University Defence Research Collaboration in Signal Processing

Edinburgh ConsortiumWhite Paper

Exploiting Visual Features for Improved Behaviour Based Target Tracking

Introduction

Video target tracking algorithms are fundamental to a wide range of defence and civilian applications including automated surveillance, traffic monitoring, human computer interaction and virtual reality. Existing algorithms tend to consider targets as point processes (e.g. Probability Hypothesis Density filter), or as re-identifiable objects following known (e.g. Kalman Filter) or unknown (e.g. Mean-shift) motion models. However, these approaches ignore the richness of video, which contains other features with the potential to vastly reduce computational requirements.

To illustrate this point, consider a pedestrian video surveillance application. Pedestrians tend to exhibit ad-hoc obstacle avoidance behaviour but to model all possible motion eventualities has high model complexity. In the Kalman Filter, for example, tracking error will increase when rapid changes in target motion occur which in dense scenes increases the possibility of data association errors. In such cases an intentional prior (a detected feature) that could be used to predict an ad-hoc change in motion is appealing. This theory also generalises to other intentional features: consider a car approaching a crossroads and the indicator light signals intention to turn; this strong visual features would enable better predictions.

Recent research at Heriot-Watt University on Deep Learning has shown that head-pose based intentional priors can be robustly extracted from video data - even in low resolution surveillance data – and used to improve pedestrian target tracking [1, 2].

Deep Learning for Head-Pose Extraction

Deep Learning is a new area of machine learning that replaces hand-crafted features with efficient algorithms for unsupervised feature learning and hierarchical feature extraction. Key to their success is their ability to learn concepts at different levels of abstraction. The most useful features are automatically identified and used for learning higher-level concepts, from which a robust classifier can then be learnt using a final stage of supervised learning.

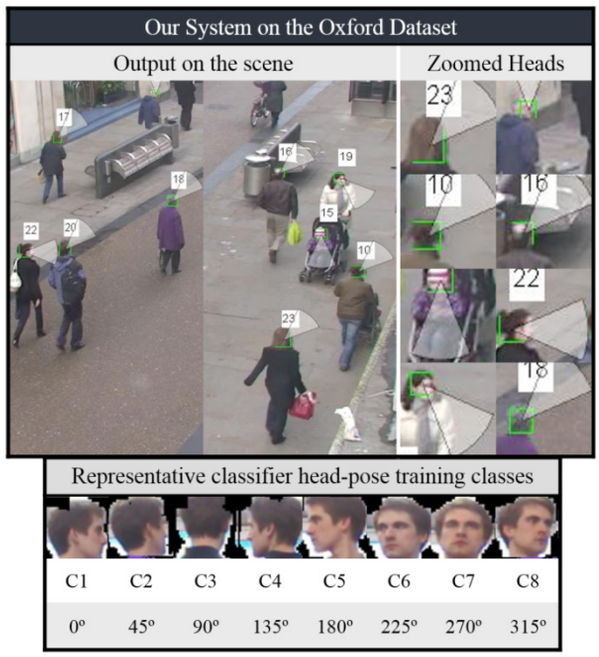
Automatic head-pose estimation has become an important feature for applications of computer vision to surveillance and human behaviour inference, with several significant works dedicated to head-pose extraction from low-resolution surveillance video. Despite these efforts, until recently there was a significant gap in the current methods for unconstrained head-pose estimation in low resolution. This was caused by a reliance on motion priors to smooth head-pose estimates, which reduced the information gain of this rich visual feature.

EPSRC
Engineering and Physical Science Research Council

[dstl]

UDRC

Novel work at Heriot-Watt University has demonstrated that Deep Belief Networks – a form of deep learning – can discriminate between head-pose angles without utilising motion priors [1]. Our state-of-the-art results on benchmark datasets have shown that head-pose angles can now be extracted to within an accuracy of 15.6 degrees, making head-pose a viable feature even in noisy, low-resolution data.



Improved Pedestrian Tracking with Instantaneous Head-Pose Features

Visual features such as head-pose have not typically been used in video tracking yet have the potential to reduce tracking error by providing additional cues about target intention. In the case of head-pose, analysis in [3] showed that it is well correlated with pedestrian direction of travel which makes it a good candidate for an intentional prior.

At Heriot-Watt we have integrated this intentional prior into a Kalman Filter by considering it as an auxiliary input during the prediction step. Initial work has shown that by doing so tracking performance can be improved by as much as 16%. However, the greatest benefits were observed for tracking targets through occlusions, where predictions based on a target's last observed head-pose were found to be significantly more reliable than standard approaches (e.g. a constant velocity model).

With promising initial results the scope for further work in this area is vast. From a feature extraction perspective there are many other potential priors to consider, as well as learning the contexts within which they are useful. Future applications could include improved situation awareness for autonomous cars, using features based on head-pose, car indicators and traffic lights for predicting both vehicle and pedestrian behaviour.

More broadly, intentional priors can be considered a special case of behaviour based tracking – that is – feeding back information about how a target is behaving to improve the underlying target tracking. For coastal surveillance applications the search/mapping strategies of autonomous underwater vehicles (AUVs) could be used for improved AUV tracking, while standard operating formations (e.g. bounding overwatch) could aid infantry tracking in complex urban environments.

References:

[1] S. S. Mukherjee et al. "Instantaneous real-time head pose at a distance", IEEE International conference on image processing, pp 3471-3475, September 2015. DOI: 10.1109/ICIP.2015.7351449

[2] S. S. Mukherjee et al. "Watch where you're going! Pedestrian tracking via head pose", IEEE International conference on computer vision theory and applications, February 2016 (to appear)