Estimating gaze direction from low-resolution faces in video

Neil Robertson[1,2] and Ian Reid[2]

[1] QinetiQ, St Andrews Road, Malvern, WR14 3PS, UK
[2] Oxford University, Dept Engineering Science, Parks Road, Oxford, OX1 3PJ, UK
{nmr,ian}@robots.ox.ac.uk

Abstract. In this paper we describe a new method for automatically estimating where a person is looking in images where the head is typically in the range 20 to 40 pixels high. We use a feature vector based on skin detection to estimate the orientation of the head, which is discretised into 8 different orientations, relative to the camera. A fast sampling method returns a distribution over previously-seen head-poses. The overall body pose relative to the camera frame is approximated using the velocity of the body, obtained via automatically-initiated colour-based tracking in the image sequence. We show that, by combining direction and head-pose information gaze is determined more robustly than using each feature alone. We demonstrate this technique on surveillance and sports footage.

1 Introduction

In applications where human activity is under observation, be that CCTV surveillance or sports footage, knowledge about where a person is looking (i.e. their gaze) provides observers with important clues which enable accurate explanation of the scene activity. It is possible, for example, for a human readily to distinguish between two people walking side-by-side but who are not "together" and those who are acting as a pair. Such a distinction is possible when there is regular eye-contact or head-turning in the direction of the other person. In soccer head position is a guide to where the ball will be passed next i.e. an indicator of *intention*, which is essential for causal reasoning. In this paper we present a new method for automatically inferring gaze direction in images where any one person represents only a small proportion (the head ranges from 20 t0 40 pixels high) of the frame.

The first component of our system is a descriptor based on skin colour. This descriptor is extracted for each head in a large training database and labelled with one of 8 distinct head poses. This labelled database can be queried to find either a nearest-neighbour match for a previously unseen descriptor or (as we discuss later) is non-parametrically sampled to provide an approximation to a distribution over possible head poses.

Recognising that general body direction plays an important rôle in determining where a person can look (due to anatomical limitations), we combine direction and head pose using Bayes' rule to obtain the joint distribution over

head pose and direction, resulting in 64 possible gazes (since head pose and direction are discretised into 8 sectors each, shown in figure 1).

The paper is organised as follows. Firstly we highlight relevant work in this, and associated, area(s). We then describe how head-pose is estimated in section 2. In section 3 we provide motivation for a Bayesian fusion method by showing intermediate results where the best head-pose match is chosen and, by contrast, where direction alone is used. Section 3 also discusses how we fuse the relevant information we have at our disposal robustly to compute a distribution over possible gazes, rejecting non-physical gazes and reliably detecting potentially significant interactions. Throughout the paper we test and evaluate on a number of datasets and additionally summarise comprehensive results in section 4. We conclude in section 5 and discuss potential future work in section 6.

1.1 Previous work

Determining gaze in surveillance images is a challenging problem that has received little or no attention to date, though preliminary work in this specific problem domain was reported in [23].

Most closely related to our work is that of Efros *et al* [6] for recognition of human action at a distance. That work showed how to distinguish between human activities such as walking, running etc. by comparing gross properties of motion using a descriptor derived from frame-to-frame optic-flow and performing an exhaustive search over extensive exemplar data. Head pose is not discussed in [6] but the use of a simple descriptor invariant to lighting and clothing is of direct relevance to head pose estimation and has directly inspired aspects of our approach.

Dee and Hogg [5] developed a system for detecting unusual activity which involves inferring which regions of the scene are visible to an agent within the scene. A Markov Chain with penalties associated with state transitions is used to return a score for observed trajectories which essentially encodes how directly a person made his/her way towards predefined goals, typically scene exits. In their work, gaze inference is vital, but is inferred from trajectory information alone which can lead to significant interactions being overlooked. In fact, many systems have been created to aid urban surveillance, most based on the notion of trajectories alone. For example [9] reports an entirely automated system for visual surveillance and monitoring of an urban site using agent trajectories. The same is true in the work of Buxton (who has been prominent in the use of Bayesian networks for visual surveillance) [2], Morellas *et al* [18] and Makris [15]. Johnson and Hogg's work [12] is another example where trajectory information is only considered.

In contrast, there has been considerable effort to extract gaze from relatively high-resolution faces, motivated by the press for better Human/Computer Interfaces. The technical aspects of this work have often focused on detecting the eyeball primarily. Matsumoto [16] computes 3-D head pose from 2-D features and stereo tracking. Perez et al. [21] focus exclusively on the tracking of the

eyeball and determination of its observed radius and orientation for gaze recognition. Kaminski et al. [13] have achieved a very similar goal but using a single image while retaining a face and eye model. Gee and Cipolla's [8] gaze determination method based on the 3D geometric relationship between facial features was applied to paintings to determine where the subject is looking. Related work has tackled expression recognition using information measures. Shinohara and Otsu demonstrated that Fisher Weights can be used to recognise "smiling" in images.

