A General Method for Human Activity Recognition in Video

Neil Robertson [a,b,*], Ian Reid [a],

[a] *University of Oxford, Department of Engineering Science, Oxford, OX2 7DD, UK*
[b] *QinetiQ, St Andrews Rd, Malvern, WR14 3PS, UK*

Abstract

In this paper we develop a system for human behaviour recognition 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 actions. Hidden Markov Models (HMM) which encode scene rules are used to smooth sequences of actions. High-level behaviour recognition is achieved by computing the likelihood that a set of predefined HMMs explains the current action sequence. Thus, human actions and behaviour are represented using a hierarchy of abstraction: from person-centred actions, to actions with spatio-temporal context, to action sequences and, finally, general behaviours. While the upper levels all use Bayesian networks and belief propagation, the lowest level uses non-parametric sampling from a previously learned database of actions. The combined method represents a general framework for human behaviour modelling. We demonstrate results from broadcast tennis sequences and surveillance footage for automated video annotation.

Key words: visual surveillance, human activity recognition, video annotation
PACS:

1 Introduction

A system capable of inferring the behaviour of humans would have many applications, from visual surveillance to automatic sports commentary. In par-

* Corresponding author
Email addresses: nmr@robots.ox.ac.uk (Neil Robertson), ian@robots.ox.ac.uk (Ian Reid).

