UDRC Edinburgh

Tracking With Intent

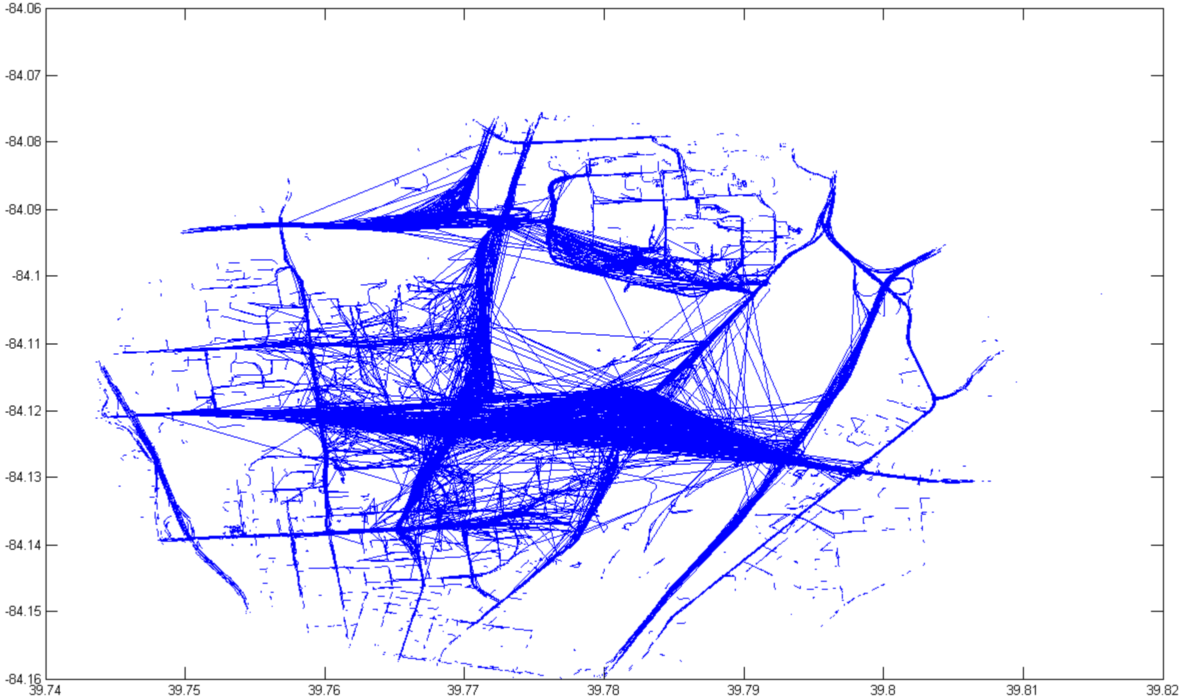
Rolf Baxter[1] , Michael Leach[1,2], Neil Robertson[1]

[1] Vision Lab, Institute of Sensors, Signals and Systems, Heriot-watt University, Edinburgh
[2] Roke Manor Research (Chemring Technology Solutions), Romsey, Hampshire

[dstl] EPSRC Engineering and Physical Sciences Research Council

Motivation: The end goal

- Identifying anomalous behaviour in the proverbial haystack to enhance situation awareness



[WPAFB 2009 Dataset]

Approach: Build better models of normality by using more of the signal

- Specifically, we propose the use of *intentional priors*
  - Priors that are indicative of future intent
  - Could be derived from different signals (e.g. car indicator, AIS, pattern of life)
  - Could be context sensitive