While this approach is most useful in HCI where the head dominates the image and the eye orientation is the only cue to intention, it is too fine-grained for surveillance video where we must usually be content to assume that the gaze direction is aligned with the head-pose. In typical images of interest in our application area (low/medium resolution), locating significant features such as the eyes, irises, corners of the mouth, etc as used in much of the work above is regularly an impossible task. Furthermore, though standard head/face-detection techniques [25] work well in medium reolution images, they are much less reliable for detecting, say, the back of a head, which still conveys significant gaze information.

The lowest level of our approach is based on skin detection. Because of significant interest in detecting and tracking people in images and video, skin detection has naturally received much attention in the Computer Vision community [3] [10] [11]. However skin detection alone is error-prone when the skin region is very small as a proportion of the image. However, contextual cues such as direction can help to disambiguate gaze using even a very coarse head-pose estimation. By combining this information in a principled (i.e. probabilistic, Bayesian) fashion, gaze estimation at a distance becomes a distinct possibility as we demonstrate in this paper.

2 Head pose detection

2.1 Head pose feature vector

Although people differ in colour and length of hair and some people may be wearing hats, beards etc. it is reasonable to assume that the amount of skin that can be seen, the position of the skin pixels within the frame and the proportion of skin to non-skin pixels is a relatively invariant cue for a person's coarse gaze in a static image. We obtain this descriptor in a robust and automatic fashion as follows. First, a mean-shift tracker [4] is automatically initialised on the head by using naive background subtraction to locate people and subsequently modelling the person as distinct "blocks", the head and torso. Second, we centre the head within the tracker window at each time step which stabilises the descriptor ensuring consistent position within the frame for similar descriptors (the head images are scaled to the same size and, since the mean-shift tracker tracks in scale-space we have a stable, invariant, descriptor). Third, despite claims in the literature to the contrary, there is no specific region of colour-space which represents skin in all sequences and therefore it is necessary to define a skin histogram

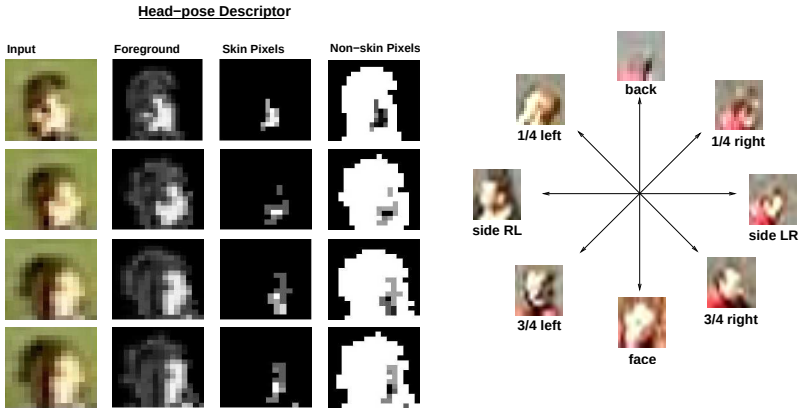

Fig. 1. The figure on the left shows the images which result from the mean-shift image patch tracker (*col. 1*) (with an additional step stabilise the descriptor by centering the head in the window), subsequent background subtraction (*col. 2*), the weight image which represents the probability that each pixel in the head is skin (*col. 3*) and non-skin (*col. 4*) (non-skin is significant as it captures proportion without the need for scaling). Thie concatenation of skin and non-skin weight vectors is our feature vector which we use to determine eight distinct head poses which are shown and labelled on the right. Varying lighting conditions are accounted for by representing the same head-pose under light from different directions in the training set. The same points on the "compass" are used as our discretisation of direction i.e. N, NE, E, etc.

for each scenario by hand-selecting a region of one frame in the current sequence to compute a (normalised) skin-colour histogram in RGB-space. We then compute the weights for every pixel in the stabilised head images which the tracker automatically produces to indicate how likely it is that it was drawn from this predefined skin histogram[3]. Using the knowledge of the background we segment the foreground out of the tracked images. Every pixel in the segmented head image is drawn from a specific RGB bin and so is assigned the relevant weight which can be interpreted as a probability that the pixel is drawn from the skin model histograms. So for every bin i (typically we use 10 bins) in the predefined, hand-selected skin-colour histogram q the histogram of the tracked image p is a weight is computed $w_i = \sqrt{\frac{q_i}{p_i}}$. Every foreground pixel in the tracked frame falls into one of the bins according to its RGB value and the normalised weight associated with that pixel is assigned to compute the overall weight image, as shown in figure 1. The non-skin pixels are assigned a weigh that the pixel is *not* drawn from the skin histogram. This non-skin descriptor is necessary because it encodes the "proportion" of the head which is skin which is essential as people vary in size not only in the sense of scale within the but physically between

[3] This will be recognised as a similar approximation to the Battacharyya coefficient as implemented in the meanshift algorithm [4].

