An adaptive motion model for person tracking with instantaneous head-pose features

Baxter, Rolf; Leach, Michael; Mukherjee, Sankha Subhra; Robertson, Neil

Published in:
IEEE Signal Processing Letters

DOI:
10.1109/LSP.2014.2364458

Publication date:
2014

An Adaptive Motion Model for Person Tracking with Instantaneous Head-Pose Features

Rolf H. Baxter, Michael J. V. Leach, Sankha S. Mukherjee, and Neil M. Robertson

***Abstract*—This letter presents novel behaviour-based tracking of people in low-resolution using instantaneous priors mediated by head-pose. We extend the Kalman Filter to adaptively combine motion information with an instantaneous prior belief about where the person will go based on where they are currently looking. We apply this new method to pedestrian surveillance, using automatically-derived head pose estimates, although the theory is not limited to head-pose priors. We perform a statistical analysis of pedestrian gazing behaviour and demonstrate tracking performance on a set of simulated and real pedestrian observations. We show that by using instantaneous 'intentional' priors our algorithm significantly outperforms a standard Kalman Filter on comprehensive test data.**

***Index Terms*— Computer vision, context awareness, deep belief networks, head pose estimation, tracking, video signal processing, video surveillance.**

I. Introduction

TRACKING error in the Kalman Filter (KF) increases when rapid changes in target motion occur. In part, this is caused by lag in adjusting the error covariance matrix. In this letter we reduce pedestrian tracking error by combining target velocity with an *intentional prior*, defined as a prior that predicts rapid changes in target motion. Specifically, we use the control input of the KF to steer the state estimate more forcefully using pedestrian gazing behaviour.

As motivation, consider a scene in which pedestrians exhibit ad-hoc obstacle avoidance (e.g. a goods-vehicle parked on the sidewalk). To model motion, two approaches are available; learning every eventuality (high model complexity), or learning a new (informative) feature. In pedestrian tracking, typical motion can be learnt by using flow vectors and clustering but often requires a strong assumption that motion patterns are stable [2]–[5]. Persistent changes can be incorporated over time but ad-hoc trajectories are still typically seen as outliers [6], [7]. The resulting models cannot accurately reflect pedestrian response to spatio-temporal context which could cause tracking failure and data-association errors, particularly if occlusions occur (Fig. 1). In such cases an intentional prior (feature) that can predict an ad-hoc change in trajectory is appealing. This theory also generalises to other intentional features: consider a car approaching a crossroads and the indicator light signals intention to turn; contextual knowledge enables better predictions. Several authors have incorporated the concept of 'personal space' and collision avoidance into pedestrian tracking [8], [9]. Others have incorporated the idea that socially grouped pedestrians will attempt to stay in close proximity [10], [11]. Both concepts represent different intentional priors. In our work we show a generic way of integrating intentional priors into a Kalman Filter and demonstrate performance with a novel head-pose prior.

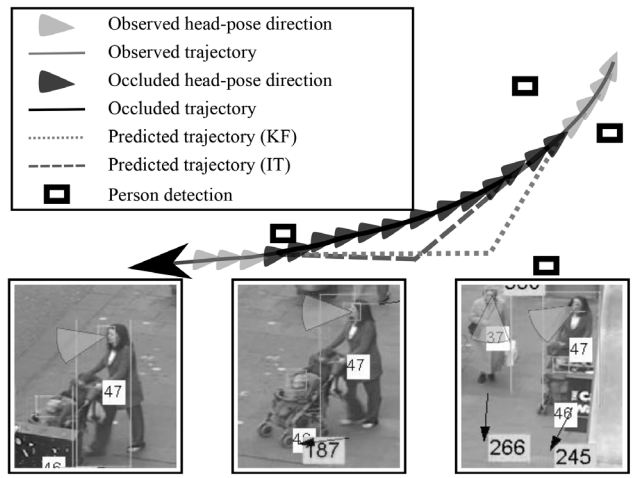

Fig. 1. (*Top*) A real person trajectory/head pose behaviour and predicted trajectory using a Kalman Filter (KF) and our intentional tracker (IT). Tracking failures can lead to target data association errors. (*Bottom*) Frames from the Benfold dataset [1] showing pedestiran head-pose.