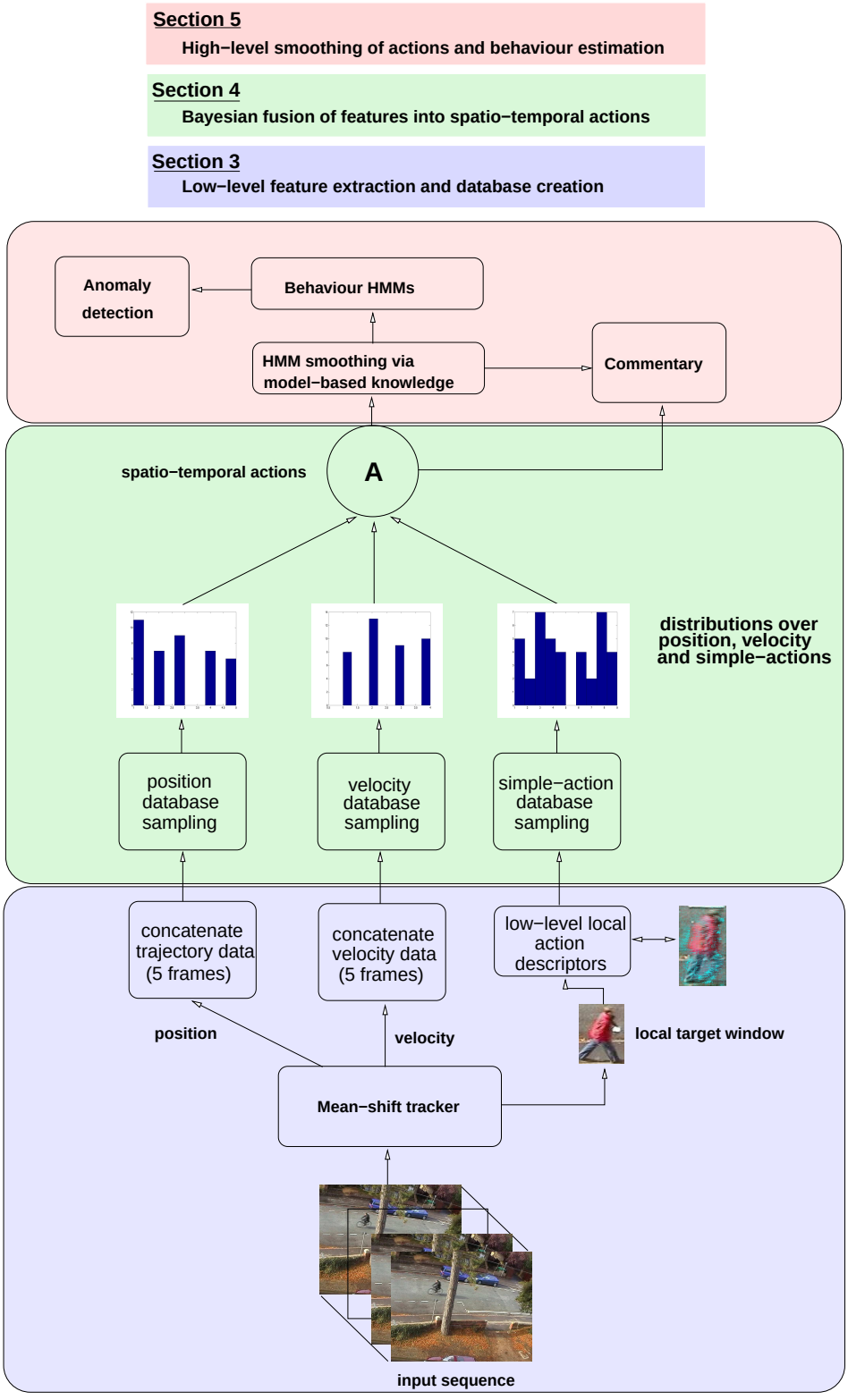


Fig. 1. This schematic diagram illustrates the relationship between image features, actions, action sequences and the high-level parameterisation of behaviour. The diagram relates to corresponding sections of the text as shown in the key at the top of the figure.

ticular, a method for classifying an instantaneous human action, or even better, determining a behaviour that may comprise several actions in sequence, would inevitably be a core building block of the system. In this paper we present progress towards such a system by demonstrating how a data-driven (non-parametric) learning and classification technique for actions can be combined with an effective, HMM (parametric) representation of action sequences, which we use to describe behaviours. An overview of the method is shown in figure 1.

The lowest level of our system, for recognising person-centred actions (e.g. *walking* versus *running*, versus *standing*) is based on the technique described by Efros *et al.* [8] who showed how action recognition can be structured as a search over a comprehensive training database. Although their work was effective for matching frames in video sequences according to similar gross properties of inter-frame motion, the instantaneous action descriptors used are only effective if the training set is very large indeed. In many applications, including our own, there is a need to achieve similar recognition rates but with a much smaller training set. Since mismatches are therefore considerably more likely it is important to match probabilistically for two reasons: (i) so that one knows how different from the input is from the returned set of candidate matches; (ii) to prevent committing to one interpretation of activity early in the process. To this end, we show how an extension to their "blurry motion channel" descriptor, when placed within a probabilistic framework, can effectively disambiguate between types of action.

Efros *et al.* deliberately used position-independent descriptors, and made no attempt to reason at a higher level about the actions. In constrast, we are explicitly interested in higher-level reasoning about action context. In particular the spatial context (where an action happened) and the temporal context (when it happened, and more interestingly, where it occurred in a sequence of actions) are vital for higher level reasoning and thus we take steps to represent both. For example an action "standing still" may be interpreted as normal behaviour in one spatial context (at a bus stop, say), while it may be considered to be the higher-level behaviour "loitering" if it occurs in an alleyway. To this end we consider position and velocity information as additional features; these too are compared against a training database to elicit qualitative position and direction labels, respectively. In an urban surveillance scenario these qualitative descriptors might be, for example, *nearside-pavement*, *on the road*, *far-side pavement* for position, *left-to-right*, *away*, *towards* (etc.) for direction. The results of the three database searches are then fused using a Bayes net to provide a distribution over possible spatio-temporal actions. An example of a spatio-temporal action might be *walking, left-to-right, near-side pavement.* Taking the maximum likelihood (ML) spatio-temporal action at each instant in a sequence yields a commentary of the estimated observed activity. If instead the action distributions are used as input to an HMM which encodes

the known "rules" of the scene then a *maximum a posteriori* (MAP) action sequence results. As a final level of abstraction, we then use further HMMs to characterise high-level behaviour which corresponds to certain patterns of activity. Our approach differs from much previous use of HMMs [4,16,10] in that our HMM input/ouputs are distributions over action types rather than low-level visual features. Abstracting the input/output variables in this way means that much less training data is required for the HMMs, or indeed they are sufficiently clear that they can be modelled manually using "expert" knowledge.