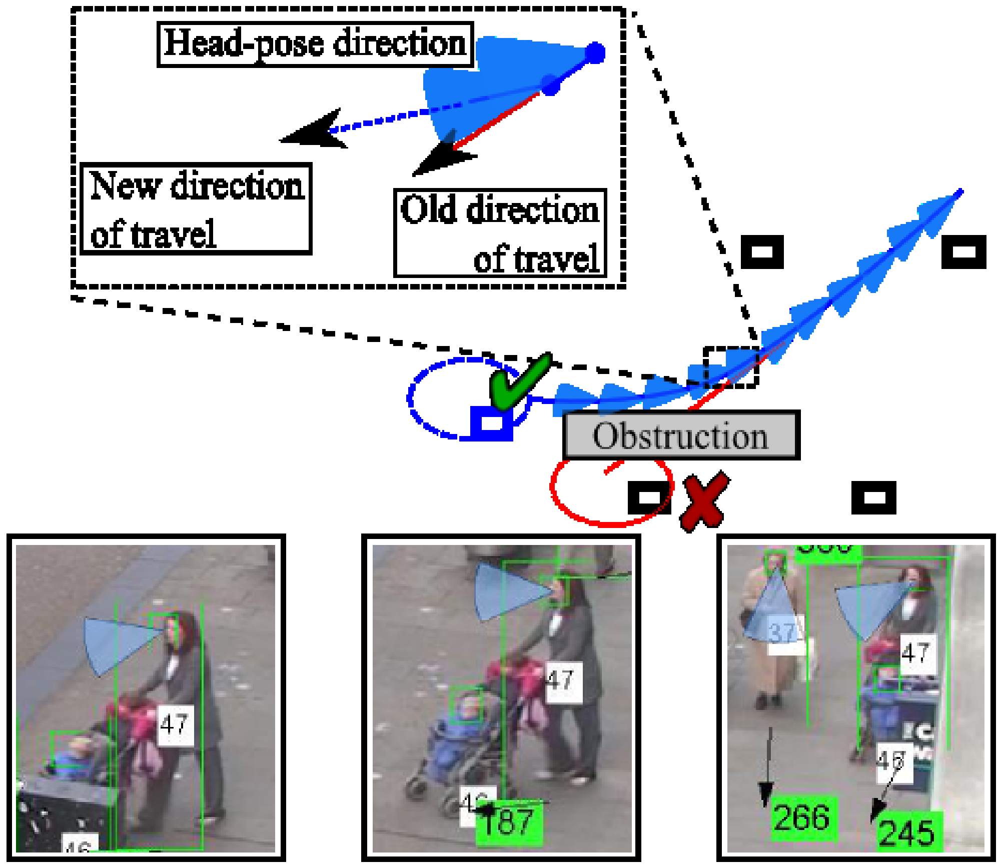
- This talk focuses on person tracking in video with head-pose priors
  - Relevant to automated visual surveillance (e.g. base protection)
  - People perform a broad range of behaviours so represent challenging targets
  - Concept is extensible to other real-world targets

Motivating Example

Intuition: Head-pose is an informative intentional prior

- Head pose can provide both spatial and social context

- We can use head-pose to build better person trackers



Related Work

- Recent work has shown that performing head-pose estimation within the outdoor built environment is reasonable
  - Odobez at Idiap [5]
  - Benfold at Oxford [6]
  - Our latest work at Heriot-Watt (IEEE Sig.Proc.Letters, to appear)

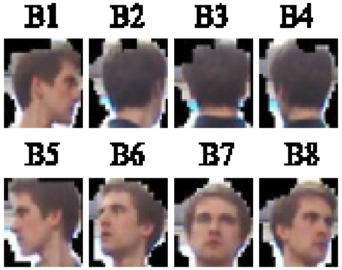
Benfold

Truth \ Classifier output	B1	B2	B3	B4	B5	B6	B7	B8	No. Ex.
B1	0.47	0.41	0.09	0.02	0.01	0.00	0.00	0.00	[3309]
B2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	[21227]
B3	0.00	0.43	0.56	0.01	0.00	0.00	0.00	0.00	[5907]
B4	0.01	0.39	0.18	0.40	0.00	0.02	0.00	0.00	[1487]
B5	0.00	0.10	0.20	0.26	0.41	0.02	0.00	0.00	[3593]
B6	0.00	0.08	0.01	0.00	0.00	0.91	0.00	0.00	[15706]
B7	0.00	0.04	0.09	0.21	0.00	0.02	0.65	0.00	[7201]
B8	0.01	0.13	0.21	0.12	0.10	0.01	0.00	0.42	[3393]

