What are you looking at? Gaze estimation in medium-scale images

Abstract

In this paper we describe a new method for estimating where a person is looking in images where the head of a person is typically 20 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 head pose. The general direction of the person is estimated based on velocity. We show that, by combining direction and head pose using a Bayesian Network 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 progress towards automatically inferring gaze direction in images where any one person represents only a small proportion (the head is around 20 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 is then queried to find either a nearest-neighbour match for a previously unseen descriptor or (as we discuss later) the above 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, 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

For human action recognition at a distance Efros [6] 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. 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 this work, clearly gaze inference is vital, but this 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. The AI Lab at MIT has developed an entirely automated system for visual surveillance and monitoring of an urban site [9] but it appears that only trajectories are utilised. 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.

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. In medium-scale images locating significant features such as the eyes and corners of the mouth as used in [8] is an impossible task. Related work has tackled expression recognition using information measures. Shinohara and Otsu demonstrated that Fisher Weights can be used to recognise "smiling" in images. Unsurprisingly, the main application focus of gaze recognition work has been Human-Computer Interfaces and the technical aspects have 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. While this is most useful in HCI where the head dominates the image and the eye orientation is the only cue to intention, this approach is too fine-grained for surveillance video where it must be assumed the eye is aligned with head-pose.

Skin detection has received much attention in the Computer Vision community [3] [10] [11], but it is clear that determining gaze in surveillance images is a challenging problem that has received little or no attention by the vision community. We recognise that skin detection alone will be too error-prone when the skin region is very small as a proportion of the image. However, additional 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

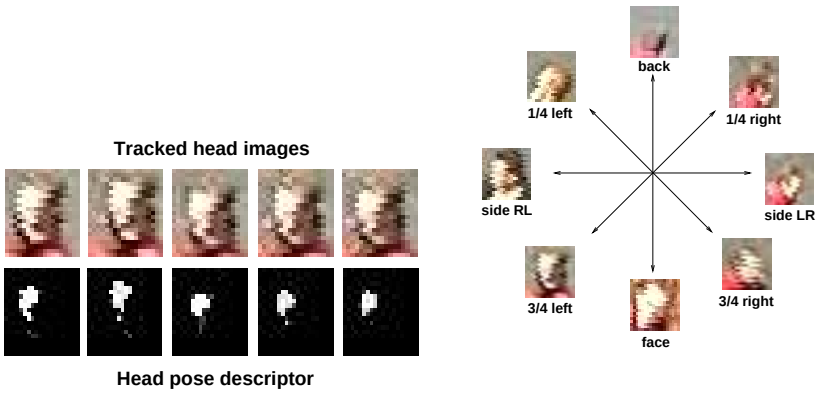

Figure 1: The figure on the left shows the images which result from the mean-shift image patch tracker and, below each, the weight image which represents the probability that each pixel is skin. This is our feature vector which we use to determine eight distinct head poses which are shown and labelled on the right. The same points on the "compass" are used as our discretisation of direction i.e. N, NE, E, etc.

2.1 Head pose feature vector

Though people differ in colour and length of hair and some people may be wearing hats it is reasonable to assume that the amount of skin that can be seen and the position of the skin pixels within the frame is a relatively invariant cue for a person's coarse gaze in a static image. To obtain this descriptor there is a small degree of manual intervention required. First, a mean-shift tracker [4] is hand-initialised on the head. While we anticipate this could be done automatically in the future by modelling the person as distinct "blocks" e.g. head and torso, in this work we concentrate on gaze estimation and assume we have a coarse estimate of which part of a moving "blob" is the head. Second, because there is no specific region of colour-space which represents skin in all sequences it is necessary to define a skin histogram for each scenario. We hand-select a region of one frame in the current sequence to compute a (normalised) skin-colour histogram in RGB-space, with 10 bins. We then compute the probability that every pixel in the head images which the tracker produces is drawn from this predefined skin histogram[1]. Each pixel in each 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 comes from the skin model. The weight image therefore defines our feature vector for head orientation per frame. An example is shown in figure 1.