1.1 Contributions of this work

In summary, we make the following contributions:

- By representing position and velocity, in addition to local motion, spatial context is given to recent results in data-driven human action recognition, which is important for higher level reasoning;
- Inspired by Sidenbladh's [23] method for generating a set of particles representing a distribution over trajectories, we structure the search over actions using a PCA decomposition of the database. This yields an efficient search which is $O(logN)$ compared with $O(N)$, (which for our application means 20x faster than for brute-force nearest-neighbour). Additionally, by including a stochastic element to the search, we can obtain a likelihood distribution over possible person-centred actions, positions and velocities (independently);
- The use of a Bayes net for fusion of non-parametric database search results for action recognition generates a distribution over spatio-temporal action;
- Smoothing of action sequences using an HMM which encodes the basic rules of the scene produces a robust text commentary of observed activity;
- Higher level reasoning about scene context by representation of behaviours as action sequences. Representation and recognition of these is achieved via HMMs. Human level descriptions are achieved by abstracting the actions as a precursor.

1.2 Paper structure

This paper is structured as follows. We begin with a review of relevant literature in section 2, then turn to a more detailed description of each of the stages of our algorithm. Section 3 deals with the low-level feature extraction stage. Section 4 describes in detail how we have implemented an efficient probabilistic search of an exemplar training database in order to sample from the action, position and direction distributions. Generating text commentaries on video from the ML action sequences is demonstrated in this section for urban

surveillance footage and broadcast tennis matches. We describe the parameterisation of action sequences for inferring higher level behaviour in section 5. We conclude in section 6 and discuss avenues for future research based on this work in section 7.

2 Review of related work

There has been much reported in the recent literature about methods for training recognition systems using large training data sets (e.g. [26]). This approach is beneficial in the case where not much is known about the class of object one wishes to detect. Nairac and Tarassenko are proponents of the idea that learning normality alone is all that is required for the detection of abnormality [24]. Their work has shown that, using a number of different similarity measures, it is possible to reliably detect unusual behaviour, e.g. dangerous abnormalities in an aircraft engine. In fact, there is an increasing number of examples in the literature describing techniques for action recognition which involve no complex models of activity. Recently Zhong *et al.* [30] demonstrated detecting unusual activity by classifying motion and colour histograms into prototypes and using the distance from the clusters as a measure of novelty. Boiman and Irani [3] detect novelty in video by highlighting activity which cannot be reproduced from other segments of the same image sequence. In all of this prior work, however, there is no attempt to describe the anomaly; the authors are content to simply detect, not classify, unusual activity.

Of most direct relevance to our work is the work of Efros *et al.* [8] which demonstrated that the general actions of people at medium scale can be distinguished by representing the action as a set of Gaussian-blurred motion channels derived from the optic flow between successive frames of the sequence. These non-parametric approaches do not exploit the spatio-temporal relationship between actions and as such do not analyse high-level behaviour. Ke *et al.* used similar optic-flow features to Efros but extended from 2-D to 3-D volumetric features as a descriptor for action recognition [14]. Their method show some promise for spatio-temporal action recognition but the results do not incorporate the spatial component within the scene in any meaningful way i.e. where the action occurs is not significant in their formulation. Grimson *et al.* have developed an entirely automated system for visual surveillance and monitoring of an urban site [10] but does not attempt to explain observed behaviour. Zelnik-Manor and Irani [29] used a distance metric to identify examples of actions in video. Xiang and Gong have addressed the important issue of how to effectively recognise action in a surveillance context when there is a sparsity of example data [27] and what rôle the high-level labelling of trajectories plays in this situation and in the general case [28]. Renno *et al.* have expended considerable effort in creating solutions for a deployable,

wide-area visual surveillance system and have addressed a variety of issues including themes such as colour constancy [19] and learning semantic models [17] in addition to the more common problems associated with surveillance, such as tracking. Notable work includes the investigation of how to track through blind regions i.e. areas between camera views which cannot be seen in any view [2]. These systems typically use trajectory information alone, however.

