Guiding Visual Surveillance by Tracking Human Attention

Ben Benfold
bbenfold@robots.ox.ac.uk
Ian Reid
ian@robots.ox.ac.uk

Department of Engineering Science
University of Oxford
OX1 3PJ
Oxford, UK

Abstract

We describe a novel method for directing the attention of an automated surveillance system. Our starting premise is that the attention of people in a scene can be used as an indicator of interesting areas and events. To determine people's attention from passive visual observations we develop a system for automatic tracking and detection of individual heads to infer their gaze direction. The former is achieved by combining a histogram of oriented gradient (HOG) based head detector with frame-to-frame tracking using multiple point features to provide stable head images. The latter is achieved using a head pose classification method which uses randomised ferns with decision branches based on both HOG and colour based features to determine a coarse gaze direction for each person in the scene. By building both static and temporally varying maps of areas where people look we are able to identify interesting regions.

1 Introduction

In ordinary day-to-day behaviour humans identify interesting objects in their surroundings by drawing on knowledge of the world that they have accumulated throughout their lifetime. In contrast, for an automatic reasoning system world knowledge is very limited, so making such inferences is extremely difficult. This work aims to measure the interest of the people present in a scene automatically using remote passive sensing in the form of visual surveillance. The resulting information can be used to direct the attention of an observer towards locations that they might find interesting.

In low resolution images, the main indicator of a person's attention is their face direction. Thus, the system that we have developed automatically locates and tracks pedestrians in surveillance-style video before measuring their head pose to estimate their gaze direction. We then demonstrate how the resulting gaze estimations can be used to identify the subject of interest in three different surveillance scenarios.

The system is based around a multi-target tracking system which is described in section 3. Section 4 describes how randomised ferns are used for head pose estimation with an analysis of how different training methods affect the accuracy. The results from experiments measuring the amount of attention received by the different locations in a scene are presented in section 5.

BMVC 2009 doi:10.5244/C.23.14

2 Related Work

The idea of automatically measuring head pose has been addressed a number of times, most notably by Robertson [14, 15], who was able to automatically recognise basic interactions using gaze estimates. Tian *et al* [16] developed a similar system which could track and estimate the head pose for a single person. Both systems located the head by separating the individual using background subtraction and using the silhouette to find the head. One of the contributions of this work is the development of a robust multi-person tracking system that does not rely on background subtraction, making it capable of tracking the heads of multiple pedestrians through complex environments where occlusions are frequent.

Estimating the pose of human heads is a difficult task due the the large amount of variation in human appearance compared to relatively subtle differences between poses. Various machine learning techniques have been applied to the problem including neural networks [6, 16, 17], nearest neighbour classifiers [9, 15] and model fitting [3, 11, 20]. Our approach is based on randomised ferns, which were recently used to estimate the pose of colour segmented images [1] but we instead make use of histogram of oriented gradient features which Dalal and Triggs [5] have shown to be particularly effective for object detection.

3 Multi-Target Tracking

The first step of processing requires the pedestrians in a scene to be tracked, with the purpose of providing stable head images for the following pose estimation step. We track only the heads of pedestrians rather than their entire bodies for two reasons. The first is that security cameras are generally positioned sufficiently high to allow pedestrian's faces to be seen, so their heads are rarely obscured. The second is that the offset between the centre of a pedestrian's body and their head changes as they walk, so tracking the head directly provides more accurately positioned head images. The general approach we take is to combine absolute location estimates from a head detector based on Histograms of Oriented Gradients (HOG) [5] with velocity estimates from feature-based tracking.

Wu and Nevatia [19] used a similar approach to track pedestrians by combining detections with mean-shift tracking to fill in the gaps between detections. Our approach is instead based on a Kalman filter but we replace the process model, which usually predicts the next state based on physics, with the velocity estimations from feature tracking. Using a Kalman filter allows the two types of measurement to be combined probabilistically and additionally the covariance can be used to limit the region in which the detector needs to be applied.

The video sequences that we used were all fully calibrated relative to a known ground plane. Using calibrated videos allowed the locations of people's feet on the ground plane to be estimated from their head locations by assuming an average human height of 1.7 metres. The calibrations also allowed the approximate head size to be calculated to limit the scale range of the HOG detector.