[1]This will be recognised as a similar approximation to the Battacharyya coefficient as implemented in the meanshift algorithm [4]: $w_{image} = \sqrt{p_{skin}/q_{image}}$.

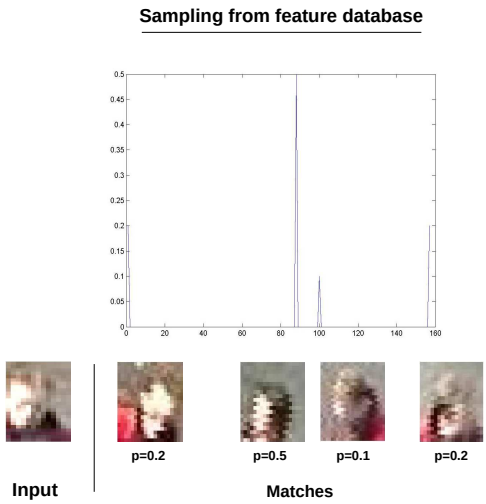

Figure 2: The original target frame (*bottom-left*) is used to compute a descriptor similar to those shown in figure 1, representing skin pixels. This descriptor is then represented as a set of Principal Components and the exemplar database formulated as a binary tree split on the sign of those components. The leaf nodes are indices into matching frames. 10 samples of the database for an input frame and corresponding feature vector are shown here with matching frames and assigned probabilities of a match with the input frame.

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. (In this first attempt we have not made provision for scale changes.) The training data we select is from a surveillance-style camera position and around 100 examples of each view are selected. The head was tracked and the example labelled accordingly. The weight image for each frame is 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 7).

2.3 Matching head poses

The descriptors for each head pose are $(20 \times 20 =)400$ element vectors. With 8 possible orientations and 100 examples of each orientation searching this dataset rapidly becomes an issue. Although linear-time nearest-neighbour search is not intractable unless near real-time performance is desired we consider a tree-search method for two reasons. We

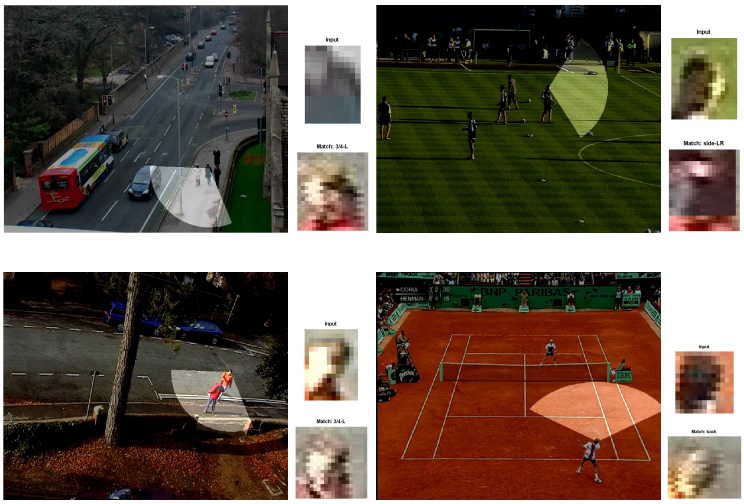

Figure 3: Detecting head pose in different scenes using the same exemplar set. The main image shows the frame with the estimated gaze angle superimposed, the pair of images directly beside each frame shows the input image that the head-pose detector uses (*top*) and the best (ML) match in the database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

elect to structure the database using a binary-tree in which each node in the tree divides the set of exemplars below 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 [23] showed how to formulate a binary tree search in a pseudo-probabilistic manner. This tehcnique was later applied to probabilistic analysis of human activity by [22]. We achieve recognition rates of 80% using this method with 10 samples. An example of such a distribution in this context is shown in figure 2. Results of sampling from this database for a number of different scenes are shown in figure 3.