A number of parametric methods have been formulated for recognising action. Brand and Kettnaker use HMMs for this purpose [4]. Buxton has used Bayesian networks for visual surveillance [6] as has Town [25]. Makris [16] also uses HMMs for detailed modelling of trajectories from learned geometric route data. Porikli and Haga [18] include object-based and frame-based features, parameterised by an HMM. Galata, Johnson and Hogg [9,11] use Vector Quantisation (VQ) to group and classify trajectory data. Indeed there is a notable attempt by Johnson and Hogg to introduce the concept of action and behaviour into classification systems [11]. While the parametric approaches demonstrate a degree of success in classifying complex activity, there is a tendency to use the parameterisation as a "black-box". Therefore a lower-level description is not derived, certainly not in human-readable terms. In this work we use intermediate levels of abstraction from simple-actions (e.g. *walking*) through spatio-temporal action (e.g. *walking-on-the pavement*) to behaviour extended through time which is composed of sequences of spatio-temporal actions (e.g. *crossing-the-road*). Bregler [5] achieved human gait recognition at multiple levels of abstraction, from image blobs to higher-level HMMs, which was demonstrated for clasification, but neither explanation or reporting on human behaviour at a human-readable level was attempted.

There have been efforts to explain behaviour in video. Attempts to describe and query video at the action and not the feature level [13] has involved combining research on Question Answering and Natural Language Understanding [12] with Computer Vision. The system presented by Katz *et al.* uses surveillance video as an example and can answer questions at the level of, "Did any cars leave the garage?" [12]. This is a rare example in the literature of an attempt to interpret video at a human level. At a similarly high level, although they deliberately ignore the challenge of extracting the required information directly from video, Rigolli and Brady have analysed driver behaviour, explaining activity using an agent and rule-based representation [20].

3 Low-level feature extraction

The main component of our human-activity recognition method is action recognition via non-parametric matching of trajectory data and instantaneous motion descriptors, fused via a Bayes net. This is split into two stages, as can

be seen in the bottom and middle sections of figure 1 (respectively, the blue and green coloured portions). The first step is based on data extracted from a colour-based image patch tracker. We describe this stage of the system in this section and the subsequent Bayesian fusion stage in section 4.

3.1 Target description

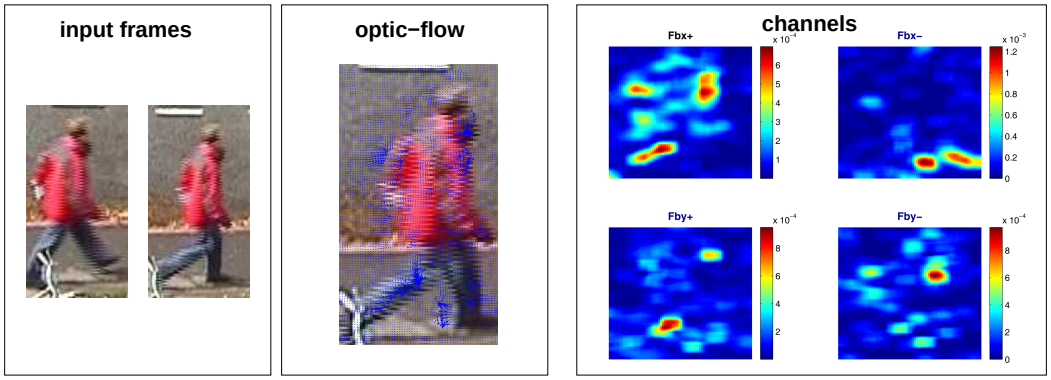
Using a mean shift tracking algorithm [7], we extract the following information for each target for each frame: position, velocity and a window around the target. In addition to the target's place and speed we are also interested in the identification of the action of the person we have tracked e.g. *walking* or *running*. An effective method to do this was proposed by Efros *et al.* [8]. In that work a local motion descriptor based on coarse optic flow is extracted from a target window. The optic flow between consecutive frames of a sequence is computed. We use the Lucas-Kanade technique which compares favourably with most other published techniques [1].

We define a "person-centred action" as the activity independent of any position in the world coordinate frame. The overall motion of the target is the motion of the object-centred coordinate frame in the scene, and is considered independently using trajectory features. The local motion is therefore the motion, within the target-centred window, relative to the object-centred coordinate frame. This local motion descriptor is compared against a dataset of previously seen local motion descriptors that have been hand-labelled with their corresponding actions. The nearest-neighbour (ML) match provides an action label for the current data.

3.2 Computing a local motion descriptor

For completeness, we now discuss in more detail the method of Efros *et al.* for computing descriptors of human action based on optic flow.