Caviar

Truth \ Classifier output	B1	B2	B3	B4	B5	B6	B7	B8	No. Ex.
B1	0.90	0.01	0.02	0.01	0.01	0.00	0.01	0.03	[2125]
B2	0.08	0.50	0.27	0.04	0.02	0.00	0.04	0.05	[253]
B3	0.01	0.01	0.91	0.01	0.02	0.01	0.02	0.01	[2527]
B4	0.02	0.02	0.42	0.25	0.17	0.05	0.07	0.00	[405]
B5	0.00	0.00	0.03	0.01	0.91	0.03	0.03	0.00	[2159]
B6	0.00	0.00	0.01	0.00	0.03	0.89	0.05	0.01	[2616]
B7	0.00	0.00	0.01	0.00	0.01	0.02	0.95	0.01	[5024]
B8	0.05	0.01	0.03	0.00	0.01	0.01	0.09	0.80	[1302]

Representative poses



	Caviar	Benfold
TPR	0.76	0.60
FPR	0.02	0.03

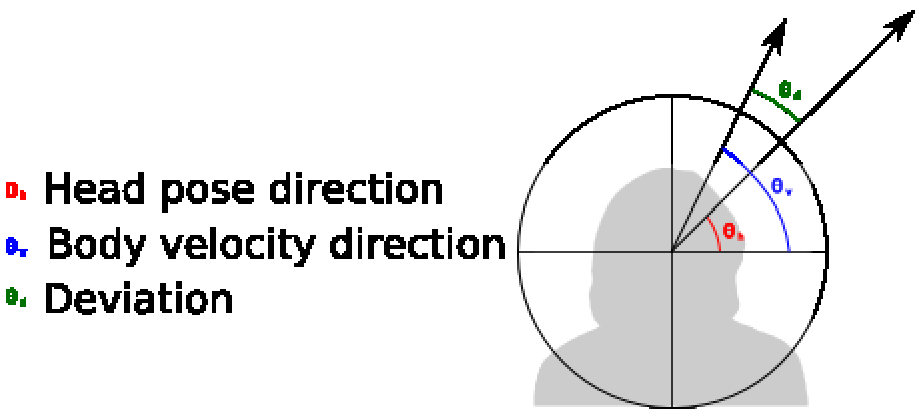
- A number of trackers consider social context when making target predictions (e.g. Pelligrini et al. [7])

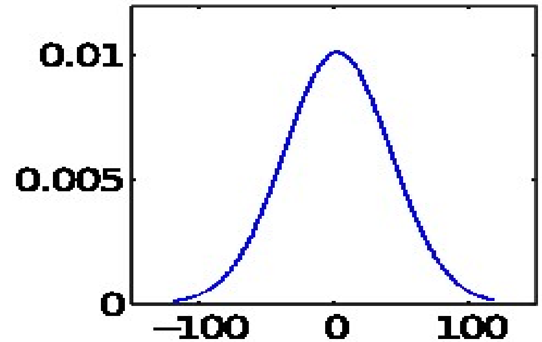
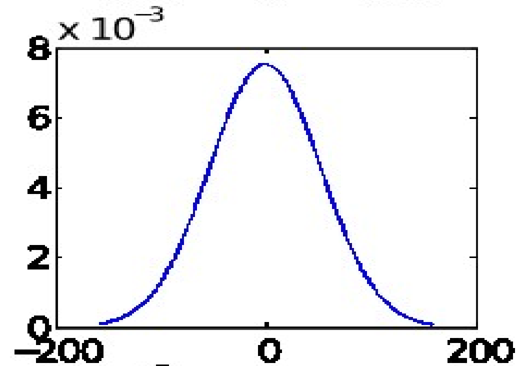
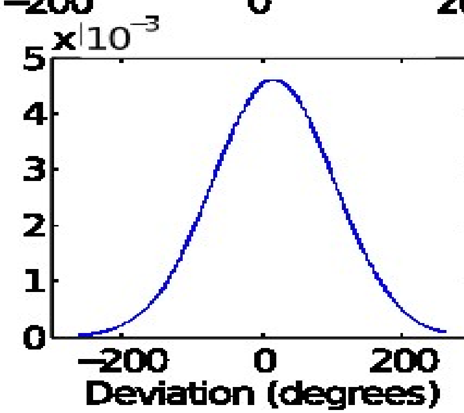
- Sankaranaraynan et al. fused person tracking with a PTZ facial tracking system, but did not use head-pose to predict target location [8].

