Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator

Ben Benfold and Ian Reid
Department of Engineering Science
University of Oxford
Oxford, UK
{bbenfold,ian}@robots.ox.ac.uk

Abstract

We present a method to estimate the coarse gaze directions of people from surveillance data. Unlike previous work we aim to do this without recourse to a large hand-labelled corpus of training data. In contrast we propose a method for learning a classifier without any hand labelled data using only the output from an automatic tracking system. A Conditional Random Field is used to model the interactions between the head motion, walking direction, and appearance to recover the gaze directions and simultaneously train randomised decision tree classifiers. Experiments demonstrate performance exceeding that of conventionally trained classifiers on two large surveillance datasets.

1. Introduction

(Preprint)[1] Our aim is to automatically identify the direction in which people are facing as a coarse estimate of their gaze direction. Doing so in unconstrained environments is particularly difficult because of the many variables affecting the appearance of an individual. Facial structure, hair style and colour, skin colour, blurry images, lighting conditions and accessories such as hats and sunglasses contribute to large variations in the appearance of different people looking in the same direction compared with potentially subtle differences between the appearance of the same person looking in different directions. Recent approaches to estimating gaze direction in surveillance scenarios have treated this as a classification problem and quantised the gaze direction into one of 8 classes, each one representing 45° of the full 360° range. These classifiers are learned using a large corpus of hand-labelled training exemplars [1, 2, 16, 17, 7, 13]. In this paper we examine the possibility of learning to estimate gaze directions in an unsupervised manner using the output of an automatic head tracking system. By observing a scene for an extended period we can have some confidence that the head tracker acquires data representative of all the classes. Furthermore we also note that people tend to look in their direction of motion. The combination of these two factors yields the potential to automatically acquire a very large set of weakly labelled data without human intervention. Furthermore, such a training set has the potential to yield classifiers customised to the specific conditions of particular installation, such as the viewpoint, focus and lighting conditions which would be difficult to train for explicitly. Nevertheless, automatic acquisition of training data in this manner will inevitably yield a high percentage of incorrect labels, not only because people do not always look in their direction of travel, but also because the images of the heads may not be well centred or could represent false positives, as shown in figure 1.

Figure 1. Randomly chosen sequences from the two video datasets. The sequences demonstrate the problems of image blur, tracking failures, incorrect detections and unusual appearances caused by clothing such as hats.

We use a head tracking [3] system to acquire data con-

[1]This is a draft version and may contain some minor errors. The full version is copyright and will be available from the IEEE following the CVPR 2011 conference.

sisting of image windows averaging 24×26 pixels in size representing putative head regions. For each image the tracker also provides an instantaneous ground plane velocity estimate for the corresponding pedestrian. In our two exemplar scenarios used to evaluate the idea, we have acquired datasets comprising 473412 images from 2258 people for the first scenario and 639581 images from 3861 people for the second.

The system is based around a Conditional Random Field (CRF) which models different aspects of gaze behaviour, such as the tendency for pedestrians to look in their direction of travel. When applied to a dataset, the CRF automatically infers the gaze directions for the images and as a side effect also trains a forest of randomised tree classifiers. The randomised forest models the interaction between the head images and the gaze direction variables within the CRF, but once trained can also be used as a standalone gaze classifier without the CRF.

The method for learning a classifier that we describe is weakly supervised when considered in isolation, however the weak supervision consists of the direction of motion from an unsupervised tracking system. When considered in combination with the tracking system, our approach is fully unsupervised. In the context of learning, our system has some similarities to that of Leistner et al [8], who used multiple instance learning with randomised trees to classify object images into categories.

Gaze estimation in visual surveillance is motivated by applications requiring inference of interactions between people [10] and frequently observed scene locations [9, 2]. Existing methods for coarse gaze estimation use manually labelled data to train various types of classifier such as Support Vector Machines [13, 7], Decision Trees [2, 1, 16], Neural Networks [17] and Nearest Neighbour classifiers [11]. In the majority of these approaches [13, 16, 17, 7], training images were obtained from the same scenes that were used for testing, which is advantageous but unrealistic unless methods are available for training without manual intervention.

Our system relies heavily on an accurate model of human gaze behaviour. A recent study of human gaze behaviour provided a set of hand labelled ground-plane velocities and gaze directions. We used this data, which we will refer to as the *model data*, to infer some of the parameters for our gaze behaviour model. These parameters are expected to generalise to any video sequence so the same values were used when testing on both datasets.

