter values $\gamma$ and $\tau$ derived via an optimisation procedure (optimisation details presented later). The $\gamma$ parameter adjusts the gradient at which the function moves from zero to one, while $\tau$ shifts the sigmoid along the x-axis. The resulting function can thus be adjusted to change the base-weight (weight given for zero strength) as well as the gradient at which the weight changes.

$$\alpha = \frac{1}{1 + exp(-\gamma(\hat{s}_t - \tau))} \tag{4}$$

To calculate $\hat{s}_t$ we use the absolute magnitude of the deviations for the last 10 time steps (arbitrarily chosen). To eliminate small fluctuations in deviation/detection inaccuracies. We use a binning procedure to partition the velocity and head pose into 8 bins (numerically numbered 1:8), where each bin represents a $45^\circ$ sector. The signal strength at time t is thus calculated as follows (where $\theta_k^g$ is the head pose direction and $\theta_k^v$ is the direction of travel):

$$\hat{s}_t = |\sum_{k=t-10}^{t} Bin(\theta_k^g) - Bin(\theta_k^v)| \tag{5}$$

Having finally defined all of the components required to generate $F_t$, the remainder of the Kalman filtering algorithm remains the same.

## IV. Experiments

We validate our approach in a visual surveillance application. Because our focus is the development of an 'Intentional Tracker' we assume that object detection and head pose estimations are provided. We compare the tracking performance of our tracker against the standard Kalman filter using both simulated and real examples from the Benfold dataset [1].

The simulated corpus contained 5 core trajectories of 200 time steps (each represented 500 times). Four trajectories contained a sharp turn at $t = 100$ ($+90^\circ$, $+45^\circ$, $-45^\circ$, $-90^\circ$), while the fifth was a straight track. Process noise and observation noise were added to all trajectories with ranges [0:0.2]m and [0:2.5]m respectively. For each trajectory head poses were generated using the Gaussian distribution extracted from the Benfold dataset (all individuals: $\mu = 3.788, \sigma = 39.504$). For $t > 1$ velocity direction was calculated to which the deviation was then added. To make the head poses match real behaviour we performed forward and backwards smoothing to simulate the person looking in the direction of the turn before the turn itself (forward window = 2, backward window = 20). Fig. 3 shows a representative track.

For simulated data we report the mean squared error (MSE) between the estimated track and ground truth. The ground truth information was not available for the real video scenarios so for those cumulative log likelihood (LL) is reported. This measures the likelihood of the innovations being drawn from the innovation covariance matrix. We report log likelihood for both simulated and real data. We use the Kalman filter (KF) as a comparative baseline and discuss the +/-% improvement of our intentional tracker over the KF.

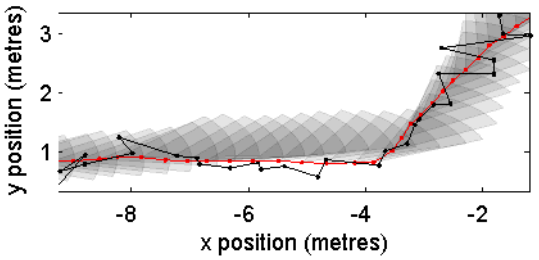


Fig. 3. Part of a simulated track showing a $45^\circ$ turn. The true target position is shown in red, observations in black, and head pose as grey sectors.

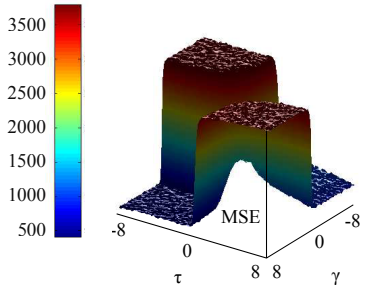


Fig. 4. Optimisation of sigmoid weighting parameters $\gamma$ and $\tau$ using a mean squared error (MSE) cost function. Optimal values are in the range $\gamma = [0, 8]$, $\tau = [-8, 0]$.

**Sigmoid optimisation:** In our optimisation procedure we minimised the MSE for $\gamma$ and $\tau$. Fig. 4 shows two local minima in the optimisation landscape for the $+90^\circ$ trajectory (all trajectories gave similar results). To understand the two local minima Fig. 5 shows the sigmoid output for values found in the two regions with lowest MSE. This figure highlights two interesting points. 1) Both regions give high weight to the intentional prior 'most' of the time (i.e. a strength of zero, which will be caused by little deviation). 2) Giving more weight to the intentional prior for strong deviation signals yields superior performance (bottom-left quadrant: $\gamma > 0$, $\tau < 0$). Furthermore, parameter values yielding the worst performance are those that give no weight to the intentional prior (not shown). The remainder of our experiments use $\gamma = 1.5$ and $\tau = -1.5$.

**Simulated corpus:** Fig. 6a shows the mean ($\mu$) and standard deviation ($\sigma$) of improvement in MSE for a corpus of 2500 trajectories with 0.01m process noise and 0.5m observation noise. For brevity we omit results for the other noise models. The figure shows that the intentional tracker delivers a mean improvement over the Kalman filter for all trajectories with range [13.46%, 23.61%]. Fig. 6b shows $\mu$ and $\sigma$ for the log likelihood (LL) improvement on the same data. Our improvements are in the range [3.8%, 6.29%].