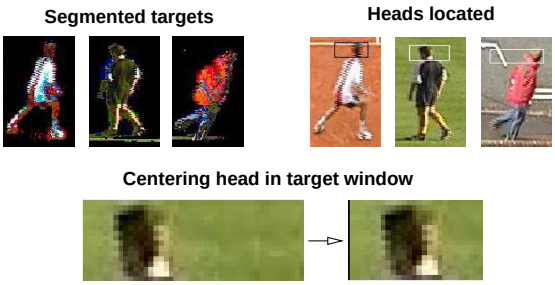

Fig. 2. Automatic location of the head is achieved by segmenting the target using simple background subtraction (*top-left*) and morphological operations with a kernel biased towards the scale of the target to identify objects. The head is taken as the top 1/7th of the entire body (*top-right*). The head is automatically centred in the bounding box at each time step to stabilise the tracking and provide an invariant descriptor for head pose, as shown in the second row.

one another. Each descriptor is scaled to a standard 20×20 pixel window to achieve robust comparison when the head sizes vary. Finally, in order to provide temporal context to our descriptor of head-pose we concatenate individual descriptors from 5 consecutive frames of tracker data for a particular example and this defines our instantaneous descriptor of head-pose.

2.2 Training data

We assume that we can distinguish head pose to a resolution of 45 degrees. There is no obvious benefit to detecting head orientations at a higher degree of accuracy and it is unlikely that the coarse target images would be amenable in any case. This means discretising the 360 degrees orientation-space into 8 distinct views as shown in figure 1. The training data we select is from a surveillance-style camera position and around 100 examples of each view are selected from across a number of different sequences and under different lighting conditions (i. e. light from left, right and above). The head was automatically tracked as described above and the example sequence labelled accordingly. The weight image for 5 consecutive frames are then computed and this feature vector stored in our exemplar set. The same example set is used in all the experiments reported (e.g. there are no footballers in the training dataset used to compute the gaze estimates presented in figure 9).

2.3 Matching head poses

The descriptors for each head pose are $(20 \times 20 \times 5 =)2000$ element vectors. With 8 possible orientations and 100 examples of each orientation searching this dataset rapidly becomes an issue. We elect to structure the database using a binary-tree in which each node in the tree divides the set of exemplars below

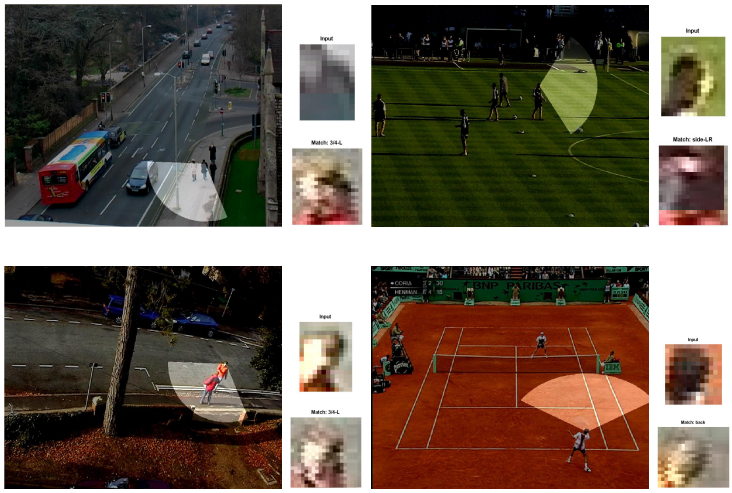

Fig. 3. Detecting head pose in different scenes using the same exemplar set. The main image shows the frame with the estimated gaze angle superimposed, the pair of images directly beside each frame shows the input image that the head-pose detector uses (*top*) and the best (ML) match in the database with corresponding label (*bottom*).