3.1 Object Velocity Estimates

In this section we describe the method used to obtain robust velocity estimations for the heads by combining the velocities of multiple tracked corner features [4, 8]. The approach works by estimating the probability that each tracked feature is part of each object based on the observed velocities up to and including the current frame. Each object velocity is then

calculated by taking a weighted average which gives a higher weight to the features that are most likely to be on the object. For our purposes, objects represent pedestrians but only corner features on the heads of the pedestrians are tracked.

The probability of a feature being associated with an object is modelled with a dynamic Bayesian network in which each state represents an object, so the probability that a tracked feature belongs to a state represents our confidence that the feature is part of the object in the current frame. Transitions between states occur in frames where a tracked feature moves between objects due to tracking errors. Modelling the probabilities of the associations between features and objects in this way helps to prevent tracked features that are on other close objects or the background from contributing to an object's velocity estimate.

We use the notation $C^{t|t}$ to represent the posterior probabilities in each frame, with elements $c_{ij}^{t|t}$ being the probability of tracked feature j being on object i at time t, given all of the observations up to and including those made at t. The posterior probabilities are updated in every frame by considering the probability of failure (features moving between objects) and the observed feature velocities.

For each new frame, the first step is to calculate the prior probabilities $C^{t|t-1}$ from $C^{t-1|t-1}$ using the transition model S^t, which changes at each time step based on the locations of the objects:

$$C^{t|t-1} = S^t C^{t-1|t-1} \tag{1}$$

The elements of S are calculated by modelling feature trackers as having a failure rate of τ, so $s_{pq} = 1 - \tau$ for $p = q$ and the remaining τ is divided between transitions to the remaining states for which $p \neq q$. Features that are incorrectly tracked are more likely to jump to close or overlapping objects, so the fraction of τ allocated to each destination object is estimated as being proportional to the area of the intersection between the bounding boxes of the two objects. The background is represented as a stationary object with an infinite bounding box.

The next step is to estimate the object velocities v_i^t by searching over the observed feature velocities u_j^t. These feature velocities have probability $c_{ij}^{t|t-1}$ of being samples from the observed velocity distribution of object i, so by taking the sum of the $c_{ij}^{t|t-1}$ for features within a small range of a candidate velocity v, we estimate the density of samples. This density represents the probability of v being the object velocity.

$$P(v|u_1 \dots u_N) \propto \sum_{k \in R_v} c_{ik}^{t|t-1} \tag{2}$$

Where R_v is the set of indices for features having velocities within a small range of v. The probability that v is correct also depends on the previous velocity of object i, so an additional term $P(v|v_i^{t-1})$ represents the prior probability of v being correct given a constant velocity model. The most likely velocity, v_{max} is then calculated by maximising the product of the two terms:

$$v_{max} = \underset{v}{\operatorname{argmax}}\, P(v|v_i^{t-1}) P(v|u_1 \dots u_N) \tag{3}$$

Since we are only able to track fairly small numbers of features for each head due to the relatively low video resolution, we test all of the feature velocities as candidates and calculate the object velocity v_i^t as the mean of the velocities from features in $R_{v_{max}}$ weighted by their corresponding $c_{ij}^{t|t-1}$. For larger sets of features mean-shift could be used to find the maximum.

The object motion is approximated as a 2D velocity because the small head images do not provide enough information to reliably constrain a model with more degrees of freedom. For larger image of more complex objects, methods such as affine transfer [13] which take object structure into account would potentially be more appropriate.

In the final step, the observations from the current frame are used to calculate the posterior probabilities of features belonging to objects, $C^{t|t}$:

$$c_{ij}^{t|t} = o_{ij} c_{ij}^{t|t-1} \tag{4}$$

The observation probability matrix O consists of the probabilities of each feature belonging to each object given the observed feature velocities u_j^t and estimated object velocities v_i^t. The probabilities are calculated by assuming that the velocities of an object's features are normally distributed about the velocity of the object. The result is that we can be confident of features being on objects with distinctive velocities after one or two observations, whereas identifying features on objects with velocities similar to those of other objects takes longer because each observation provides less information. The elements of O also include a weak location term which assigns a very low probability to features that are far from the bounding box of an object.