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 5. However, it is clear the crucial interaction between the two people is missed. To address this issue we compute the joint probability over direction, d, and head-pose, h. The distribution $p(h_m|h_i)$ is estimated by sampling from the databases where h_m and h_i are the ML matches and the input respectively. $P(d_m|d_i)$ is computed using a linear function $p(d) = 1 - \frac{d\theta}{45}$ where $d\theta = |\theta_{true} - \theta_{nearest}|$ where $\theta_{nearest}$ is the projection to the nearest discrete compass point and θ_{true} is the heading computed from the trajectory of the tracked target. For any given data D (here h and d), $p(g|D) = \frac{p(D|g)p(g)}{p(D)}$. The prior $p(D)$ is uniform and the likelihood $p(D|g)$ is defined

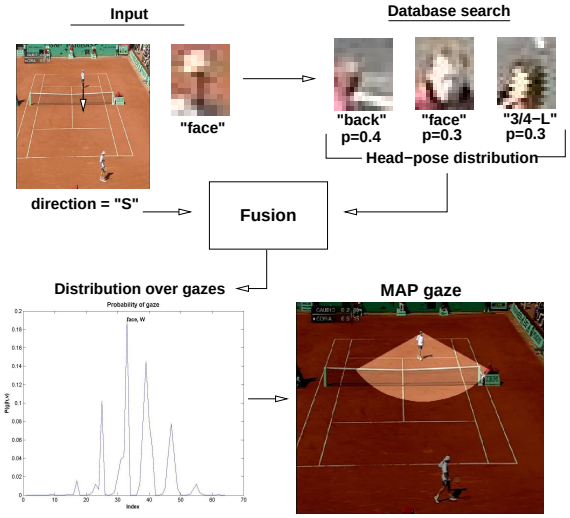

Figure 4: In this example, the ML match for head pose is incorrectly chosen as "back" and the direction correctly identified as "S". Due to the low prior on this combined gaze (it is not possible to turn the head through 180 degrees) this gaze is rejected as the most likely at the fusion stage. The MAP gaze is chosen as "Face, " as shown in the bottom-left image which is a very good approximation to the true gaze as observed in the video sequence.

as a zero-mean Gaussian centred on the current best estimate of head-pose and direction of motion. The prior on gaze is defined using a table which lists expected (i.e. physically possible) gazes and unexpected (i.e. non-physical) gazes. Typically $P(g = g_{expected}) = 0.8$ and, for non-physical gazes where $|\theta_{head} - \theta_{direction}| > 90$, $P(g = g_{unexpected}) = 0.2$.

Now we compute a distribution over 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 (i.e. $p(h_{match}|h_{input})$ can be incorporated in a mathematically sound way. An example is shown in figure 4.

4 Results

We have tested this method on various datasets (see figures 5, 6 and 7). 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 5 we show significant improvement over using head-pose or direction alone to compute gaze (c.f. figure 5, *top-left*). The crucial interaction which conveys the information that the people in the scene are together is the frequent turning of