the node into roughly equal halves. Such a structure can be searched in roughly $\log n$ time to give an approximate nearest-neighbour result. We do this for two reasons: first, even for a modest database of 800 examples such as ours it *is* faster by a factor of 10; second, we wish to frame the problem of gaze detection in a probabilistic way and Sidenbladh [24] showed how to formulate a binary tree (based on the sign of the Principal Components of the data) search in a pseudo-probabilistic manner. This technique was later applied to probabilistic analysis of human activity in [22]. We achieve recognition rates of 80% (the correct example is chosen as the ML model 8/10 queries) using this pseudo-probabilistic method based on Principal Components with 10 samples. An illustrative example of such a distribution in this context is shown in figure 4. Results of sampling from this database for a number of different scenes are shown in figure 3. In order to display where the person is looking in the images angles are assigned to the discretised head-poses shown in figure 1 according to the "compass" e.g. $N : 0^o$ etc. The angles are then corrected for the projection of the camera at each time step (depending on the location of the person on the ground-plane in the image) as defined in figure 5.

3 Gaze estimation

3.1 Bayesian fusion of head-pose and direction

The naive assumption that direction of motion information is a good guide as to what a person can see has been used in figure 6. However, it is clear the

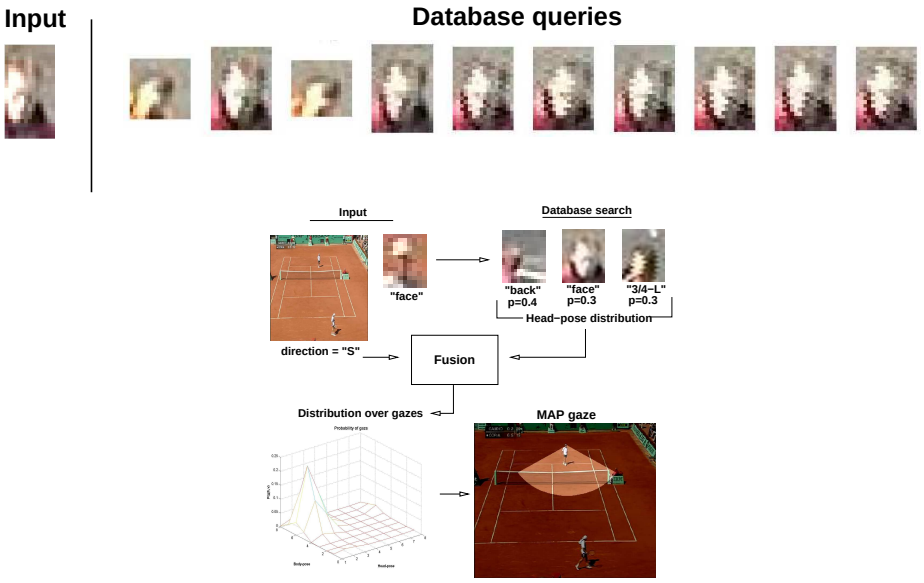

Fig. 4. (*Left*) The distribution over head-poses resulting from 10 queries of the database for this input frame is shown in the graph above. The leaf nodes of the database contain indices into matching frames and the matching frame images and assigned probabilities of a match are shown below the graph. (*Right*) Fusing head-pose and direction estimates improves gaze estimation. Here, the ML match for head pose would be incorrectly chosen as "back". The body-direction is identified as "S" which, since it is not possible to turn the head through 180^o relative to the body, this gaze has a low (predefined) prior and is rejected as the most likely at the fusion stage. The MAP gaze is identified as "Face" which is a very good approximation to the true gaze.

crucial interaction between the two people is missed. To address this issue we compute the joint posterior distribution over direction of motion and head pose. The priors on these are initially uniform for direction of motion, reflecting the fact that for these purposes there is no preference for any particular direction in the scene, and for head pose a centred, weighted function that models a strong preference for looking forwards rather than sideways. The prior on gaze is defined using a table which lists expected (i.e. physically possible) gazes and unexpected (i.e. non-physical) gazes.

We define g as the measurement of head-pose, d is the measurement of body motion direction, G is the true gaze direction and B is the true body direction, with all quantities referred to the ground centre. We compute the joint probability of true body pose and true gaze:

$$P(B, G|d, g) \propto P(d, g|B, G)P(B, G) \tag{1}$$

Now given that the measurement of direction d is independent both true and measured gaze G, g once true body B pose is known, $P(d|B, G, g) = P(d|B)$ and