Our pedestrian tracker takes as input the results of object detection and head-pose estimation. These areas are themselves challenging, especially in the presence of occlusions, camera motion and illumination changes [12]. Head-pose estimation is a thriving research topic producing ever increasing accuracy levels: [1] reports an error rate of 24° degrees on real surveillance video. [13]–[15] report similar accuracy and model anatomical constraints using joint body and head-pose estimation. None of this prior work estimates pedestrian position conditioned on head-pose, as we do in this letter.

Manuscript received August 08, 2014; revised October 08, 2014; accepted October 09, 2014. Date of publication October 22, 2014; date of current version November 03, 2014. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K014277/1 and the MOD University Defence Research Collaboration in Signal Processing. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ba-Tuong Vo.

R. H. Baxter, S. S. Mukherjee, and N. M. Robertson are with the Institute for Sensors, Signals and Systems, Heriot-Watt University, Edinburgh EH14 4AS, U.K. (e-mail: r.h.baxter@hw.ac.uk; sm794@hw.ac.uk; n.m.robertson@hw.ac.uk).

M. J. V. Leach is with Chemring Technology Solutions, Romsey, Hampshire SO51 0ZN, U.K. (e-mail: michael.leach@chemringts.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2364458

Robertson and Reid [16] have already shown that head pose can facilitate behaviour explanation in low/medium resolution images. Sankaranarayanan *et al.* proposed to use head-pose for pedestrian tracking in [17], and presented an algorithm for obtaining high-resolution face images of pedestrians on-the-fly using Pan-Tilt-Zoom cameras. Separate work by Tung *et al.* [12] and Dee and Hogg [18], [19] consider a target's goal location when making motion prediction, but in all cases rely on learnt goal and trajectory change locations. In contrast, we propose that target motion can be predicted from head-pose.

No prior work has used head-pose to predict pedestrian position and applied it to real video data. In this letter we present for the first time a full derivation and evaluation, following encouraging early work [20].

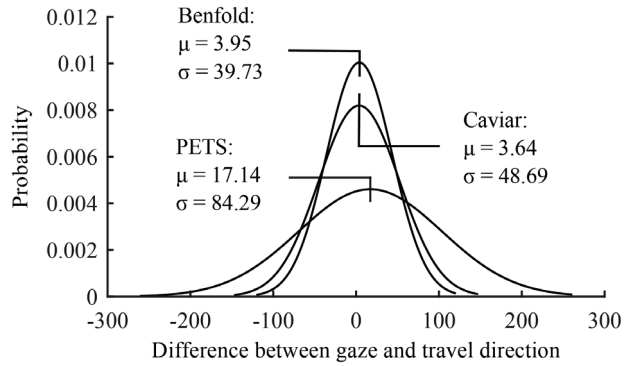

Fig. 2. Probability density functions and associated statistics for head pose deviations extracted from three video datasets.

II. Using Head Pose to Predict Behaviour

We consider the application of pedestrian surveillance and tracking to demonstrate the efficacy of our method. The assumption 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 is essential for dealing with occlusions and missing detections. We return to head-pose extraction in more detail in Section III-C.

We performed a statistical analysis of pedestrian trajectory and head pose behaviour to validate our hypothesis on three benchmark video datasets: Benfold [1], Caviar [21] and PETS 2007 [22]. We used manual annotations of person location (bounding box), head location (bounding box) and head pose direction (angle). We calculated the difference in angle between head-pose and the travel bearing for each pedestrian. For the Caviar and PETS datasets travel bearing 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 Direct Linear Transformation with point correspondances [23], 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.

Formally, denote a person's velocity direction at frame t as θ_t^v and their head pose direction as θ_t^g. The head pose/direction deviation can then be calculated as the error $\varepsilon_t = \theta_t^v - \theta_t^g$. The extracted deviations were then analysed to expose their statistical properties which were analytically compared. Mean and variance were extracted for 37 pedestrians from the caviar dataset, 34 pedestrians from the PETS dataset, and 154 pedestrians from the Benfold dataset.

Fig. 2 shows the probability density functions (PDFs) generated using the extracted statistics (underlying histograms were approximately Gaussian). The PDFs show clear support for the intuition that people look where they are going, showing high probability of head pose deviations close to zero. However, there are clear variations in behaviour between the datasets which suggests that any head pose based intentional tracker would need to be optimised for the scene to balance the reliability of the feature. Given these results, we use the remainder of this letter to present our approach for integrating an intentional priors into the KF with a head-pose based implementation.