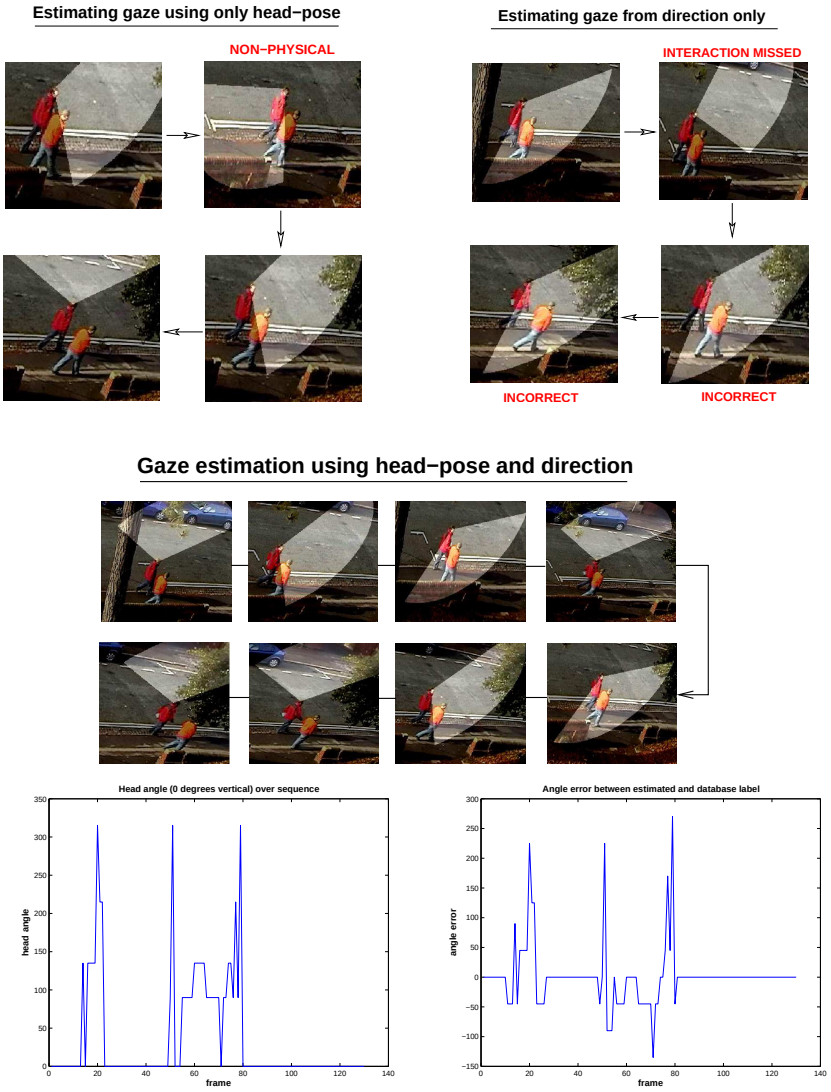

Figure 5: First we estimate gaze using the head-pose estimate alone (*top-left*), resulting in a non-physical gaze (second frame) since the head must turn through more than 90 degrees but there is no mechanism for rejecting such a gaze (even if the likelihood is objectively very low) when the direction of the person is not incorporated into the gaze estimation calculation. Computing gaze using only the direction-of-motion estimated from the trajectory results in critical interactions being missed (*top-right*). In this example the tracked person turns to look at the other person on his right twice. This is not detected (see frame 2) and, moreover, it is estimated that the second person *is* in view in images 3 and 4, which is incorrect. By fusing directon-of-motion information and head-pose estimates the MAP gaze is much improved and the crucial interaction is captured (*middle*). We show the error between the MAP estimate and the ground truth in the graphs below. The mean of the absolute value errors here is 22.27 degrees, the median zero degrees. This error corresponds to one half of our discretisation of head angle (which is 45 degrees), and is a clear improvement on using the ML head-pose or direction estimate alone. Moreover the errors are isolated and Markov smoothing (either Kalman Filter or HMM) of the head-pose could well improve the results even further. The frames for which the head is turned are clearly evident since the angle relative to zero degrees vertical in the image plane increases.

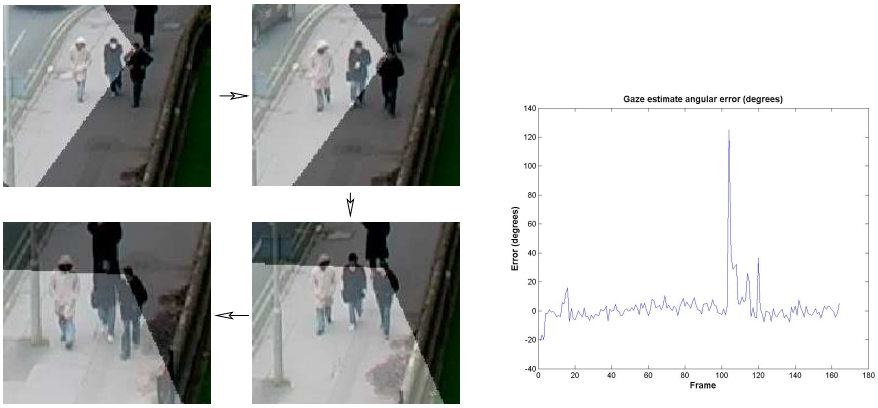

Figure 6: Second surveillance sequence. The same training data set as used to obtain the results above is used to infer head pose in this video. The ground truth is estimated by hand from the images. The mean error is 5.64 degrees, the median 0.5 degrees.