- **No prior work has used head-pose to aid target tracking**
  - ➔ **The video signal is being under utilised**

Is the signal present?

- Is head-pose is well correlated with direction of travel?
- Analysis: 3 datasets - Caviar, PETS and Oxford [10,11,6]
- Using automatic detections [12] and ground truth head-pose annotations



Dataset	Example frame	PDF of head pose/velocity error
Benfold		
Caviar		
PETS 2007		

The Kalman Filter

Time update ("predict")

1) Project the state ahead

$$\hat{\boldsymbol{x}}_t^- = \boldsymbol{F}_{t-1}\hat{\boldsymbol{x}}_{t-1} + \boldsymbol{B}\boldsymbol{u}_{t-1}$$

2) Project the error covariance ahead

$$\boldsymbol{P}_t^- = \boldsymbol{F}_{t-1}\hat{\boldsymbol{x}}_{t-1}\boldsymbol{F}_{t-1}^T + \boldsymbol{Q}_t$$

Measurement update ("correct")

1) Compute the innovation

$$\boldsymbol{\epsilon}_t = \boldsymbol{z}_t - \boldsymbol{H}\hat{\boldsymbol{x}}_t^-$$

2) Compute the Kalman Gain K_t

$$\boldsymbol{K}_t = \boldsymbol{P}_t^-\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_t^-\boldsymbol{H}^T + \boldsymbol{R})^{-1}$$

3) Update estimate with Z_t

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_t^- + \boldsymbol{K}_t\,\boldsymbol{\epsilon}_t$$

4) Update the error covariance

$$\boldsymbol{P}_t = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}_t)\boldsymbol{P}_t^-$$

State vector: $x_t = [x, y, \dot{x}, \dot{y}]^T$

Observation matrix: $H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$

Mixing component: $\alpha_t = (1 + \exp(-\rho(s_t - \tau)))^{-1}$

$\gamma_t = 1 - \alpha_t$

Motion model:

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

Intentional prior : $B_t = [\alpha_t dx,\ \alpha_t dy, \alpha_t dx,\ \alpha_t dy]^T$

$$\left.\begin{array}{l} d_t \;::\; \text{Distance}(\hat{x}_{t-1}, \hat{x}_t) \\ \theta_t^h \;::\; \text{head–pose direction at t} \end{array}\right\} \quad \begin{array}{l} dx = d_t \cos\theta_t \\ dy = d_t \sin\theta_t \end{array}$$

Mixing component: $\alpha_t = (1 + \exp(-\rho(s_t - \tau)))^{-1}$

Strength of prior:

$$S_t :: \left| \sum_{k=\max(0,t-10)}^{t} Bin(\theta_k^h) - Bin(\theta_k^v) \right|$$

θ_k^v :: Travel direction

θ_k^h :: Head–pose direction

ρ :: Sensitivity

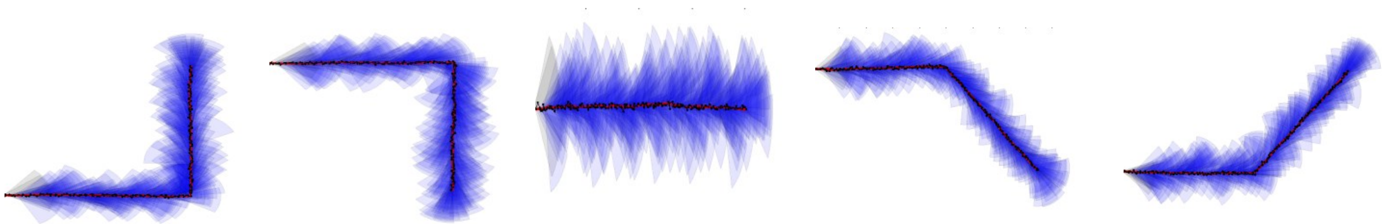
τ :: Base weight

Head-pose binning:

[4]