III. Kalman Filter Adaptation

We now show how to integrate head pose information into a tracker. Note that although the algorithm is applied to pedestrian tracking our approach remains generic and different intentional priors could be used (e.g. car indicator). As a basis for our tracker we use the KF [24] due to its clear assumptions, wide spread use and efficiency.

A. Kalman Filter Preliminaries

For brevity we only highlight pertinent aspects of the KF (for a thorough introduction see [24], [25]). The KF estimates the state $x \in \Re^n$ of a discrete-time controlled process governed by the linear equation $x_t = F_t x_{t-1} + B u_{t-1} + w_{t-1}$ with measurements $z_t = H x_t + v_t$ (where t indicates time).

We represent the position and velocity of a target by the state vector $x_t = [pos^x, pos^y, \dot{x}, \dot{y}]^T$, where $\dot{x}$ and $\dot{y}$ represent the target's velocity with respect to its position.

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). F_t relates the state of the process at $t-1$ to t, B is the process control input model, u_{t-1} is the control vector (set to 1 in the experiments) and H is the observation matrix:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{1}$$

That is, we measure target position but not velocity, where a measurement z_t consists of the tuple $[pos^x, pos^y]_t$ and x_0 is initialised as: $x_0 = (H \times z_0)^T$.

B. Integrating Intentional Priors

We fuse intentional priors into the KF, firstly, by calculating the strength of the prior, denoted $\hat{s}_t$, using the absolute magnitude of the deviations for the last 10 time steps (arbitrarily chosen). This allows $\hat{s}_t$ to combine both the magnitude and persistence of the prior signal. Rather than using the raw angles,

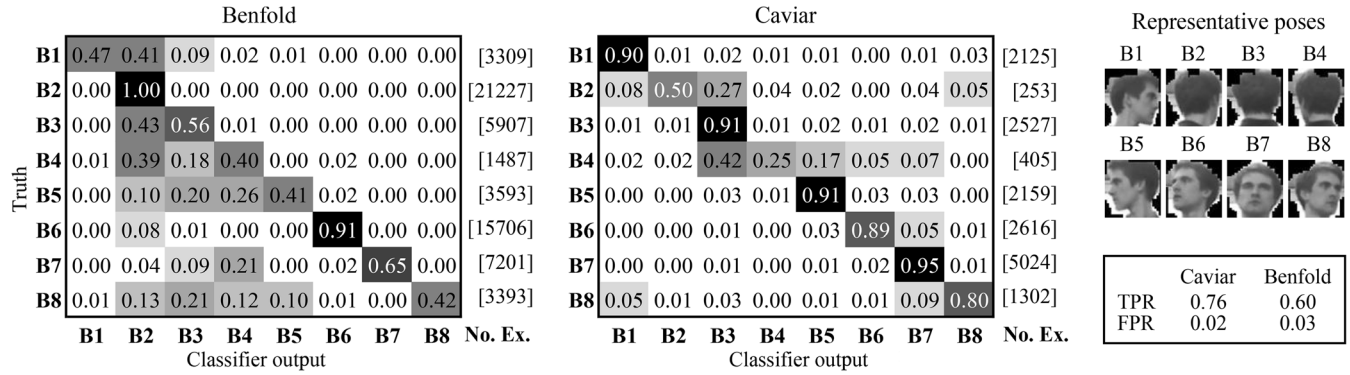

Fig. 3. Confusion matrices and true positive/false positive rates (TPR/FPR) for our deep belief network head-pose classifier on the Benfold and Caviar datasets. In square brackets: Number of head-pose examples.

we eliminate small fluctuations in deviation/detection inaccuracies by using a binning procedure to partition the velocity and head pose into 8 bins (numerically numbered 1:8). Each bin represents a 45° sector (see Fig. 3). This procedure allows a smoothed estimate of the head-pose deviation signal (discussed in Section II) to be generated. The signal strength at time t is then calculated as follows (where θ^g_k is the head pose direction and θ^v_k is the direction of travel):

$$\hat{s}_t = \left| \sum_{k=t-10}^{t} Bin(\theta^g_k) - Bin(\theta^v_k) \right| \tag{2}$$