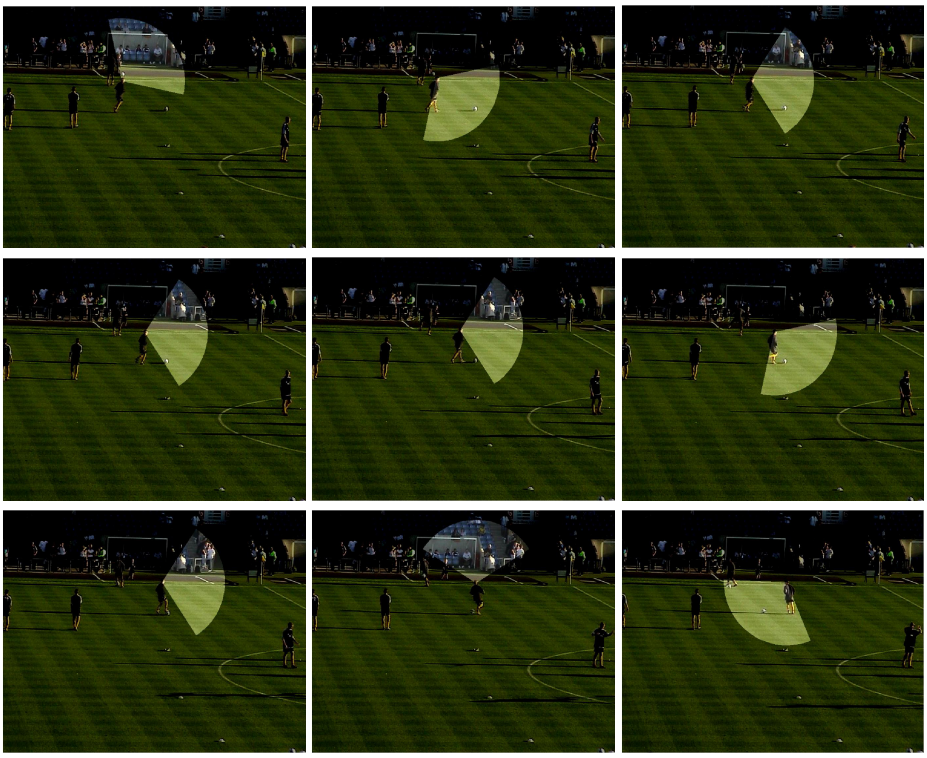

Figure 7: This final example demonstrates the method in soccer footage.

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. Finally we demonstrate the method on sports video in figure 7.

5 Conclusions

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

6 Future work

One source of error is the video tracker which can produce inconsistency in the positions of the skin pixels in the target frame. Matches are to some degree dependent on the location of the skin pixels in the centre of the frame and tracking inconsistency can cause discrepancies to arise. This needs to be investigated.

It seems to us 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 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 in the process of 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 pp.10001006, San Juan, PR, June 1997

[2] H. Buxton *Learning and Understanding Dynamic Scene Activity* ECCV Generative Model Based Vision Workshop, Copenhagen, Denmark, 2002

[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 Recognition (FG'98), Nara, Japan, pp. 124-129, Apr. 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 p1379-1384, 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, 144-150.

[12] N. Johnson and D. Hogg. *Learning the Distribution of Object Trajectories for Event Recognition* Proc. British Machine Vision Conference, volume 2, pages 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] Authors own work. Blank for review.

[23] 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