Tracked features are removed whenever they are significantly more likely to be on the background than on any foreground object and are replaced with new features to maintain a minimum number on each object.

3.2 Tracking Analysis

Two experiments were carried out to test the performance of the tracking algorithm, both using a three minute video in which all 71473 ground truth head regions had been labelled. The first experiment measured the accuracy gained from using the dynamic Bayesian network to model the probabilities of tracked features being on objects. The heads of pedestrians were tracked for twenty second intervals without guidance from the HOG detector and compared with ground truth data to measure the rate of drift. For comparison, similar methods were tested by considering all of the tracked features within the bounding box of each head and calculating the velocity by either taking the mean or using equation 2 with feature velocities having equal weights. The results (figure 1) show that the method we use results in significantly less drift.

The second experiment tested the ability of the tracking algorithm to locate the heads in each frame of the video sequence compared with HOG detections alone. The precision and recall were calculated by thresholding the detection strength from the HOG detector and the Kalman filter covariance from the tracker. In addition to providing more accurate head locations, the tracking allows the detector to be applied to only a small subset of each frame, reducing the processing time by up to a factor of 50.

4 Head Pose Estimation

Randomised trees and ferns have been successfully applied to the tasks of object detection and classification in small image patches [1, 2, 7, 10, 18]. Ferns differ from standard decision trees in that the same set of branch tests are applied to each image, regardless of previous outcomes. The results of the branch tests identify a histogram to which training data should be added and from which the probability distributions for test images are obtained. Ferns

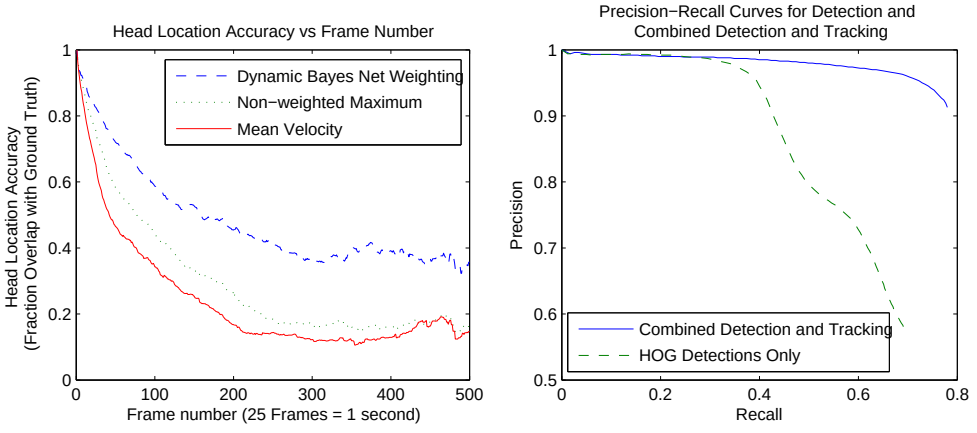

Figure 1: Left: Drift from cumulative errors in the motion estimation without guidance from HOG detections over twenty seconds of tracking (500 frames) Right: Comparison of the combined tracking algorithm to detection alone. The tracking locates approximately twice as many heads before having a significant drop in precision.

essentially store the class distribution for every combination of possible branch outcomes. The decisions in a randomised fern are randomly selected from the set of all possible decisions, representing a subset of the available information. The classification accuracy can be improved by combining the output from a *forest* of ferns.

To estimate the head pose we discretised the gaze direction by dividing the full 360° range into a fixed number of classes. Following classification, the gaze direction for an example was estimated by interpolating around the largest class.

4.1 Fern Decisions

The choice of decisions for a gaze direction classifier is critical; decision outcomes must be able to recognise general properties of each direction class irrespective of the large variations in appearance between people. Decisions based on two different feature types were tested, both compare values from different image locations against one another rather than against a fixed threshold which makes them robust to brightness variations and colour tints.