2. Model Formulation

The observations used as input to our system consist of a set of image sequences $I = \{\mathbf{i}_x\}$ where every image has a corresponding movement direction and magnitude v^t representing the individual's ground plane velocity. There are

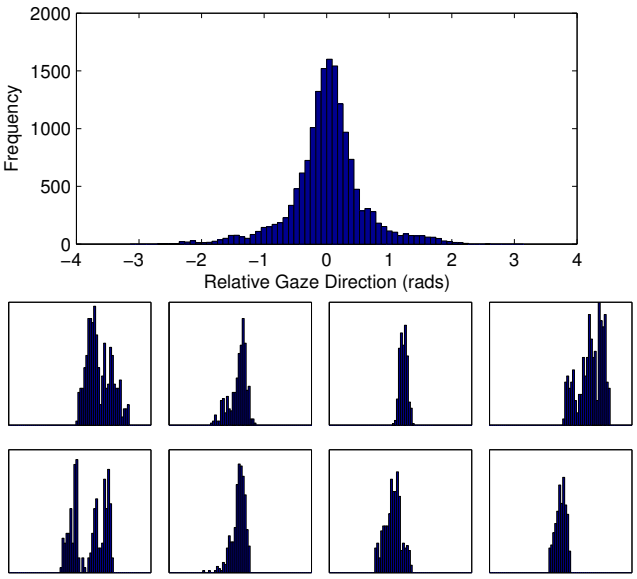

Figure 2. Distribution of gaze direction relative to walking direction over all people (top) and eight individual people (bottom).

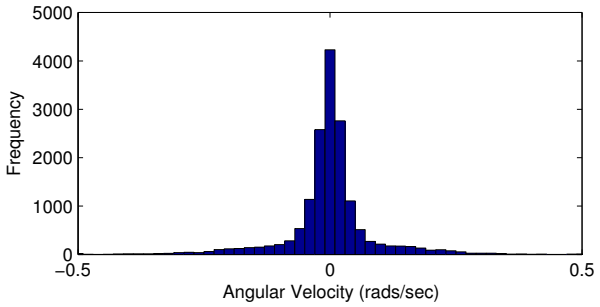

Figure 3. The distribution of head angular velocity relative to the walking direction over all of the people in the model data.

three key properties of the tracked images that we harness to infer the gaze directions. The first is that people tend to look most frequently in their direction of travel, an intuition which is confirmed by the model data shown in figure 2. The second property is that people usually move their heads slowly, so we expect sequential gaze directions to be reasonably similar. Again, this is confirmed by the plot of angular velocities from the model data shown in figure 3. Lastly, we expect the appearance of people to be more similar when they are looking in the same direction compared to when they are not.

The overall estimation is based around the optimisation of a CRF, the structure of which is represented as a factor graph in figure 4. Using the clique template representation of Sutton and McCallum [18], we divide the factors ψ_c into four sets $\mathcal{C} = \{C_T, C_F, C_\omega, C_I\}$ depending on which random variables they combine. The overall conditional prob-

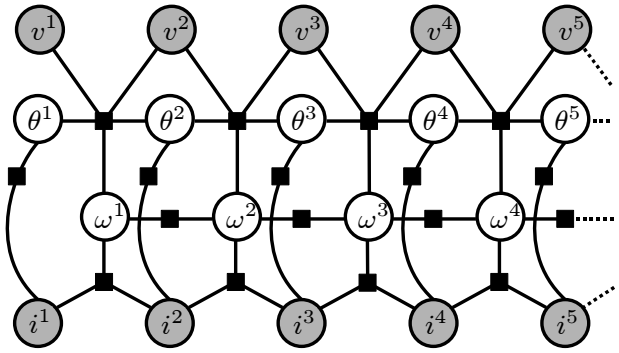

Figure 4. A factor graph showing how cliques and variables interact in the CRF. The angle estimate at time t is represented by θ_x^t, the angular velocity between times t and $t+1$ is represented by ω^t, and the observed image information and walking velocity are represented by i^t and v^t respectively.