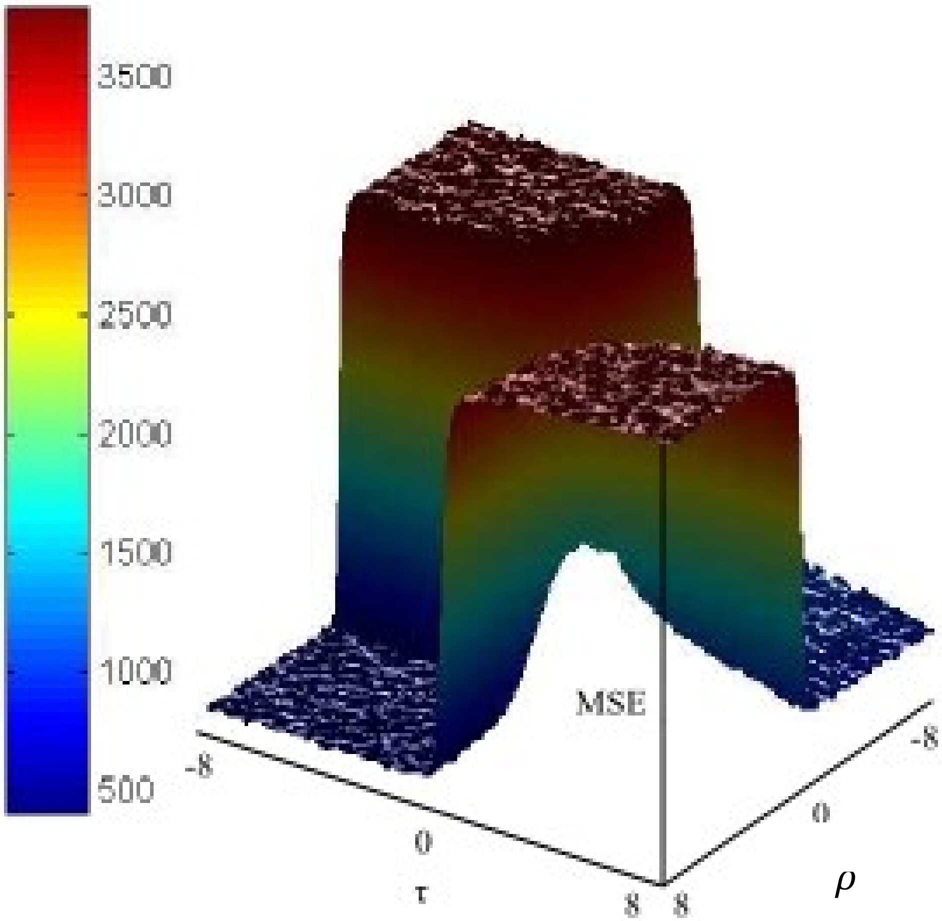
Evaluation

- 5 core trajectories:

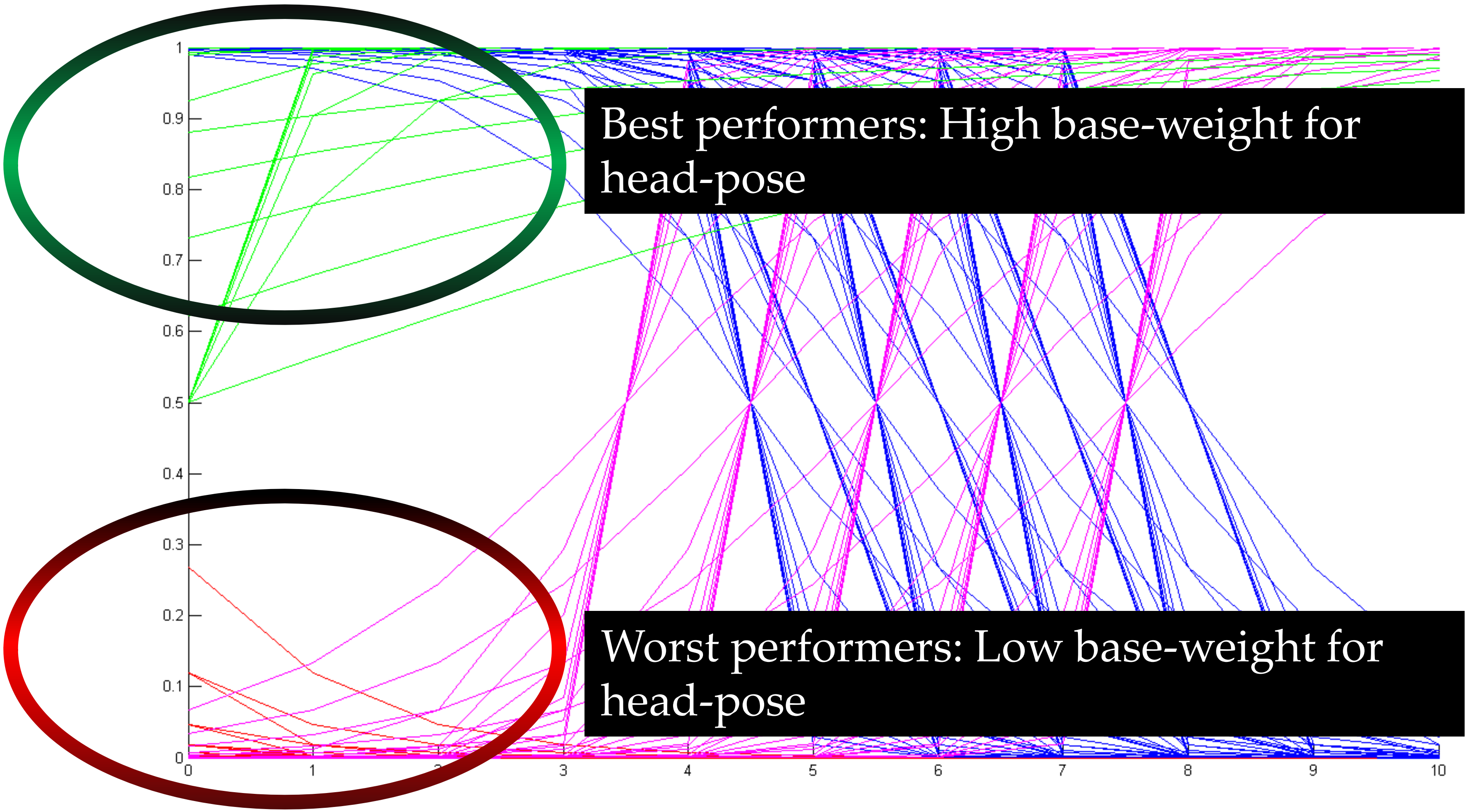


- Gaussian noise added to positions and head-poses
- Comparative baseline:
  - Standard Kalman Filter (i.e. without head-pose)

Sigmoid optimisation



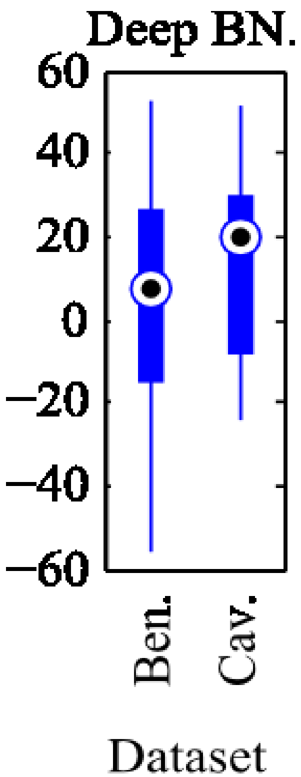
Sigmoid optimisation



Performance with Deep BN. head-pose

- **Median improvements:**
  - **Ben: 7.21%**
  - **Cav: 19.5%**

- **Optimising the head-pose classifier for the scene could improve this.**



(b)

Conclusions & Future Work

- Shown how to integrate an intentional prior into the Kalman Filter (KF)

- Validation has shown that using head-pose intentional priors we can make better target predictions

- Key next steps:
  - How do we integrate anomaly detection?
  - How do we learn and use different kinds of intentional prior & context?

Questions

?

References

[1] **Baxter,** Rolf H., Robertson, Neil M., Lane, David M. "Real-time Event Recognition from Video via a Bag-of-activities''. In proceedings of the UAI Bayesian Modelling workshop, ***2011***.

[2] **Li,** J., Hospedales, T. M., Gong, S. and Xiang, T. "Learning Rare Behaviours'. Lecture Notes in computer Science, Volume 6493, pp 293-307, ***2011***.