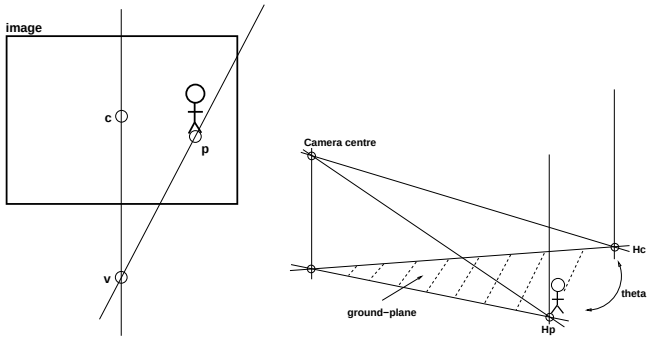

Fig. 5. When assigning angles to the matched discretised head-poses one must compensate for the camera projection since "North" (see figure 1) does not in general correspond to vertical in the image plane. In order to choose the correct frame of reference we do not perform full camera calibration but compute the projective transform (**H** : image→ground-plane) by hand-selecting 4 points in the image. The vertical vanishing point (**v**, *left*) is computed from 2 lines normal to the ground plane and parallel in the image. The angle *theta* between the projection of the optic-rays through the camera centre (**Hv**, *right*) and the image centre (**Hc**, *left*) and the point at the feet of the tracked person ((**Hp**, *right*) is the angle which adjusts vertical in the image to "North" in our ground plane reference frame i. e. $\cos^{-1}[(\mathbf{Hc} \times \mathbf{Hv}).(\mathbf{Hv} \times \mathbf{Hp})]$.

similarly that the measurement of gaze g is independent of true body pose B given true gaze G, $P(g|B,G) = p(g|G)$, then we have

$$P(B,G|d,g) \propto P(g|G)P(d|B)P(G|B)P(B) \tag{2}$$

We assume that the measurement errors in gaze and direction are unbiased and normally distributed around the respective true values

$$P(g|G) = \mathcal{N}(G,\sigma_G^2), P(d|B) = \mathcal{N}(B,\sigma_B^2) \tag{3}$$

(actually, since these are discrete variables we use a discrete approximation).

The joint prior, $P(B,G)$ is factored as above into $P(G|B)P(B)$ where the first term encodes our knowledge that people tend to look straight ahead (so the distribution $P(G|B)$ is peaked around B, while $P(B)$ is taken to be uniform, encoding our belief that all directions of body pose are equally likely, although this is easily changed: for example in tennis one player is expected to be predominantly facing the camera).

While for single frame estimation this formulation fuses our measurements with prior beliefs, when analysing video data we can further impose smoothness constraints to encode temporal coherence: the joint prior at time t is in this case taken to be $P(G_t,B_t|G_{t-1},B_{t-1}) = P(G_t|B_t,B_{t-1},G_{t-1})P(B_t|B_{t_1})$ where we have used an assumption that the current direction is independent of previous

gaze[4], and current gaze depends only on current pose and previous gaze. The former term, $P(G_t|B_t, B_{t-1}, G_{t-1})$, strikes a balance between between our belief that people tend to look where they are going, and temporal consistency of gaze via a mixture $G_t \sim \alpha\mathcal{N}(G_{t-1}, \sigma_G^2) + (1-\alpha)\mathcal{N}(B_t, \sigma_B^2)$.

Now we compute the joint distribution for all 64 possible gazes resulting from possible combinations of 8 head poses and 8 directions. This posterior distribution allows us to maintain probabilistic estimates without committing to a defined gaze which will be advantageous for further reasoning about overall scene behaviour. Immediately though we can see that gazes which we consider very unlikely given our prior knowledge of human biomechanics (since the head cannot turn beyond 90 degrees relative to the torso [20]) can be rejected in addition to the obvious benefit that the quality of lower-level match can be incorporated in a mathematically sound way. An illustrative example is shown in figure 4.

4 Results

We have tested this method on various datasets (see figures 6, 7, 8, 9) and 10. The first dataset provided us with the exemplar data for use on all the test videos shown in this paper. In the first example in figure 6 we show significant improvement over using head-pose or direction alone to compute gaze. The crucial interaction which conveys the information that the people in the scene are together is the frequent turning of the head to look at each other. We reliably detect this interaction as can be seen from the images and the estimated head angle relative to vertical. The second example is similar but in completely different scene. The skin histogram is recomputed for this video but the training data remains the same. Once more the interaction implied by the head turning to look at his companions is determined. We demonstrate the method on sports video in figure 9 and on a standard vision sequence in figure 10. It is shown in figure 7 how useful this technique can be in a causal-reasoning context where we identify two people looking at one another prior to meeting. Finally we discuss the failure mode in figure 11 which is found to be where the size of the head falls below 20 pixels and the gaze becomes ambiguous due to the small number of skin pixels.