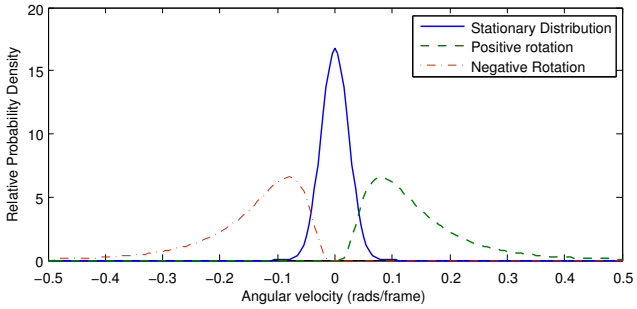

Figure 5. The three components of the mixture model used to represent the angular velocity.

ability is represented as the product of the individual factor functions:

$$p(\boldsymbol{\theta}, \boldsymbol{\omega} | \boldsymbol{i}, \boldsymbol{v}) = \frac{1}{Z(x)} \prod_{C_p \in \mathcal{C}} \prod_{\psi_c \in C_p} \psi_c(\boldsymbol{i}_c, \boldsymbol{v}_c, \boldsymbol{\theta}_c, \boldsymbol{\omega}_c) \quad (1)$$

The function $Z(x)$ represents a normalising constant which is required to ensure that the distribution sums to one given the observed image $\boldsymbol{i}$ and velocity $\boldsymbol{v}$. The labels, which consist of the estimated gaze directions θ^t and angular velocities ω^t, are both discretised to allow efficient inference. Gaze directions are represented as a distribution over 32 bins each representing an 11.25 degree range of angles.

The angular velocity ω^t is represented as a vector of three weights $\boldsymbol{\omega}^t = (\omega^+, \omega^0, \omega^-)^\mathsf{T}$ which correspond to the probability of the angular velocity being represented by each of three components, which are shown in figure 5. The first component is a Gaussian to represent the peak in the centre of the distribution corresponding to no head rotation. The other two components are log-Gaussians to represent rotations in the positive and negative directions. We determined the parameters for these three components from the

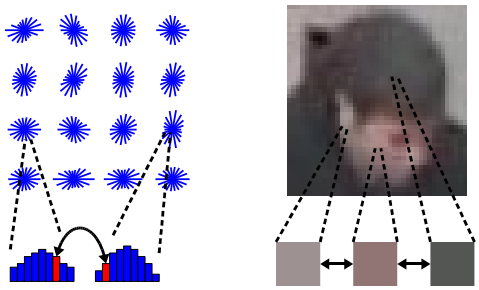

Figure 6. The two feature types that were used in the randomised tree classifiers. The first feature type (left) is determined by comparing two bins from normalised HOG descriptors and the second feature type (right) compares three different colour samples

model data using Expectation Maximisation.

The following sections will define how the four types of factor function model the interactions between random variables.

2.1. Angular Velocity Factors

We begin by defining the factor representing transitions between angular velocities. The parameters consist of a single matrix A representing the probability of transitioning from one angular velocity state to another:

$$\psi_c(\omega^t, \omega^{t+1}) \propto P(\omega^{t+1} | \omega^t) \quad (2)$$

$$\propto (\boldsymbol{\omega}^t)^\mathsf{T} A \boldsymbol{\omega}^{t+1} \quad (3)$$

The angular acceleration matrix A was estimated from the model data.

2.2. Image Classification Factors

The next factor is that which relates head images to directions. Initially we do not have any information on the mapping between images and gaze directions, however we do know that similar images are more likely to represent the same direction. This property is modelled using a forest of randomised tree classifiers, with each leaf node containing a histogram over the 32 direction bins, initially representing a uniform distribution. The histograms are scene-specific parameters that are inferred during the automatic learning process.

Randomised trees were constructed with two types of decisions and were trained by splitting branches until there were fewer than one hundred examples in each leaf node. The branch decisions were selected from the two types used in our past work [2], which are illustrated in figure 6 and recapitulated briefly below. Since the true labels for images were not known in advance, the decisions were selected at random for each branch. We make all decisions at the same depth equal, which gives our trees some of the performance advantages of ferns [19, 14]. Since trees only expand leaves

containing data, we retain the advantages of trees in terms of storage requirements, which allows training to a greater depth than would be possible with ferns.

The first type of decision compares randomly chosen bins from Histograms of Oriented Gradients (HOGs) with spatial normalisation as used by Dalal and Triggs [4] for their pedestrian detector. Descriptors consist of gradient histograms constructed from images that have been divided into a 4×4 grid of cells and normalised spatially across 2×2 blocks of cells.

The second type of decision is the Colour Triplet Comparison (CTC) which samples colours from pixels at three different locations in the head image and makes a binary decision based on whether the first and second colours are more similar than the second and third colours. Similarity was measured as the sum of the differences in each of the RGB components (i.e. the L1 Norm of the vector difference).