**Video examples:** Finally, table II show tracking performance for 6 examples from the Benfold dataset [1]; 3 with sudden/large changes in trajectory, 3 without. The table shows

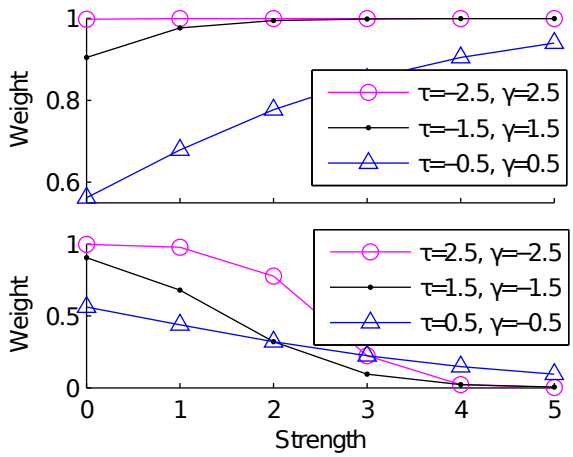


Fig. 5. Sigmoid output corresponding to the bottom-left quadrant of Fig. 4 (top), and bottom-right quadrant of Fig. 4 (bottom)

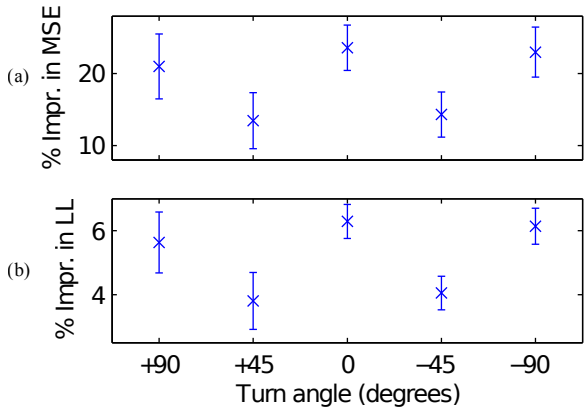


Fig. 6. Mean and +/- 1 Standard Deviation for the improvement in a) Mean Square Error (MSE) and b) Log Likelihood (LL) using the 'Intentional Tracker' over the Kalman Filter

TABLE II
PERCENTAGE IMPROVEMENT IN LOG LIKELIHOOD USING THE 'INTENTIONAL TRACKER' ON REAL DATA FROM THE BENFOLD VIDEO DATASET ([1]). TURN EXEMPLARS 1:3 HAVE APPROXIMATE TRAJECTORY CHANGES OF $-90°$, $-40°$, AND $-45°$ RESPECTIVELY.

| Trajectory | Ex. 1 | Ex. 2 | Ex. 3 | Mean |
|---|---|---|---|---|
| Turn | 18.50% | 9.18% | 16.33% | 14.67% |
| No-turn | 16.51% | 12.59% | 15.28% | 14.79% |

that for all six exemplar the intentional tracker out performs the Kalman filter, although further experiments are clearly required before any confident conclusions can be drawn.

## V. CONCLUSION

This work has presented justification for, and the novel theory of intentional priors which improve models of target behaviour. We then proposed that head pose is a good example of an intentional prior for pedestrian surveillance, and performed a statistical analysis of head pose behaviour across three video datasets. We showed that head pose and direction of travel are well correlated and provided statistical evidence that the intuition 'people look where they are going' is true. The results of our pedestrian tracking experiments confirm that our 'Intentional Tracker' is able to outperform the Kalman filter by as much as 23.61% on the simulated sample trajectories by means of reduced MSE. We also demonstrated performance on a sample of real pedestrian trajectories from the Benfold dataset ([1]), where the 'Intentional Tracker' achieved a mean improvement of 14.73% in log likelihood.

Better behaviour models allow us to make better behaviour predictions, from which our ultimate goal is to detect anomalies through observed inconsistency. **In future work** there are several key limitations that could be addressed: the algorithm needs further evaluation against a large set of real pedestrian trajectories, and needs to be demonstrated within a full signal processing chain in which all detections are determined algorithmically. Furthermore, the 'Intentional Tracker' needs to be applied to different target types to show cross-domain application.

## ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014277/1 and the MOD University Defence Research Collaboration in Signal Processing.

## REFERENCES

[1] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," *2011 International Conference on Computer Vision*, pp. 2344–2351, Nov. 2011.

[2] S. Pellegrini and L. Van Gool, "Tracking with a mixed continuous-discrete Conditional Random Field," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1215–1228, Oct. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1077314212001683

[3] K. Sankaranarayanan, M. Chang, and N. Krahnstoever, "Tracking gaze direction from far-field surveillance cameras," in *IEEE Workshop on Applications of Computer Vision*, 2011, pp. 519 – 526.

[4] B. Tordoff and D. Murray, "Resolution vs. tracking error: zoom as a gain controller," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Comput. Soc, 2003, pp. I–273–I–280.

[5] N. M. Robertson and I. D. Reid, "Automatic Reasoning about Causal Events in Surveillance Video," *EURASIP Journal on Image and Video Processing*, vol. Special Is, Sep. 2011.

[6] R. Mukherjee and N. Robertson, "Unconstrained head-pose estimation in real-time via low-resolution depth features," in *to appear at Computer Vision and Pattern Recognition (accepted paper)*, 2014.

[7] J. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1544–1551, Jun. 2012.

[8] M. Leach, E. Sparks, and N. Robertson, "Contextual Anomaly Detection in Crowded Surveillance Scenes," *Pattern Recognition Letters*, no. In Press, Dec. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865513004625

[9] CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. Edinburgh University Informatics Department.

[10] J. Ferryman and D. Tweed, "An overview of the pets 2007 dataset," in *Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS*, 2007.

[11] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," vol. 82, no. Series D, pp. 35–45, 1960.

[12] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," Department of Computer Science, University of North Carolina, Tech. Rep., 2006.