Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator

Ben Benfold and Ian Reid Department of Engineering Science University of Oxford

UNIVERSITY OF OXFORD

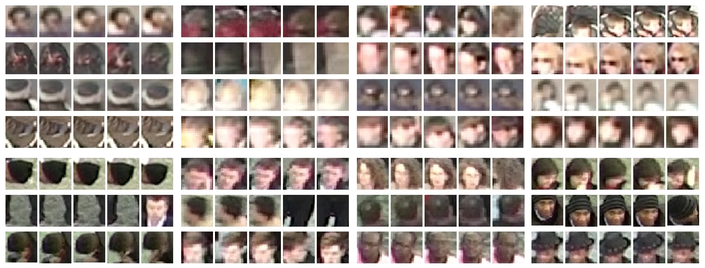
Introduction

Problem: How can we estimate gaze directions for unknown people in unknown scenes?

Scenes have different combinations of viewpoints and lighting conditions and people have different skin and hair styles and colours combined with accessories such as hats and sunglasses. It would be infeasible to collect and label a training dataset representing all possible appearances.

Solution: Automatically learn a classifier using only the output from a head tracking system.

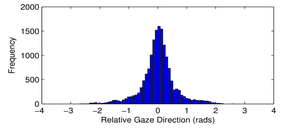
An automatic tracking system provides large (~500,000) datasets of head images for individual scenes. We know that people look most frequently in their direction of travel, so by carefully modelling the gaze behaviour we can infer gaze directions using weak supervision from the walking directions estimated by the tracker.



Conditional Random Field Model

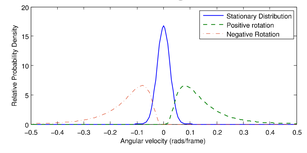
Head Motion Factors

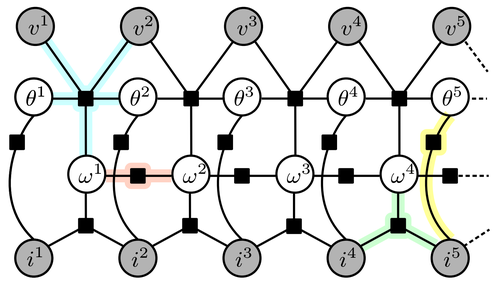
Walking pedestrians are most likely to look in their direction of motion. A state transition matrix specifies the transition probabilities, with the steady state constrained so that the distribution of relative gaze directions tends towards a known prior.



Angular Velocity Factors

People tend to rotate their heads slowly, so the angular velocities of the head at consecutive time steps are correlated. Changes to the angular velocity are modelled using an acceleration matrix.



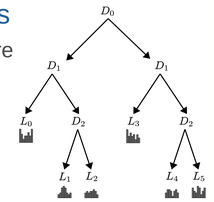


Notation

- θ The gaze direction represented as a distribution over 32 discrete direction classes, each representing an 11.25 degree range.
- ω The angular velocity of the head, represented as a mixture of three components corresponding to clockwise, anticlockwise and no movement.
- v The observed walking velocity, consisting of both direction and speed
- i The observed head image region

Image Classification Factors

Images with similar appearances are likely to represent similar gaze directions. Randomised decision trees are used to model the distribution over gaze directions given the outcome from a set of binary tests. A combination of gradient and colour based decisions are used.



Changing Image Factors

When a person rotates their head we expect consecutive images to be more different than when their head is stationary. A movement probability is estimated based on the number of randomised trees where consecutive images reach different leaves.

Inference

Overview

For each dataset we have scene-specific parameters which consist of the leaf histograms for the decision trees and the movement probabilities from the changing image factors. We would also like to infer the gaze directions for all of the pedestrians for the scene, which are considered to be latent variables. The latent variables and parameters are optimised simultaneously using the Expectation Maximisation algorithm.

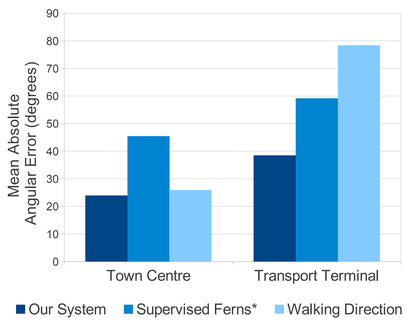
Expectation

The CRFs corresponding to each of the pedestrians contain cycles, so we approximate the expectation over the latent variables using Loopy Belief Propagation (LBP). Messages between nodes are passed in alternating forwards and backwards passes, similar to the Forwards-Backwards algorithm for HMMs.

Maximisation

The leaf histograms and movement probabilities are both maximised by taking the mean over the corresponding expected distributions from the CRFs across all of the people in the scene.

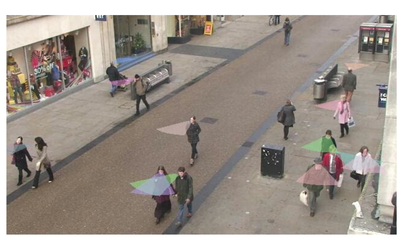
Overall Results



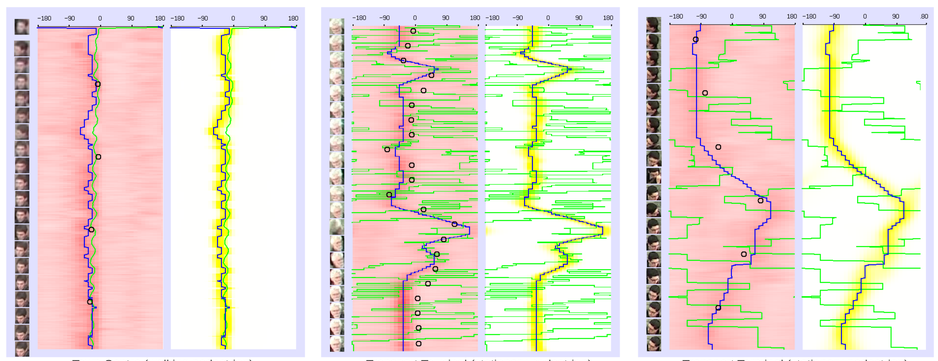
*Benfold & Reid, BMVC 2009

Performance

The system was tested using two datasets – one from a busy town centre street where most people are walking and the other from a busy transport terminal where most people are stationary. On both datasets, the unsupervised system outperforms the walking direction baseline and our previously published randomised fern based classifiers, which required hand labelled training data.



Individual Results



Town Centre (walking pedestrian) Transport Terminal (stationary pedestrian) Transport Terminal (stationary pedestrian)

Estimated State Sequences

In the Town Centre dataset, most pedestrians are walking so their direction of motion is a good indicator of the gaze direction. The Transport Terminal dataset has few moving people, however the learned appearance model for the image classification factors provides enough information for relatively accurate gaze direction estimates to be made.

Key

- Walking direction
- Estimated gaze direction
- Ground truth
- Full CRF model estimates
- Distribution from learned appearance model only

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
http://www.robots.ox.ac.uk/ActiveVision/

This project was generously supported by Oxford Risk
We also thank Joe Hale, Terry Thomson and Eric Sommerlade for their contributions.