A forest of forty randomised trees was used to represent the angle distributions, with the factor function averaging over all of the outputs to obtain the required probability estimate:

$$\psi_c(\theta^t, i^t) = \frac{1}{n}\sum_{k=1}^{n} P(\theta^t | D^k(i^t)) \tag{4}$$

The notation $D^k(i^t)$ is used to represent the branch decision outcomes from passing i^t down the kth tree in the forest, from which the conditional probability is estimated using the histogram at the corresponding leaf.

Randomised tree classifiers were chosen because they require very little time to retrain if the image data remains the same and only the corresponding class distributions change.

2.3. Changing Image Factors

The next type of factor is intended to represent the correlation between the gaze direction changing and the observed image changing. The distance between two head images $d(i^t, i^{t+1})$ is measured as the number of randomised trees in which the two images reached different leaves, if δ represents the Kronecker delta function then this is defined as:

$$d(i^t, i^{t+1}) = \sum_{k=1}^{n} \delta(D^k(i^t), D^k(i^{t+1})) \tag{5}$$

The vector ϕ of length $n+1$ represents the probability of the head being stationary for each possible number of different bins, resulting in the following factor definition:

$$\psi_c(\omega^t, i^t, i^{t+1}) = P(\omega^t | i^t, i^{t+1}) \tag{6}$$

$$\propto \omega^0 \phi_{d(i^t, i^{t+1})} + (1 - \omega^0)\frac{1 - \phi_{d(i^t, i^{t+1})}}{2} \tag{7}$$

The elements of ϕ are scene-specific parameters that are learned automatically and ω^0 is the element of $\boldsymbol{\omega}^t$ representing the probability of the ω^t being represented by the stationary component.

2.4. Head Motion Factors

Lastly we describe the factors C_T which cover the transitions between pairs of gaze directions. The factor function is defined in terms of a prior transition matrix T and a matrix of angular velocity marginals M:

$$\psi_c(\theta^t, \theta^{t+1}, v^t, v^{t+1}, \omega^t) \propto P(\theta^{t+1} | \theta^t, v^t, v^{t+1}, \omega^t) \tag{8}$$

$$\propto (\theta^{*t})^{\mathsf{T}} (T \oplus M) \theta^{*t+1} \tag{9}$$

The operator $\oplus$ is used here to denote the element-wise product of two matrices and θ^{*t} represents the vector θ^t rotated so that it is measured relative to v^t. M is a cyclic matrix where elements m_{ij} represent the probability of the angular velocity required to transition from state i to state j given the estimated angular velocity parameters ω^t.

The matrix T represents our prior knowledge of how the gaze direction should change between each pair of frames. Many of the elements in T could be learned from the observations, however some of the elements represent transitions that are expected to occur very infrequently, such as rapid head movements or transitions between backward facing directions. These elements would be impractical to estimate empirically. To avoid estimating all 32^2 elements of T directly, the number of degrees of freedom in was reduced by parameterising T, as described in section 2.4.1, before fitting to the transitions in the model data using a constrained optimisation.

In the event that either v^t or v^{t+1} has a magnitude of less than half the mean human walking speed ($0.7ms^{-1}$), the walking direction is considered to be unreliable and the head motion factor is evaluated in the absolute frame of reference. This differs from equation 9 in that θ^t and θ^{t+1} are used rather than the relative versions and a small modification is made to the steady state for T, which will be described in the next section.

2.4.1 Transition Matrix Parameterisation

From any state, we represent the probability of travelling to a destination state as a mixture $\mathbf{z}$ of the three components illustrated in figure 5. It is these mixture coefficients on the three components of the velocity distribution that constitute the parameterisation.

When an individual is facing their direction of travel, there is an equal probability that they will move their gaze direction to the left or to the right, however if they are already looking sideways it is more likely that they will move their gaze direction towards the direction of travel. This observation has been previously used to aid high resolution

head tracking [5]. We model this by allowing the weights for the three components to vary depending on the current angle of the head, so the parameterisation consists of three vectors of n weights $\mathbf{z}^+$, $\mathbf{z}^0$ and $\mathbf{z}^-$ corresponding to positive, stationary and negative rotation probabilities respectively. Each of the three components were quantised to give a probability distribution q over discrete states so the members of T could be easily calculated efficiently:

$$t_{ij} = z_i^+ q_{j-i \pmod n}^+ + z_i^- q_{j-i \pmod n}^- + z_i^0 q_{j-i \pmod n}^0 \quad (10)$$