5 Conclusions

In this paper we have demonstrated that a simple descriptor, readily computed from medium-scale video, can be used robustly to estimate head pose. In order to speed up non-parametric matching into an exemplar database and to maintain probabilistic estimates throughout we employed a fast pseudo-probabilistic

[4] though we recognise that this may in fact be a poor assumption in some cases since people may change their motion or pose in response to observing something interesting while gazing around

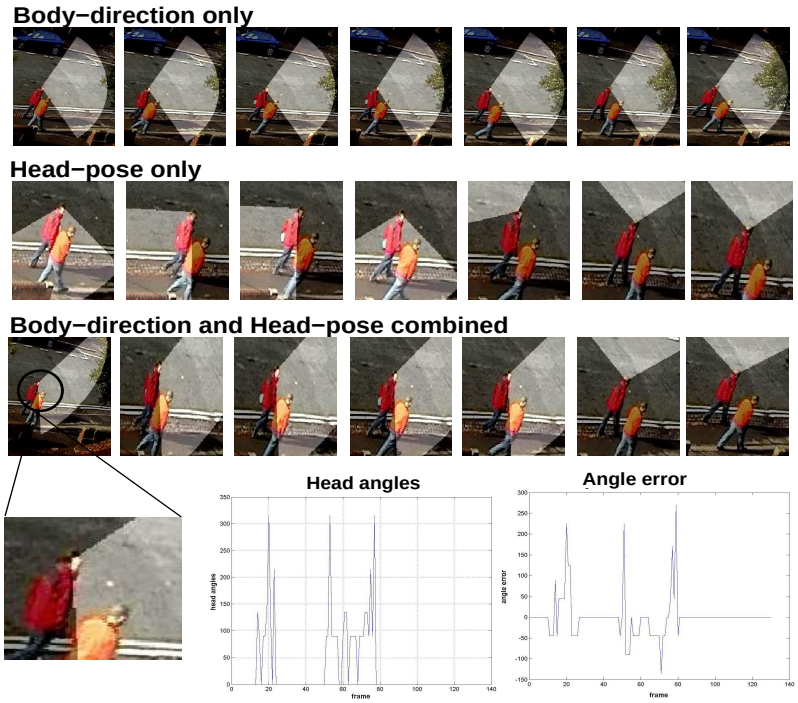

Fig. 6. In this video there is an interaction between the two people where the fact they look at each other the prime indicator that they are "together". On the first row we estimate gaze from body direction alone, on the second row using head-pose alone, which is improved but prone to some errors. We see that (*third row*) fusing the head-pose and body-direction estimates gives the correct result.

binary search based on Principal Components. To resolve ambiguity, improve matching and reject known implausible gaze estimates we used a simple application of Bayes' Rule to fuse priors on direction-of-motion and head-pose, evidence from our exemplar-matching algorithm and priors on gaze (which we specified in advance). We demonstrated on a number of different datasets that this gives acceptable gaze estimation for people being tracked at a distance.

The Bayesian fusion method we have used in this work could be readily extended to include other contextual data. We used body direction in this paper but information such as the silhouette is equally interesting. Moreover the descriptor for head-pose could be extended to include information from multiple cameras. The work reported here would be most useful in a causal reasoning context where knowledge of where a person is looking can help solve interesting

Fig. 7. Two people meeting could potentially be identified by each person being in the other's gaze (in addition to other cues such as proximity), as we show in this example.

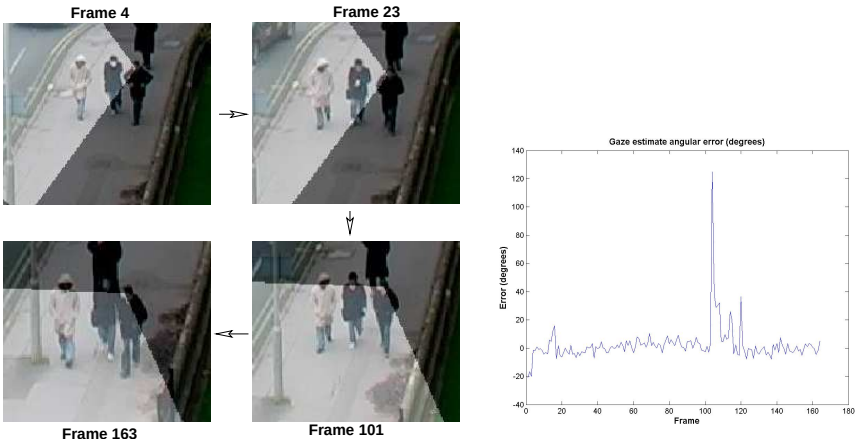

Fig. 8. Second surveillance sequence. The same training data set as used to obtain the results above is used to infer head pose in this video without temporal smoothing. The ground truth has been produced by a human user drawing the line-of-sight on the images. The mean error is 5.64 degrees, the median 0.5 degrees.