Next, we weight the influence of the prior. Intuitively, the weight (α_t) should increase in line with the strength of the prior $\hat{s}_t$. A sigmoid function applied to $\hat{s}_t$ is a simple and effective way to achieve this. The sigmoid is parameterised by ρ and τ and could be optimised for the scene to reflect the reliability of the prior, where ρ adjusts the rate at which the function moves from zero to one and τ adjusts the 'base-weight (weight given for zero strength). Rather than optimising for any particular scene, we use values for ρ and τ that were empirically derived in [20] (see Section IV for further details).

$$\alpha_t = (1 + exp(-\rho(\hat{s}_t - \tau)))^{-1} \tag{3}$$

Having determined α_t, the transition model (F_t) is adjusted to reduce the influence of the target's previous motion. Denote F_{t-1} as the motion model at time $t-1$ and $\gamma_t = 1 - \alpha_t$. The motion model is then updated as follows:

$$F_t = \begin{bmatrix} 1 & 0 & \gamma_t & 0 \\ 0 & 1 & 0 & \gamma_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

This has the effect of reducing the influence of $\dot{x}$ and $\dot{y}$ by a factor of γ_t during the prediction step of the algorithm. The influence of the intentional prior is asserted using the control matrix B_t:

$$B_t = [\alpha_t dx, \alpha_t dy, \alpha_t dx, \alpha_t dy]^T \tag{5}$$

$$dx = d_t cos(\theta_p), dy = d_t sin(\theta_p) \tag{6}$$

Where d_t is the geometric distance travelled by the target between $t-1:t$ and θ_p is the predicted travel direction based on head pose angle θ^d_{t-1}. Two approaches could be used for calculating d_t: It could be estimated from $[\dot{x}_{t-1}, \dot{y}_{t-1}]$, which is an estimate of the targets velocity given observations $z_{0:t-1}$. Alternatively, a smoothed velocity could be calculated from $[pos^x_{t-k:t-1}, pos^y_{t-k:t-1}]$, where $2 \leq k \leq t$. In practice the second approach was found to give better performance using empirically derived $k = 5$.

Having finally defined all of the components required to generate F_t, the remainder of the KF algorithm remains the same. Predictions are now based on a target's previous motion (with weight γ_t) and the intentional prior (with weight α_t).

C. Head-Pose Extraction

We validate our approach in a visual surveillance application. Although not the focus of our work, we briefly discuss the novel head-pose extraction procedure used within our validation. We trained a Deep Belief Network [26] using the combined datasets: [1], [21], [27]. Heads were binned into 45° poses and reflected in the y axis to reduce dataset bias. The histogram equalised raw image data (cropped head bounding boxes) were scaled to 32×32 pixels each and provided to a Deep Belief Network parameterised as follows: Number of units per layer; 1024, 400, and 8 (first, second and third layers respectively), dropout rate; 20% (layer 1 only), unit type; rectified linear (layers 1 and 2), softmax (classifier layer (3)). All layers were trained for ≈ 1000 Epochs each using variable learning rates. For the third layer we used an 80:20 train/test split for the Benfold dataset, and a 50:50 split for the caviar data. The resulting confusion matrices are shown in Fig. 3.

IV. Experiments

We compare performance of our tracker against the standard KF (by which we mean having no head-pose information) using the Benfold [1] and Caviar [21] video datasets. To overcome the limited presence of obstacle avoidance behaviours within these datasets (manifested as trajectories with sharp turns), we also compare performance on a corpus of simulated trajectories and hope to obtain additional video examples in the future. Our focus is the development of an intentional tracker so primarily use hand annotated head-poses under the assumtion that object detection/head-pose estimation is provided by the current

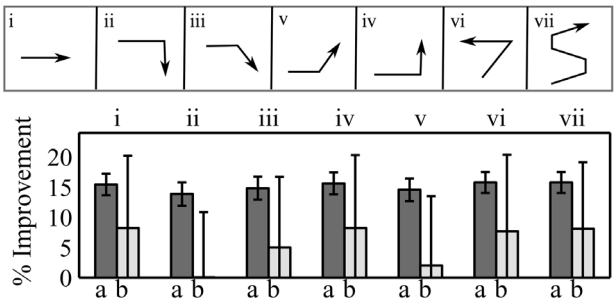

Fig. 4. (*Top*) Trajectories for (i) basic and (ii-vii) obstacle avoidance behaviour. (*Bottom*) Improvement gain by our intentional tracker vs. a standard KF on the simulated corpus. Detection rate:100% (showing median ±1 std. deviation). (a) Cumulative Log Likelihood, (b) Mean Square Error.