The values for the weights are optimised to ensure that they represent a transition matrix that is consistent with both the prior angular velocity mixture weights ω_{pr}^t and some general properties of head motion, both of which were learned from the hand labelled motion data. Specifically, the following constraints must be met for a transition matrix to be valid:

Total Weight To ensure that the probability mass represented by the velocity distribution sums to one, each set of three component weights must total one:

$$C_i^\Sigma(\mathbf{z}) = z_i^+ + z_i^0 + z_i^- - 1 = 0 \quad (11)$$

Positive Weights To prevent negative transition probabilities, inequality constraints must be added to ensure that each of the $3n$ component weights are not negative:

$$C_i^{p^+}(\mathbf{z}) = \max(-z_i^+, 0) = 0 \quad (12)$$

$$C_i^{p^0}(\mathbf{z}) = \max(-z_i^0, 0) = 0 \quad (13)$$

$$C_i^{p^-}(\mathbf{z}) = \max(-z_i^-, 0) = 0 \quad (14)$$

Steady State The gaze directions of pedestrians are highly correlated with their walking directions, since people tend to look in their direction of travel and rarely look behind themselves. The gaze distribution of an individual observed for long enough should tend towards the distribution s shown in figure 2. We expect the transition matrix to satisfy this condition if s is its steady state:

$$T^\mathsf{T} s = s \quad (15)$$

To enforce the steady state distribution, constraints were introduced with the following form:

$$C_i^S(\mathbf{z}) = \frac{\sum_j t_{ji} s_j}{s_i} - 1 = 0 \quad (16)$$

If an individual is not walking and the motion factor is evaluated in the absolute frame of reference, s is set to be uniform to reflect the lack of prior knowledge on potential gaze directions.

Objective Function The requirements above constrain $2n$ degrees of freedom, however $\mathbf{z}$ has $3n$ variables. A constraint on the relative values of z_i^0 could be introduced to provide an additional $n-1$ constraints, but doing so would make it often impossible to satisfy all of the constraints. Instead, an objective function was used to regularise the solution by imposing the preference of having the stationary weights equal to the angular velocity mixture weight priors:

$$f(\mathbf{z}) = -\sum_i (z_i^- - \omega_{pr}^-)^2 + (z_i^0 - \omega_{pr}^0)^2 + (z_i^+ - \omega_{pr}^+)^2 \quad (17)$$

2.4.2 Transition Matrix Optimisation

To find the optimal parameter values for the matrix T, the constraints were combined with the objective function using the quadratic penalty method [12]. Every constraint introduces a penalty term which is zero when satisfied.

$$F(\mathbf{z};\mu) = f(\mathbf{z}) - \frac{1}{2\mu}\sum_i \Big(C_i^\Sigma(\mathbf{z})^2 + C_i^S(\mathbf{z})^2 + C_i^{p^+}(\mathbf{z})^2 + C_i^{p^0}(\mathbf{z})^2 + C_i^{p^-}(\mathbf{z})^2 \Big) \quad (18)$$

Although the constraints have the same quadratic form as the objective function, the parameter μ is reduced during the optimisation to make the penalty functions dominate over the objective function to ensure that the constraints are satisfied. The objective function has discontinuous second derivatives resulting from the inequality constraints, so nonlinear conjugate gradients was used to perform the optimisation instead of a Newton-based method.

3. Model Optimisation

Having described the structure of the individual factor functions, we now consider the optimisation of θ and ω, which we consider to be latent variables for the purpose of learning the scene parameters. The learning uses the Expectation Maximisation algorithm, which alternates between calculating an expectation over the latent variables and maximising the parameters given the expectation.

Since our CRF contains cycles the expectation cannot be calculated exactly in any reasonable amount of time, so we approximate it using Loopy Belief Propagation (LBP), an extension of Belief Propagation [15] to graphs with cycles. A detailed description of probabilistic LBP for factor graphs in terms of the general sum-product algorithm is described by Kschischang et. al. [6]. Since our CRF is chain-structured, we use a message passing schedule which propagates messages in alternating forwards and backwards passes in a similar way to the Forwards-Backwards algorithm for Hidden Markov Models.