The first decision type was based on the same HOG features that Dalal and Triggs [5] used to train human detectors. Head images were divided into squares of sixteen cells, each having an orientation histogram with nine bins to which the corresponding intensity gradient at every pixel in the cell was added. The orientation histograms were then normalised both individually and across overlapping square blocks of four cells. Fern decisions were made by comparing the magnitudes of pairs of histograms bins.

The second type of decision was based on Colour Triplet Comparisons (CTCs). Each CTC decision sampled colours from pixels at three different locations within the tracked head region and made a binary decision based on whether the first and second colours were more similar than the second and third colours in an RGB colourspace.

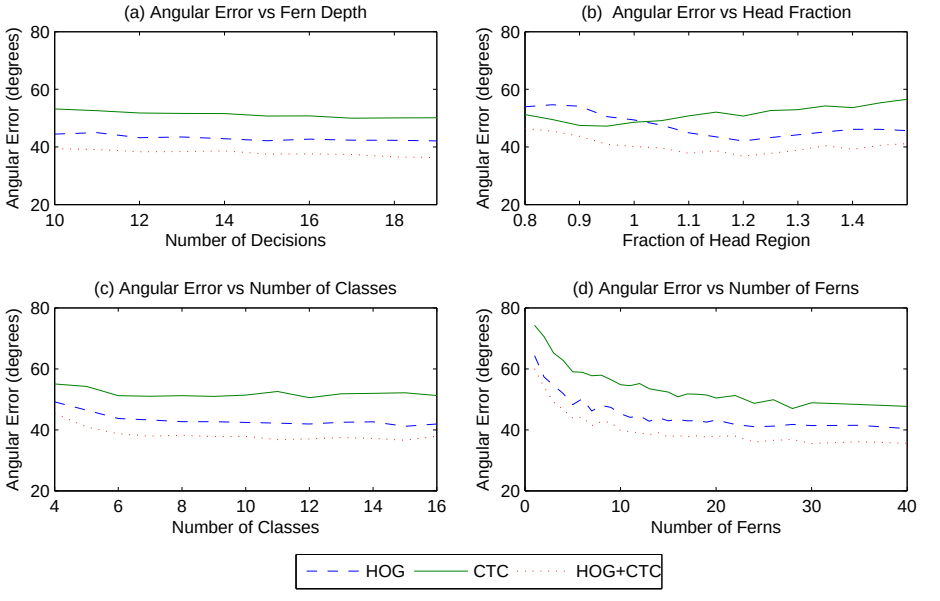

Figure 2: Effects of different training parameters on the estimation accuracy of randomised ferns using the two feature types both individually and combined. Unless otherwise specified, 20 ferns with 16 decisions and 8 classes were trained using a region 1.2 times the size of the head. The errors specified are the mean absolute difference between the angles interpolated from the ferns and the ground truth.

4.2 Training Ferns

The effects of altering the basic parameters when training the randomised ferns using both feature types were tested using a set of cropped head images. Ferns were trained with approximately 1200 labelled example images that were evenly distributed around the 360 degree range and tested on a further 300 images. The results (figure 2) show that the HOG features performed better than the CTC features alone, but a combination of the two gave the best performance.

There is inevitably some amount of translational error in the location estimations from the head tracker which introduces error in the gaze direction. Some experiments were carried out to test different ways of improving the accuracy in the presence of such error. Three different approaches were tried:

Training with artificial errors: Classifiers were trained with example images to which translational error with a standard deviation of up to half the image width was introduced.

Local Maximum: An additional classifier was trained using head/non-head examples and used to search for the most likely head location in a small region around the given position. The most likely region was then classified to estimate the gaze direction.

Weighted Mean: Similar to the local search, except that classification was performed at every location within the search region. The resulting gaze estimate was the mean of the classifications weighted by the likeliness from the head/non-head classifier.

