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

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 have developed a system which automatically locates and tracks pedestrians in surveillance-style video before measuring their head pose as an estimate of 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 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. In contrast to similar systems, we have developed 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. 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 head tracking algorithm combines absolute location estimates from a head detector with velocity estimates from feature-based tracking to provide stable head images for the subsequent pose estimation step. A head detector was trained using the Histogram of Oriented Gradients based method of Dalal and Triggs [2] to provide absolute position estimates. The velocity measurements were made by tracking a number of corner features [1, 3] and learning which were representative of the head velocity using a dynamic Bayesian network. The individual feature velocity estimates were then probabilistically combined to give robust velocity estimates for the head. The two types of measurement were combined using a Kalman filter with the process model, which usually predicts the next state based on physics, replaced 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 next stage of processing uses the stable head regions provided by the tracking to estimate the direction in which the person is facing. Randomised ferns, a type of randomised tree classifier, were trained using labelled head images and used to estimate the probability that a given head image belonged to each of eight direction classes. The decisions in the ferns were based on two types of comparison, both of which were designed to be robust against contrast and brightness variations. The first decision type was based on the same HOG features that Dalal and Triggs used to train human detectors and the second was based on a comparison of colours sampled at different locations within the head region.

The tracking and head pose estimation were combined to make a fully automatic system (figure 1) 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. Using a GPU implementation of the HOG head detector, 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. In the first two experiments, static regions receiving attention were identified by accumulating gaze estimates over a long period of time. The third experiment involved locating a transient subject of attention by combining gaze estimates from multiple people, the results of which are shown in figure 2.

The results demonstrate that the system is capable of both automatically tracking a number of pedestrians in the presence of occlusions and estimating the amount of attention that the pedestrians give to different areas of the scene.

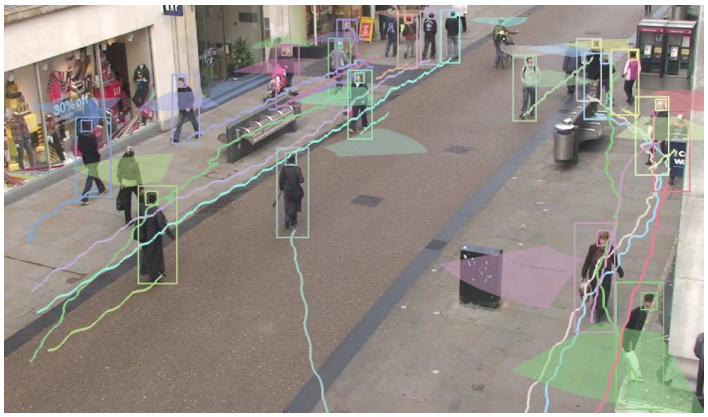

Figure 1: A frame showing the gaze direction estimates and the paths along which pedestrians were tracked.

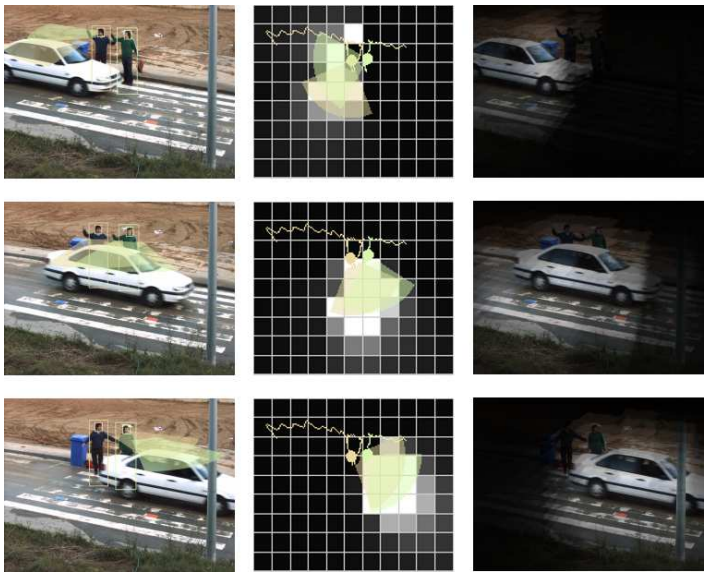

Figure 2: Sequence showing how the attention map can be used to highlight transient areas of interest. The left column shows video frames with annotated gaze directions, the middle column shows the corresponding attention maps and the third column shows the video frame modulated with the projected attention map

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

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

[2] 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`.

[3] 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.