The two sets of parameters to be estimated in the maximisation step are the probability histograms in the leaves

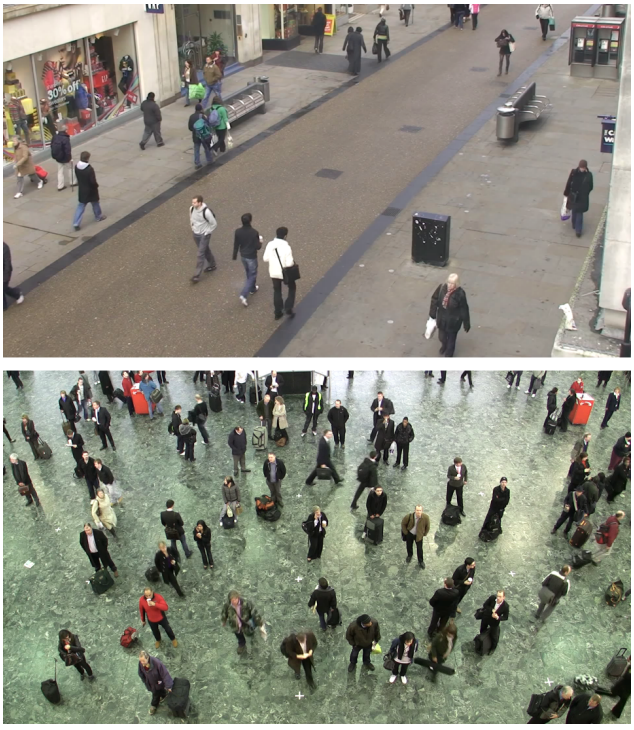

Figure 7. Sample frames from the two scenes on which the system was tested. In the first scene, most people are moving but in the second scene most people are stationary. The second scene also exhibits significant distortion, which we train for implicitly.

of the randomised trees and the vector ϕ of motion probabilities from the image differences factors. The leaf histograms are maximised when they are equal to the mean of the expected gaze directions for all of the images reaching the leaf, which is the standard training process for trees. The motion probabilities are maximised by marginalising over the factors for all pairs of images in the dataset.

4. Evaluation

The system was evaluated on the tracking output from two large video sequences of different scenes, shown in figure 7. Both scenes are public places where pedestrians exhibit a wide variety of appearances due to hats, sunglasses and different clothing. The mean size of the head images in both cases was 24×26 pixels.

The first dataset is from an outdoor town centre scene where the majority of pedestrians are walking and consists of 473412 images from 2259 people. Every one hundredth image in the dataset was hand labelled to provide ground truth, a total of 4347 images. This sequence is publicly available and we will make our images with ground truth available to enable future comparisons. The second dataset covers a busy transport terminal where the majority of people are stationary and consists of 639581 images from 3861 people.

	Test Dataset	
Method	Town Centre	Terminal
Our Unsupervised System	23.9	38.5
Supervised Ferns [2]	45.5	59.2
Walking Direction	25.9	78.4

Table 1. Gaze estimation performance (MAAE in degrees) of our unsupervised system compared with that of conventionally trained classifiers and the walking direction baseline. Our system outperforms the other approaches on both datasets.

The performance of the system was measured using the Mean Absolute Angular Error (MAAE), which is stated in degrees. A baseline performance measure was obtained by assuming that the gaze direction is the same as the walking direction. This baseline provides very good estimates for the Town Centre dataset, since most people look in their direction of travel, however in the Transport Terminal dataset there are many stationary people so the direction of travel is often incorrect.

The system was also compared with a classifier that we developed in our previous work [2] which was based on randomised ferns and was trained using approximately 1500 head images that were cropped from still photos of different people that were acquired from other datasets and internet image searches.

The results of applying the system to the two large video datasets are shown in table 1. In both datasets, our system significantly outperforms the supervised ferns, which demonstrates the value of learning the scene-specific classifier. Since most people in the Town Centre dataset look in their direction of travel, there is not much to be learned, so we only marginally outperform the walking direction baseline. In the Transport Terminal dataset there are many stationary people so our system performs considerably better than the walking direction baseline, which is almost random.

Although one of the benefits of the unsupervised learning approach is the ability to learn scene-specific classifiers, to provide some additional insight the learned randomised forests were tested in isolation (without the CRF) on each of the two video datasets as well as the still images that were used to train the supervised classifier. It should be noted that this is not the intended usage of the system, since for any practical purpose the combination of the CRF model and the learned trees would be used rather than just the trees alone.