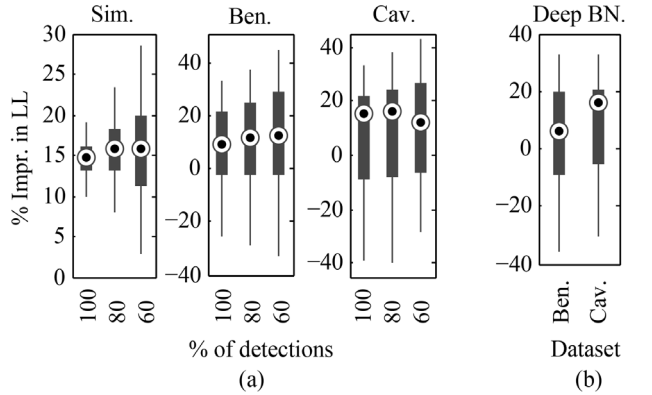

Fig. 5. Improvement in Cumulative Log Likelihood (LL) by our intentional tracker vs. a standard KF. (a) Using the simulated, Benfold, & Caviar datasets under three head/body detection rates & hand-annotated head-pose. (b) Using headpose classifications from our deep belief network (Deep BN).

state-of-the-art. However, as final validation we also report performance using detections from our deep belief network (Deep BN.) head-pose classifier.

Throughout the experiments $Q = 0.1 \times I_4$ and $R = 0.5 \times I_2$, where I_n indicates the n-dimensional identity matrix.

We use improvement in mean squared error (MSE) and improvement in cumulative log likelihood (CLL) as our evaluation metrics, where MSE is the sum of squared differences between predicted and real trajectories and improvement is defined as: MSE_{KF}/MSE_{IT} (IT: intentional tracker). CLL is based on the measurement innovation and is defined as $CLL_{KF} = \sum_{k=1}^{T} LL_k^{KF}$ and $CLL_{IT} = \sum_{k=1}^{T} LL_k^{IT}$. Improvement in CLL is: CLL_{KF}/CLL_{IT}. CLL measures how well the innovation covariance is modelled and is a useful metric when MSE cannot be calculated. We used the optimal sigmoid parameters derived empirically in [20] throughout our experiments ($\rho = 1.5, \tau = -1.5$) which gives high weight to the head-pose prior. No further parameter optimisation or tuning was performed for any of the datasets. For both the simulated and real datasets we synthesised pedestrian detection errors at different rates by withholding 0-40% of observations (uniform distribution).

A. Obstacle Avoidance

We consider obstacle avoidance trajectories using a simulated corpus containing 3500 trajectories of 200 time steps (typical track length in a surveillance video). Representative trajectories are shown in Fig. 4 to which Gaussian noise was added to the true target positions ($\mu = 0, \sigma = 0.1$) and to observations ($\mu = 0, \sigma = 0.5$). The direction of travel between t and $(t + 5)$ was used to simulate head-pose direction to which Gaussian distributed noise was also added ($\mu = 4, \sigma = 20$).

Fig. 4 shows that our approach outperforms the baseline for each trajectory. Performance is degraded by the sharpness of trajectory changes, with worst performance obtained for trajectories *ii* and *v*.

B. Annotated Detections

Fig. 5(a) shows performance on the video datasets when using annotated detections. This consisted of person head-pose for the Intentional Tracker and body bounding box for the standard KF. Our approach out performs the standard KF under all conditions.

TABLE I
PERCENTAGE IMPROVEMENT (REDUCTION) IN MEAN SQUARED ERROR (MSE) DURING OCCLUSION FOR 7 TRAJECTORIES

Traj. No.	7	8	9	10	11	15	22
% Imp.	64.0	75.5	84.5	12.9	73.2	62.0	68.3

At a detection rate of 60% we maintained improvements of 12.7% (Benfold) and 12.2% (Caviar). The video datasets contained fewer challenging (e.g. sharp turn) trajectories than the simulated corpora, but head-pose behaviour was occasionally effected by distractions (e.g. shop windows) making all datasets equally challenging.

