Automatic Video Surveillance

Neil Robertson, Ian Reid and Michael Brady

Active Vision Group, Dept. of Engineering Science, University of Oxford

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

Objective

What does it mean to *loiter suspiciously*? Can we provide a computer with the intelligence to detect such an activity? **The ability to pre-empt suspicious, anomalous or dangerous human behaviour is the goal of automatic video surveillance.** In this work we develop a system for human behaviour recognition and anomaly detection in video sequences. Human behaviour is modelled as a stochastic sequence of actions. Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. Action recognition is achieved via probabilistic search of image feature databases representing previously seen, i.e. normal, actions. Behaviour recognition is achieved by computing the likelihood that a set of predefined Hidden Markov Models (HMMs) explains the current action sequence. **This approach allows human-level descriptions of behaviour to be obtained while retaining the benefits of compact models.** This represents a general framework for human behaviour modelling, and we apply it to two application areas: (i) surveillance and in particular anomaly detection; (ii) sports sequences for automated video annotation.

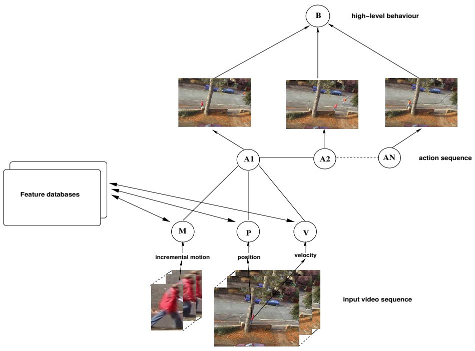
Human Action Recognition

Using the mean-shift tracking algorithm, we extract position, velocity and a target-centred image for each person at each frame. In addition to the target's place and speed we are also interested in the identification of the action of the person we have tracked e.g. *walking* or *running*. An effective method to do this was derived by Efros *et al* (ICCV 2003) which demonstrated the ability to distinguish between types of action and match specific frames within an action sequence.

The **optical-flow** between consecutive frames of a sequence is computed which is ideal as it is **invariant to lighting changes, clothing and appearance.** Invariance is essential as we are seeking a general description of the incremental motion of a person to match the action between different "actors". The Efros *et al* method is only suitable where there is a large, comprehensive dataset from which to choose the matching frame. If there is only a small number of examples of a certain action there is potential confusion between frames from another (incorrect) action sequence. To add **temporal context**, the optical-flow based motion descriptors from a number of consecutive frames, typically 5, are concatenated to form a motion feature vector at each frame. We structure the position, velocity and optical-flow information independently as **databases of principal components**. The best match in each database is found using a pseudo-probabilistic binary-tree search where the tree is split based on the sign of the principal components. This is 20x faster than a nearest-neighbour search: $O(logN)$ compared with $O(N)$).

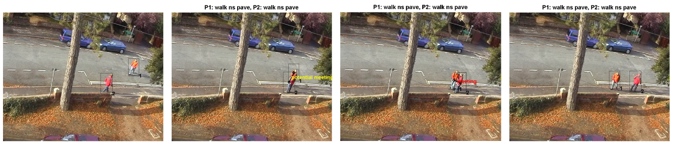
The probability of a certain action is found by fusing the search likelihoods using a Bayesian network. For any given input data d, composed of position (x), velocity (v) and motion-descriptors (m), the probability that a certain action explains the observed data is $p(d|a) = p(a|d)p(d)/p(a)$. The conditional probabilities $p(a|d)$ are hand-coded and the prior $p(a)$ is learned from the frequency of occurrence in training data. A set of HMMs is subsequently learned using the action sequence distribution for a certain, known behaviour. A newly observed sequence of actions $[a_1 \ldots a_n]$ and associated probabilities $[p(a_1 \ldots p(a_n)]$ is used to find the most likely behaviour at frame n by computing the likelihoods of predefined behaviour HMMs. Whether the HMM states are meaningful or not **the maximum-likelihood action sequence provides a rich description of behaviour**.

Behaviour Understanding



Urban Surveillance

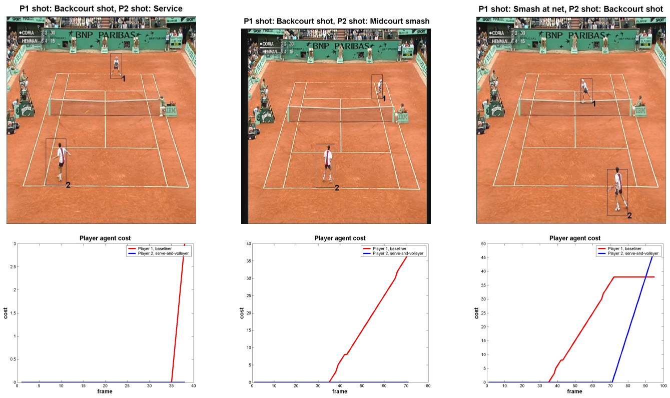
At each frame we obtain an estimation of the action (e.g *walking-on-pavement*) and the behaviour (e.g. `crossing-road`) of each person in the scene.


P1: walk ns pave, P2: walk ns pave
P1: walk ns pave, P2: walk ns pave
P1: walk ns pave, P2: walk ns pave
potential meeting

Potential meeting **Meeting**

It is now possible to **explain interactions**. In order to describe a scenario involving two people in close proximity we use a rule-based system to determine mutual behaviour such as `meeting` or `passing`. Our system reasons in terms of the human-readable descriptions of the individual action sequences by updating a set of known facts according to predefined rules.

Sports Video



Coupled HMMs capture the causal nature of the interaction between players. CHMMs show classification rates of 93% compared to 57% using HMMs. We are able to classify which type of overall play is taking place e.g. `baseline-rally`, demonstrating the generality of our method. By creating a cost-function for *baseliner* and *serve-and-volleyer* types of player, we can automatically determine whether a player is playing out his game-plan: **a higher cost is incurred when deviations from the game-plan take place.**

Conclusions

- By propagating uncertain visual information and incorporating expert domain knowledge we have shown it is possible to classify human action without recourse to large training datasets,
- Our behaviour recognition method can deal with ambiguity and perform novelty detection,
- We can extract a rich, high-level description of behaviour while retaining the benefits of compact models,
- The method is a general framework for video annotation and human behaviour recognition.