The second two approaches were tested both using separate ferns for detection and clas-

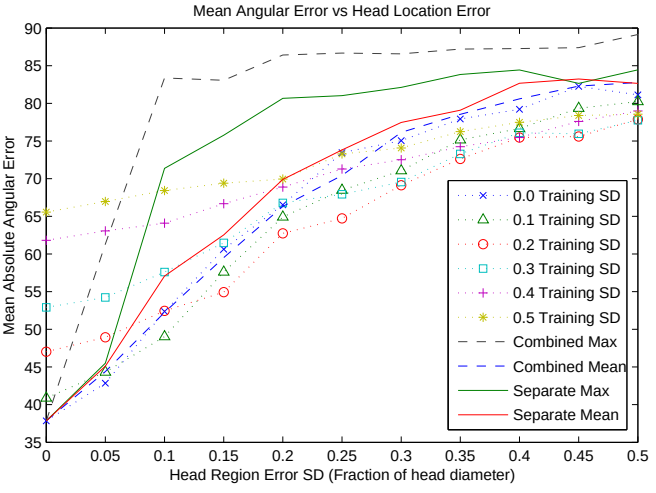

Figure 3: Results from attempts to improve classification accuracy in the presence of random translational error.

sification and also using a combined detector and classifier which used an additional class to represent non-heads. The same 1200 training and 300 test images were used, but normally distributed random error was introduced with a standard deviation of up to half the head width. The results of the experiments are shown in figure 3. Neither the local search or weighted mean improved the accuracy but training with small artificial errors improved the performance when test region error was larger or approximately equal to the training error.

5 System Evaluation

The tracking and head pose estimation were combined to make a fully automatic system which could be used to measure the amount of attention received by different areas of a scene. When applied to video sequences, the direction estimates from the randomised ferns were smoothed using a hidden Markov model to enforce temporal constraints. The gaze direction estimates were also limited to a 180° field of view around the direction of motion when people were moving at more than $0.63ms^{-1}$ (half the mean human walking speed). Using a GPU implementation of the HOG head detector [12], the complete system runs at 15fps on 640×480 video.

For three different video sequences, the locations and gaze directions of the pedestrians were projected onto a 2D ground plane and used to build up an *attention map* representing the amount of attention received by each square metre of the ground. The projected attention density was reduced linearly with the distance from the pedestrian to correct for the increasing field of view width. The first experiment involved the analysis of a video sequence of a busy town centre street with up to thirty pedestrians visible at a time. The aim was to identify areas receiving attention by accumulating gaze estimates over twenty-two minutes of video. The results from tracking approximately 2200 people are shown in figure 4.

In the second experiment, we attempted to artificially draw the attention of people to a particular location in the scene by attaching a light to the wall at eye level. For this experi-

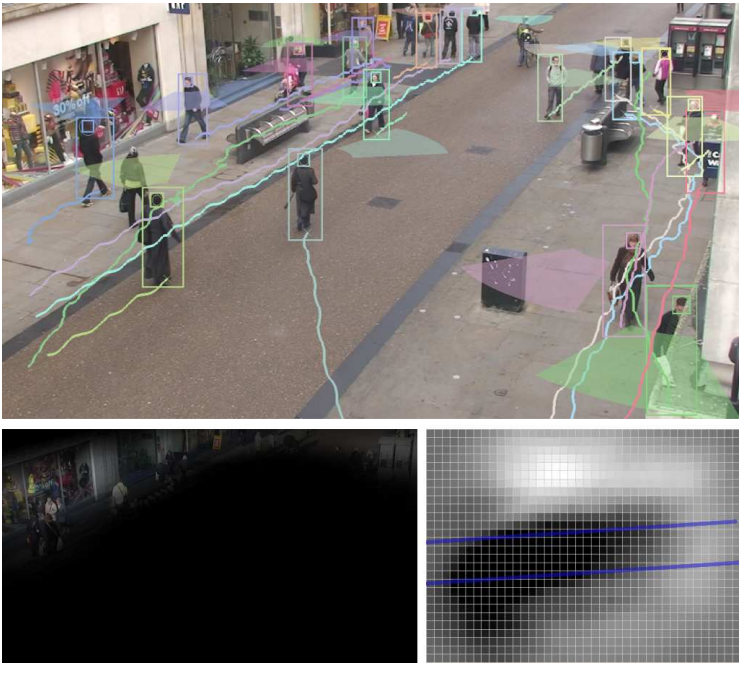

Figure 4: A frame showing the gaze direction estimates and the paths along which pedestrians were tracked. The lower images show the resulting attention map and the result of projecting it onto a video frame, identifying the shop window as a popular subject of attention. The blue lines on the attention map show the edges of the road.

ment the attention map was calculated as the difference between the attention received both with and without the light stimulus to correct for the stimuli normally present in the scene. The attention map resulting from tracking a total of 477 people over 200 minutes of video is shown in figure 5.