C. Real Detections

We next evaluate tracker performance with real head-pose classifications. For the Benfold dataset head detections were provided by a re-implementation of [1]. For Caviar the hand-annotated head-detections we used. For both datasets detected heads were classified using our novel Deep BN. head-pose classifier (Fig. 3).

Fig. 5(b) shows that we achieved median improvements of 5.9% on the Benfold data and 15.8% on Caviar. Since there are only 7 examples of sudden trajectory changes in the Benfold dataset (none are occluded), we synthesised occlusions on these trajectories. Specifically, for each change in trajectory we withheld a window of observations from each tracker to occlude the change (see Fig. 1). Table I shows the improvement (i.e. reduction) in mean squared error (MSE) between the predicted and withheld pedestrian observations. A mean reduction of 62.9% was achieved across the 7 trajectories.

V. CONCLUSION

This work has shown that head-pose and direction of travel are well correlated in some environments and we have proposed head pose as a good intentional prior for pedestrian surveillance. Our experimental validation showed that our intentional tracker could significantly outperform the standard KF on both video, and synthetic datasets containing sudden changes in behaviour. In the future we intend to use contextual information to switch between different intentional priors.

REFERENCES

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

[2] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image Vis. Comput.*, vol. 14, pp. 583–592, 1996 [Online]. Available: http://www.bmva.org/bmvc/1995/bmvc-95-057.html

[3] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, 2000 [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=868677

[4] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–64, Sep. 2006 [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16929731

[5] B. Morris and M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1114–1127, Aug. 2008 [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4543858

[6] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *Syst., Man, Cybern. B*, vol. 35, no. 3, pp. 397–408, 2005.

[7] C. Piciarelli and G. Foresti, "On-line trajectory clustering for anomalous events detection," *Patt. Recognit. Lett.*, vol. 27, no. 15, pp. 1835–1842, Nov. 2006 [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865506000432

[8] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1805–19, Nov. 2005 [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16285378

[9] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th Int. Conf. Computer Vision. IEEE*, Sep. 2009, pp. 261–268 [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5459260

[10] J. C. S. Jacques, A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using voronoi diagrams," *Patt. Anal. Applicat.*, vol. 10, no. 4, pp. 321–332, Apr. 2007.

[11] S. Pellegrini and L. Van Gool, "Tracking with a mixed continuous-discrete conditional random field," *Comput. Vis. Image Understand.*, vol. 117, no. 10, pp. 1215–1228, Oct. 2013.

[12] F. Tung, J. S. Zelek, and D. a. Clausi, "Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance," *Image Vis. Comput.*, vol. 29, no. 4, pp. 230–240, Mar. 2011.

[13] S. Ba and J. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *Int. Conf. Pattern Recognition*, 2004, pp. 264–267.

[14] C. Chen, A. Heili, and J.-M. Odobez, "A joint estimation of head and body orientation cues in surveillance video," in *IEEE Int. Conf. Computer Vision Workshops*, Nov. 2011, pp. 860–867.

[15] C. Cheng and J. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *IEEE Conf. Comput. Vis. Patt. Recognit.*, Jun. 2012, pp. 1544–1551.

[16] N. M. Robertson and I. D. Reid, "Automatic reasoning about causal events in surveillance video," *EURASIP J. Image Video Process.*, vol. Special Is, Sep. 2011.

[17] 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.

[18] H. M. Dee and D. C. Hogg, "On the feasibility of using a cognitive model to filter surveillance data," in *IEEE Conf. Advanced Video and Signal Based Surveillance*, 2005, pp. 34–39.

[19] H. Dee and D. Hogg, "Navigational strategies in behaviour modelling," *Artif. Intell.*, vol. 173, no. 2, pp. 329–342, Feb. 2009.

[20] R. H. Baxter, M. Leach, and N. M. Robertson, "Tracking with intent," *Sensor Signal Process. Defence*, 2014.

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

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

[23] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd Ed. ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. D, pp. 35–45, 1960.

[25] G. Welch and G. Bishop, An Introduction to the Kalman Filter Dept. Comput. Sci., Univ. North Carolina, Chapel Hill, NC, USA, Tech. Rep., 2006.

[26] G. E. Hinton and S. Osindero, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[27] L. Bazzani, M. Cristani, and D. Tosato, "Social interactions by visual focus of attention in a three dimensional environment," *Expert Syst.*, vol. 30, no. 2, pp. 115–127, 2013 [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0394.2012.00622.x/full