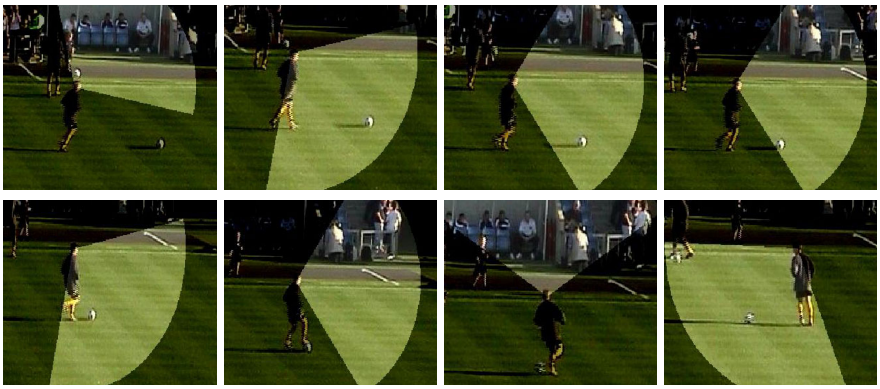

Fig. 9. This example demonstrates the method in soccer footage. The skin histogram is defined only at the start of this sequence to compensate for lighting changes, but the exemplar database remains the same as that constructed initially and used on all the sequences i.e. it contains no examples from this sequence.

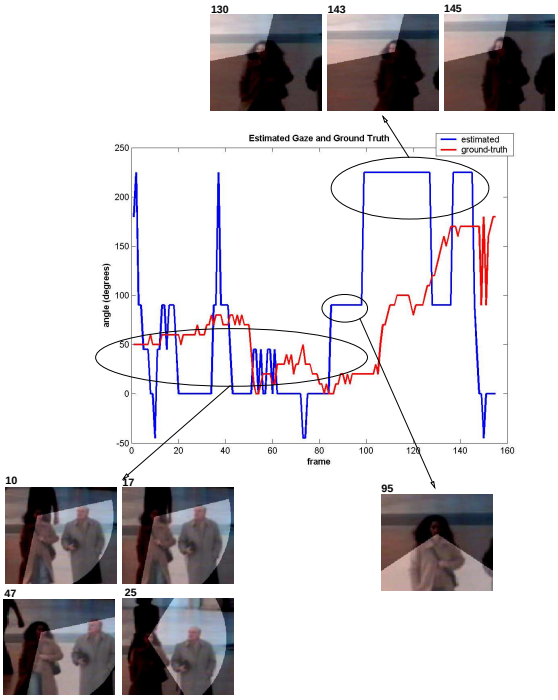

Fig. 10. This figure shows the method tested on a standard sequence (see http://groups.inf.ed.ac.uk/vision/CAVIAR/). The errors are exacerbated by our discretisation of gaze (accurate to 45^o) compared to the non-discretised ground truth (computed to 10^o from a hand-drawn estimate of line-of-sight which we take to be the best-estimate a human can make from low-resolution images) and tend to be isolated (the median error is 5.5^o). In most circumstances it is more important that the significant head-turnings are identified, which they are here, as evidenced by the expanded frames.

questions such as, "Is person A *following* person B?" or determine that person C looked right because a moving object entered his field-of-view. We are currently combining this advance with our reported work on human behaviour recognition [22] to aid automatic reasoning in video.

References

1. J. S. Beis and D. G. Lowe *Shape indexing using approximate nearest-neighbour search in high-dimensional space* IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, PR, June 1997
2. H. Buxton *Learning and Understanding Dynamic Scene Activity* ECCV Generative Model Based Vision Workshop, Copenhagen, Denmark, 2002

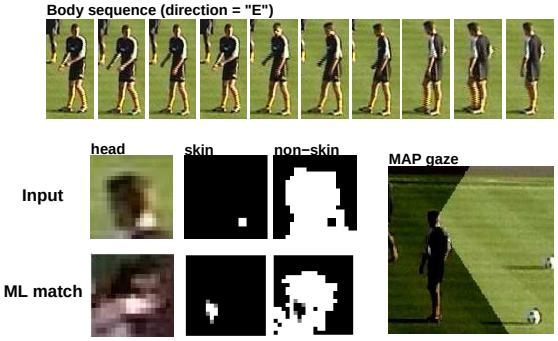

Fig. 11. We show an example here where our method can fail. The mean body direction of the player (in the frames prior to the frame for which we estimate the gaze) is East, since he is moving backwards as his head rotates. The ML match is clearly not correct because the neck has been detected and there is no representation of gaze where the neck is visible in the training dataset. Fusing the direction and head-pose estimate results in the MAP gaze "side-LR", as expected, but incorrect. The reasons for failure are clear: body direction is not a good guide to gaze in this case and there is an unusual input which results in an incorrect match. Either of these can be compensated for on their own with the Bayesian representation we devised but a scenario which combines both is likely to fail.