Where the purpose of the first two experiments was to measure the attention received by persistent locations, in the third the aim was to identify a transient source of interest. To resolve the ambiguities caused by not knowing the distance between the pedestrians and the subject of their attention, the gaze estimates from both people were multiplied and combined over a sliding window of three frames. The resulting intersection, shown in figure 6 identifies the subject of attention.

6 Conclusions

We have demonstrated a system that is capable of both automatically tracking a number of pedestrians in the presence of occlusions and which can estimate the amount of attention that the pedestrians give to different areas of the scene. In a simple scenario we demonstrated the measurement of transient interest, which could be used to guide a dynamic camera to observe the most interesting areas.

The attention maps from the three experiments demonstrate the potential of the system to provide useful information which could be used for higher level reasoning or camera control,

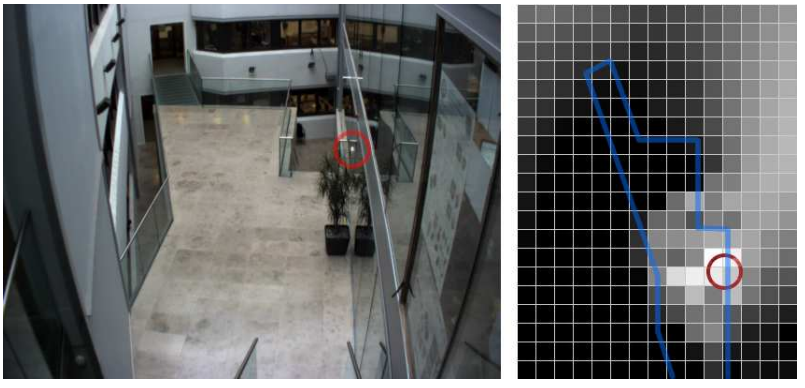

Figure 5: Gaze map (right) resulting from attempts to artificially attract attention using a light mounted at eye level, indicated by the red circle. Blue lines show the approximate floor outline.

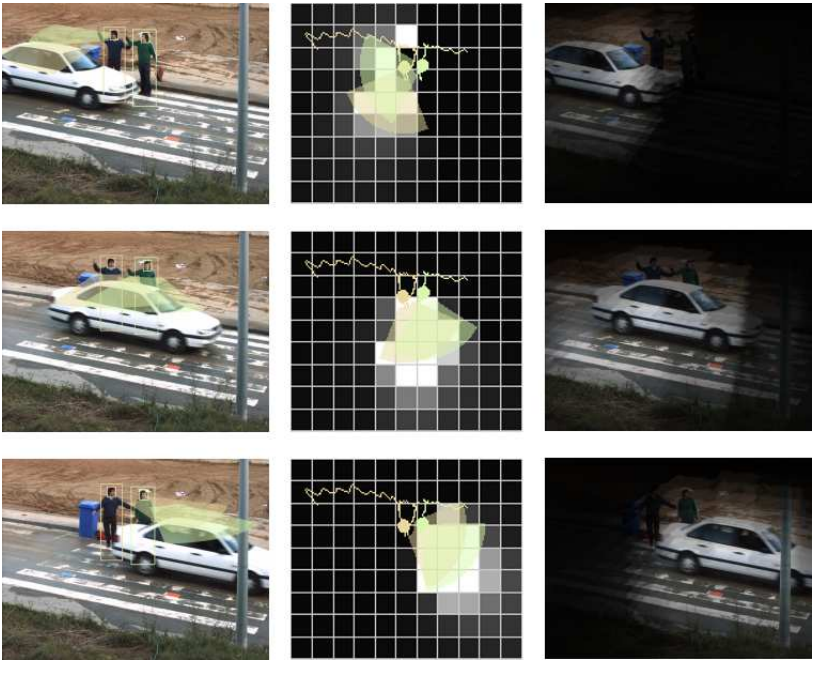

Figure 6: Sequence showing how the attention map can be used to highlight transient areas of interest. The left column shows video frames with annotated gaze directions, the middle column shows the corresponding attention maps and the third column shows the video frame modulated with the projected attention map, under the assumption that the subject of interest is between 0 and 2 metres above the ground

but there is still significant room for improvement. If 3D representations of the scenes were available or vehicles were tracked as well as people then the subject of attention could be more accurately identified.