The results from testing the randomised forest in isolation are shown in table 2. An important conclusion from these results is that we can use the walking direction to learn a randomised forest classifier, but the classifier remains effective even when the walking direction is not used in the

Method	Training Dataset	Testing Dataset		
		Town Centre	Transport Terminal	Still Images
Unsupervised Trees	Town Centre	25.6	47.8	60.2
Unsupervised Trees	Terminal	64.9	42.9	71.4
Supervised Ferns [2]	Still Images	45.5	59.2	43.5

Table 2. Comparison of the performance (MAAE in degrees) of the learned forest of randomised tree classifiers when trained and tested on the two video datasets and the still image dataset. For these experiments the CRF model was not used for testing. When the supervised ferns were tested on the still image dataset, 80% of the data was used for training and the remaining 20% for testing. The results show that the learned classifiers still outperform the supervised ferns even in the absence of motion information.

estimation at testing time.

In the absence of the CRF model, the randomised forests that were learned from the Town Centre and Transport Terminal datasets each perform best on the dataset from which they were learned, however the Town Centre forest appears to generalise better and is more accurate when tested on the still image dataset. A likely reason for this is that although both video datasets are large, most of the information for learning the classifiers comes from people who are walking. The Town Centre dataset has 441048 images from moving people, compared to only 69512 images from moving people in the Transport Terminal dataset. Many of these images will be similar, since we acquire around 200 images from each person, making approximately 420 different informative people in the Transport Terminal dataset in comparison to approximately 2100 in the Town Centre dataset. It is likely that this variation in appearances during training is responsible for the better performance of the Town Centre classifier, however this is not a limitation of the approach in general because a deployed system would be able to constantly improve the classifier over many days or years.

Many of the errors in the Transport Terminal dataset were caused because people often walk backwards to improve their view of the departure board. Since the data that was used to learn the model parameters did not include situations like this, our model considers this to be almost impossible. The result is that an incorrect gaze estimate (usually the direction of motion) is fed back into the randomised forest, where it has a negative affect on the estimations for other people on subsequent iterations. This issue could be resolved either by using training data from a wider variety of scenes to learn the model parameters, or by modifying the tracker to detect abnormal walking patterns so that they could be omitted.

Sample observation sequences and smoothed probabilities for two individuals are shown in figure 8.

5. Conclusion

We have developed a system which is capable of learning a gaze classifier from surveillance video without any human intervention. Our evaluation has shown that the performance of these scene-specific classifiers exceeds that of a conventionally trained classifier on the scene where they were learned.

Our implementation is able to learn classifiers from batches of data in approximately double the amount of time that is required to generate it, so a real-time online implementation is feasible. An online implementation would allow a potentially unlimited amount of training data to be used, which would almost certainly result in improved performance.

If the system were installed in a number of different locations, the results from the learning could be automatically shared with other systems to improve performance, allowing far more training data to be incorporated than would otherwise be possible. The learned classifier from multiple forests of decision trees could be easily combined if the same set of random decisions were chosen across all installations.

Acknowledgements

The authors would like to thank Joe Hale and Terry Thomson for collaborating with the collection and labelling of the datasets, Eric Sommerlade for many fruitful discussions, and Oxford Risk for financial support.

References

[1] B. Benfold and I. Reid. Colour invariant head pose classification in low resolution video. In *Proceedings of the 19th British Machine Vision Conference*, September 2008.

[2] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, September 2009.

[3] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, June 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, June 2005.

[5] J. Heuring and D. Murray. Modeling and copying human head movements. *Robotics and Automation, IEEE Transactions on*, 15(6):1095 –1108, Dec. 1999.