3. D. Chai and K. N. Ngan *Locating facial region of a head-and-shoulders color image* Third IEEE International Conference on Automatic Face and Gesture Recognitions, Nara, Japan, pp. 124-129, April 1998
4. D. Comaniciu and P. Meer *Mean Shift Analysis and Applications* Proceedings of the International Conference on Computer Vision-Volume 2, p.1197, September 20-25, 1999
5. H. Dee and D. Hogg *Detecting Inexplicable Behaviour* Proceedings of the British Machine Vision Conference, 2004
6. A.A. Efros, A. Berg, G. Mori and J. Malik *Recognising Action at a Distance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003
7. A. Galata, N. Johnson, D. Hogg *Learning Behaviour Models of Human Activities* British Machine Vision Conference, 1999
8. A.H. Gee and R. Cipolla. *Determining the gaze of faces in images.* Image and Vision Computing, 12(10):639-647, December 1994
9. W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. *Using Adaptive Tracking to Classify and Monitor Activities in a Site* Computer Vision and Pattern Recognition, June 23-25, 1998, Santa Barbara, CA, USA
10. K. Hidai et al. *Robust Face Detection against Brightness Fluctuation and Size Variation* International Conference on Intelligent Robots and Systems, vol. 2 pp. 1379-1384, Japan, October 2000
11. T.S. Jebara and A. Pentland *Parametrized Structure from Motion for 3D Adaptive Feedback Tracking of Faces* Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, pp. 144-150

12. N. Johnson and D. Hogg. *Learning the Distribution of Object Trajectories for Event Recognition* Proc. British Machine Vision Conference, vol. 2, pp. 583-592, September 1995
13. J. Y. Kaminski, M. Teicher, D. Knaan and A. Shavit *Three-Dimensional Face Orientation and Gaze Detection from a Single Image*, CoRR, cs.CV/0408012, 2004
14. B.D. Lucas and T. Kanade *An Iterative Image Registration Technique with Application to Stereo Vision* DARPA Image Understanding Workshop, 1981
15. D. Makris and T.Ellis *Spatial and Probabilistic Modelling of Pedestrian Behaviour* British Machine Vision Conference 2002, vol. 2, pp. 557-566, Cardiff, UK, September 2-5, 2002
16. Y. Matsumoto and A. Zelinsky *An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement* Proceedings of IEEE Fourth International Conference on Face and Gesture Recognition, pp. 499-505, 2000
17. J. McNames *A Fast Nearest-Neighbor Algorithm Based on a Principal Axis Search Tree* IEEE Pattern Analysis and Machine Intelligence, vol. 23, September 2001, pp. 964-976 ISSN:0162-8828
18. V.Morellas, I.Pavlidis, P.Tsiamyrtzis *DETER: Detection of Events for Threat Evaluation and Recognition* Machine Vision and Applications, 15(1):29-46, October 2003
19. S. A. Nene and S. K. Nayar *A Simple Algorithm for Nearest Neighbor Search in High Dimensions* IEEE Transactions on Pattern Analysis and Machine Intelligence vol.19, September 1997, p. 989-1003
20. D. Pang, M.D. and V. Li, M.D. *Atlantoaxial Rotatory Fixation: Part 1-Biomechanics OF Normal Rotation at the Atlantoaxial Joint in Children.* Neurosurgery. 55(3):614-626, September 2004
21. A.Perez, M.L. Cordoba, A. Garcia, R. Mendez, M.L. Munoz, J.L. Pedraza, F. Sanchez *A Precise Eye-Gaze Detection and Tracking System* Proceedings of the 11th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2003
22. N.M. Robertson and I.D. Reid *Behaviour understanding in video: a combined method* Proceedings of the International Conference on Computer Vision, October 2005, Beijing, China
23. N.M. Robertson, I.D. Reid and J.M. Brady *What are you looking at? Gaze recognition in medium-scale images* Human Activity Modelling and Recognition, British Machine Vision Conference, Oxford, UK, September 2005
24. H. Sidenbladh M. Black, L. Sigal. *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking* European Conference on Computer Vision, Copenhagen, Denmark, June 2002
25. P. A. Viola, M. J. Jones *Robust Real-Time Face Detection* International Journal of Computer Vision, 2004, 57(3) pp. 137-154