[3] **Leach,** Michael J. V., Sparks, Edward., Robertson, Neil M. "Contextual Anomaly Detection in Crowded Surveillance Scenes" *Pattern Recognition Letters (to appear)*, Volume 44, pp 71-79, ***2014.***

[4] **Mukherjee**, R., Robertson, Neil M., "Unconstrained head-pose estimation in real-time via low-resolution depth features". In proceedings of *CVPR (to appear)*. ***2014.***

[5] **Odobez,** J. M., Chen, C., "We are not Contortionists: couples Adaptive Learning for Head and Body Orientation Estimation in Surveillance Video". In proceedings of *CVPR*. ***2012.***

[6] **Benfold,** B., Reid, I., "Colour Invariant Head Pose classification in Low Resolution Video". In proceedings of *the 19th British Machine Vision Conference*. ***2008.***

[7] **Pellegrini,** S., Gool, L. V., "Trackin with a mixed continuous-discrete Conditional Random Field". Computer Vision and Image Understanding. Volume 117, No. 10, pp 1215-1228, ***2013.***

[8] **Sankaranarayana,** K., Chang, M., Krahnstoever, N. "Tracking gaze direction from far field surveillance cameras". In IEEE workshop on Applications of Computer Vision, pp 519-526. ***2011.***

[9] **Tordoff,** B., Murray, D. "Resolution vs. tracking error: zoom as a gain controller". In proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp273-280. ***2003.***

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

References

[11] **Ferryman,** J., and Tweed, D. "An overview of the PETS 2007 dataset''. In proceedings of the IEEE workshop on Performance Evlauation of Tracking and Surveillance, ***2007***.

[12] **Leach,** Michael J., Baxter, Rolf H., V., Sparks, Edward., Robertson, Neil M. "Social Grouping Using Visual Attention in Crowded Surveillance Video'. To appear in proceedings of the CVPR workshop on Computational Models of Social Interactions and Behaviour, ***2014***.

[13] **Kalman,** R. E. "A New approach to Linear Filtering and Prediction Problems" Journal of Basic Engineering, No. 82 (series D), pp 35-45. ***1960.***

Deep Belief Net. Classifier

Benfold

Truth \ Classifier output	B1	B2	B3	B4	B5	B6	B7	B8	No. Ex.
B1	0.47	0.41	0.09	0.02	0.01	0.00	0.00	0.00	[3309]
B2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	[21227]
B3	0.00	0.43	0.56	0.01	0.00	0.00	0.00	0.00	[5907]
B4	0.01	0.39	0.18	0.40	0.00	0.02	0.00	0.00	[1487]
B5	0.00	0.10	0.20	0.26	0.41	0.02	0.00	0.00	[3593]
B6	0.00	0.08	0.01	0.00	0.00	0.91	0.00	0.00	[15706]
B7	0.00	0.04	0.09	0.21	0.00	0.02	0.65	0.00	[7201]
B8	0.01	0.13	0.21	0.12	0.10	0.01	0.00	0.42	[3393]

Caviar

Truth \ Classifier output	B1	B2	B3	B4	B5	B6	B7	B8	No. Ex.
B1	0.90	0.01	0.02	0.01	0.01	0.00	0.01	0.03	[2125]
B2	0.08	0.50	0.27	0.04	0.02	0.00	0.04	0.05	[253]
B3	0.01	0.01	0.91	0.01	0.02	0.01	0.02	0.01	[2527]
B4	0.02	0.02	0.42	0.25	0.17	0.05	0.07	0.00	[405]
B5	0.00	0.00	0.03	0.01	0.91	0.03	0.03	0.00	[2159]
B6	0.00	0.00	0.01	0.00	0.03	0.89	0.05	0.01	[2616]
B7	0.00	0.00	0.01	0.00	0.01	0.02	0.95	0.01	[5024]
B8	0.05	0.01	0.03	0.00	0.01	0.01	0.09	0.80	[1302]

Representative poses

	Caviar	Benfold
TPR	0.76	0.60
FPR	0.02	0.03

- **Unlike competing approaches, we do not use motion or body information to 'classify' head-pose**
- **Poorest performance collates with fewest training examples**