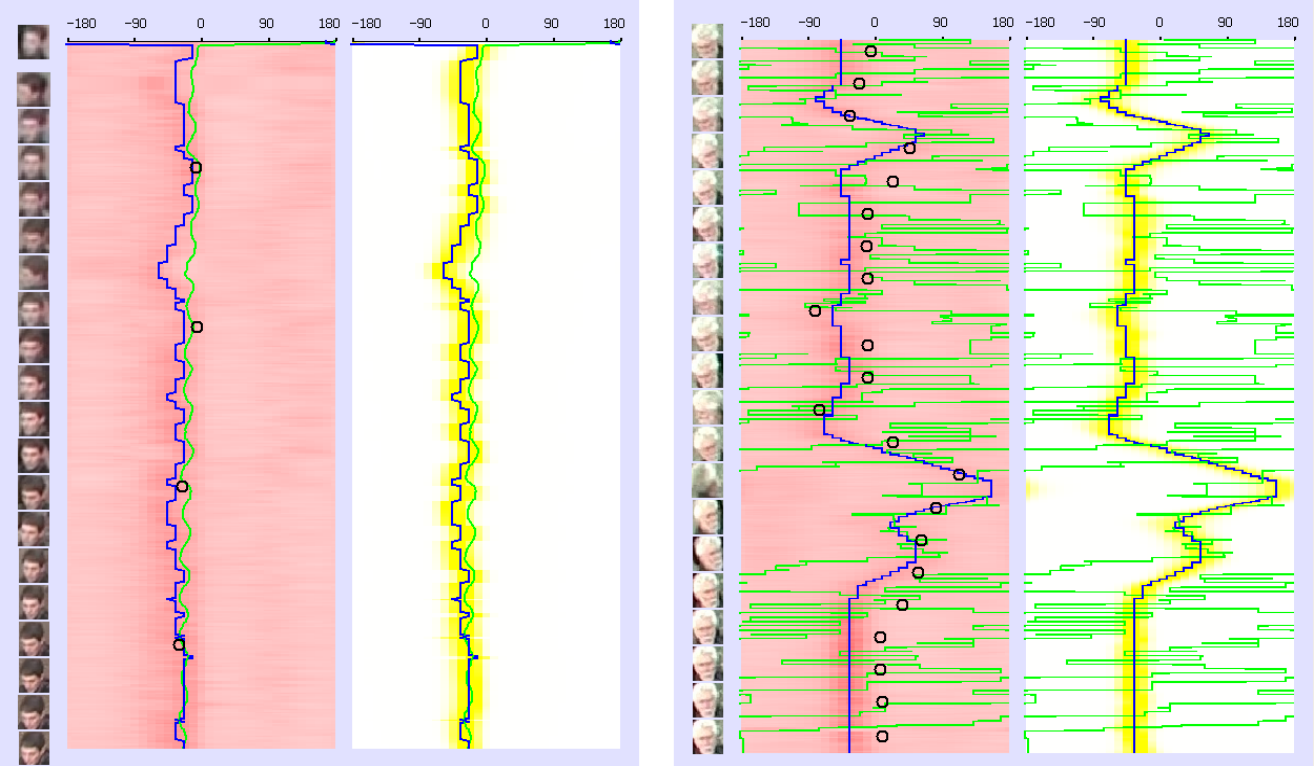

Figure 8. Two example sequences showing how head movements are correctly identified. The horizontal axis corresponds to the 360° range of gaze directions, with the centre representing 0 (looking at the camera). The vertical axis represents time, with the first frame in the sequence at the top and the last frame at the bottom. The red backgrounds show the observation probability (red is high) and yellow backgrounds show smoothed probabilities from the latent variables (yellow is high). In both images the green line is the direction in which the person is walking, blue is maximum of the marginal distribution and circles represent ground truth labels. The left sequence shows a pedestrian who is moving, so the walking direction provides a strong prior for the gaze direction. The right sequence shows a pedestrian who is standing still, so the direction of motion is not helpful, but the learned randomised forest ensures that the estimation is still reasonably accurate.

[6] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498 –519, Feb. 2001.

[7] A. Launila and J. Sullivan. Contextual features for head pose estimation in football games. In *ICPR*, pages 340–343. IEEE, 2010.

[8] C. Leistner, A. Saffari, and H. Bischof. Miforests: Multiple-instance learning with randomized trees. In *ECCV (6)*, volume 6316 of *LNCS*, pages 29–42. Springer, 2010.

[9] X. Liu, N. Krahnstoever, T. Yu, and P. H. Tu. What are customers looking at? In *AVSS*, pages 405–410. IEEE Computer Society, 2007.

[10] A. T. M. Farenzena, L. Bazzani, D. Tosato, G.Paggetti, G. Menegaz, V. Murino, and M.Cristani. Social interactions by visual focus of attention in a three-dimensional environment. In *PRAI-HBA*, December 2009.

[11] S. Niyogi and W. T. Freeman. Example-based head tracking. In *FG*, pages 374–378. IEEE Computer Society, 1996.

[12] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, USA, August 2000.

[13] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *Proceedings of the 20th British Machine Vision Conference*, September 2009.

[14] M. Özuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):448–461, 2010.

[15] J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982.

[16] N. Robertson and I. D. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV (2)*, volume 3952 of *LNCS*, pages 402–415. Springer, 2006.

[17] R. Stiefelhagen. Estimating head pose with neural networks - results on the pointing04 icpr workshop evaluation data. In *Pointing '04 ICPR workshop*, August 2004.

[18] C. Sutton and A. McCallum. An introduction to conditional random fields. Nov. 2010.

[19] B. Williams, G. Klein, and I. Reid. Real-time SLAM relocalisation. In *Proc. International Conference on Computer Vision*, 2007.