The experiments that were carried out determined the optimal training parameters for the face direction classifier and identified the error in head localisation as a weak point in the system. It is likely that future research will focus on this issue by integrating the classification and tracking more closely.

Acknowledgements: This research was funded by the EU project HERMES (IST-027110)

References

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

[2] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007. ISSN 1550-5499. doi: 10.1109/ICCV.2007.4409066.

[3] Cristian Canton-Ferrer, Josep R. Casas, and Montse Pardàs. Head orientation estimation using particle filtering in multiview scenarios. In Rainer Stiefelhagen, Rachel Bowers, and Jonathan G. Fiscus, editors, *CLEAR*, volume 4625 of *Lecture Notes in Computer Science*, pages 317–327. Springer, 2007. ISBN 978-3-540-68584-5.

[4] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005. URL http://lear.inrialpes.fr/pubs/2005/DT05.

[6] Nicolas Gourier, Jérôme Maisonnasse, Daniela Hall, and James L. Crowley. Head pose estimation on low resolution images. In Rainer Stiefelhagen and John S. Garofolo, editors, *CLEAR*, volume 4122 of *Lecture Notes in Computer Science*, pages 270–280. Springer, 2006. ISBN 978-3-540-69567-7.

[7] Vincent Lepetit, Pascal Lagger, and Pascal Fua. Randomized trees for real-time keypoint recognition. In *CVPR (2)*, pages 775–781. IEEE Computer Society, 2005. ISBN 0-7695-2372-2.

[8] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In Patrick J. Hayes, editor, *IJCAI*, pages 674–679. William Kaufmann, 1981.

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

[10] Mustafa Özuysal, Pascal Fua, and Vincent Lepetit. Fast keypoint recognition in ten lines of code. In *CVPR*. IEEE Computer Society, 2007.

[11] Ravikanth Pappu and Paul A. Beardsley. A qualitative approach to classifying gaze direction. In *FG*, pages 160–165. IEEE Computer Society, 1998.

[12] Victor Prisacariu and Ian Reid. fastHOG - a real-time GPU implementation of HOG. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009.

[13] Ian D. Reid and David W. Murray. Active tracking of foveated feature clusters using affine structure. *International Journal of Computer Vision*, 18(1):41–60, 1996.

[14] N. Robertson, I. Reid, and M. Brady. Behaviour recognition and explanation for video surveillance. *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, pages 458–463, June 2006.

[15] Neil Robertson and Ian D. Reid. Estimating gaze direction from low-resolution faces in video. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *ECCV (2)*, volume 3952 of *Lecture Notes in Computer Science*, pages 402–415. Springer, 2006. ISBN 3-540-33834-9.

[16] Ying-Li Tian, Lisa Brown, Jonathan H. Connell, Sharath Pankanti, Arun Hampapur, Andrew W. Senior, and Ruud M. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *AMFG*, pages 92–99. IEEE Computer Society, 2003. ISBN 0-7695-2010-3.

[17] Michael Voit, Kai Nickel, and Rainer Stiefelhagen. A bayesian approach for multi-view head pose estimation. *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, pages 31–34, Sept. 2006. doi: 10.1109/MFI.2006.265627.

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

[19] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR (1)*, pages 951–958. IEEE Computer Society, 2006. ISBN 0-7695-2597-0.

[20] Ying Wu and Kentaro Toyama. Wide-range, person- and illumination-insensitive head orientation estimation. In *FG*, pages 183–188. IEEE Computer Society, 2000. ISBN 0-7695-0580-5.