The rationale behind the Efros descriptor is clear. Action is almost always correlated with motion. Human action, in particular, is readily identified by the motion of limbs. The optic flow between consecutive, person-centred, images is a measure of that motion, independent of where in the scene that motion is taking place. By blurring this optic flow descriptor, the prominent areas of motion are identified e.g. hands, feet etc. Further, by splitting this blurry descriptor into "channels" (shown in figure 2) the characteristic components of the motion are described (i.e. which direction the hands, feet etc. are moving). Optic flow is ideal for this purpose because it is photometrically invariant and invariant to clothing or appearance [15]. Invariance is essential as we are



(a) (b) (c)

Fig. 2. The target-centred action descriptor: (a) The pair of input images; (b) The flow vectors superimposed on one frame, (c) The blurred optic flow in the x and y direction is further split into the four (Gaussian blurred) non-negative channels.

seeking a general description of the incremental motion of a person to match the action between different "actors".

The optic flow vector-field, $\mathbf{F}$, is split into two scalar fields which are the horizontal and vertical components of the optic flow field, F_x and F_y. These are then half-wave rectified into positive channels F_x^-, F_x^+, F_y^- and F_y^+ such that:

$$F_x = F_x^+ - F_x^- \tag{1}$$

$$F_y = F_y^+ - F_y^- \tag{2}$$

Each of the channels is blurred with a Gaussian kernel and normalised, producing the four motion descriptors for every frame of the sequence $\hat{F}b_x^+$, $\hat{F}b_x^-$, $\hat{F}b_y^+$ and $\hat{F}b_y^-$. An example is shown in figure 2.

3.3 Limitations of the optic flow descriptor when data is sparse

In the original work, Efros *et al.* had a comprehensive training dataset and did not, therefore, encounter any difficulties due to sparse training data. Additionally, while temporal context seems to be accounted for in their version of cross-correlation, this important aspect is not discussed in their work in any detail.

We find that problems due to the lack of comprehensive data are revealed by the fact that significant mismatches occur for even subtle variations in motion. For example, figure 3 shows that a "wandering" motion cannot be reliably

Fig. 3. A slight variation of normal activity is introduced in the new input sequence (*row 1*). Despite this being only subtly different from the normal activity found in the training database, significant mismatches occur such as motion reversing (see frame 7).

correlated with "walking, left-to-right" despite this being, for an experienced observer, the most obvious choice.

To obtain the experimental results shown in figure 3 we captured four exemplar sequences from surveillance data which represent the typical motion in the scene, as decided by an expert user. By matching the same person at different times we minimise errors which could potentially arise due to variations in size and shape. The database is small, containing 800 frames comprised of four distinct activities. These examples have been labelled by a user as "walking, left-to-right", "walking, right-to-left", "walking, away-from-camera" and "walking, towards-camera".

3.4 Extending the motion descriptor

Figure 4 shows a comparison of matching when the descriptor has no temporal context, and when the descriptor is the concatentation of the optic flow channels from 5 consecutive frames. The match in each case is the ML match from 10 samples of the motion-descriptor training database. In the first example shown in figure 4(a), walking in one direction has been confused with walking in the opposite. This is because, instantaneously (i.e. frame-to-frame) the movement of arms and legs is dominant whereas the body direction is not. We concatenate the motion-descriptor data from 5 consecutive frames which provides temporal context and results in the ML matching exemplar being less ambiguous (as shown in figure 4(b)). Importantly, the motion does not reverse in this case.

In the case of walking, the body direction becomes significant and clearer over 5 frames relative to the swifter arm or leg motion which can be computed accurately per frame. This helps to reduce some of the ambiguity which arises when the exemplar data may not be comprehensive, as in a surveillance

(a) (b)

Fig. 4. Increasing the temporal extent of the motion descriptor gives improved performance when data is sparse.

scenario where it is not the same person who is being repeatedly viewed.

An additional issue not addressed by Efros *et al.*, but one that is highly significant in an urban surveillance context, is that of cluttered environments. The examples provided in [8] are exclusively from the sports domain. This assumption does not translate to urban environments where frequently a person's limbs are obscured by immovable static objects (e.g. lamposts, trees) and by non-static objects (e.g. people, cars). As we have seen, this is likely to cause considerable problems for choosing the best-matching model even if training data were comprehensive. Therefore, a motivating factor for the approach we now describe is the need to model *how similar* the current observation is to what has been seen before. In a surveillance context (in contrast to a football

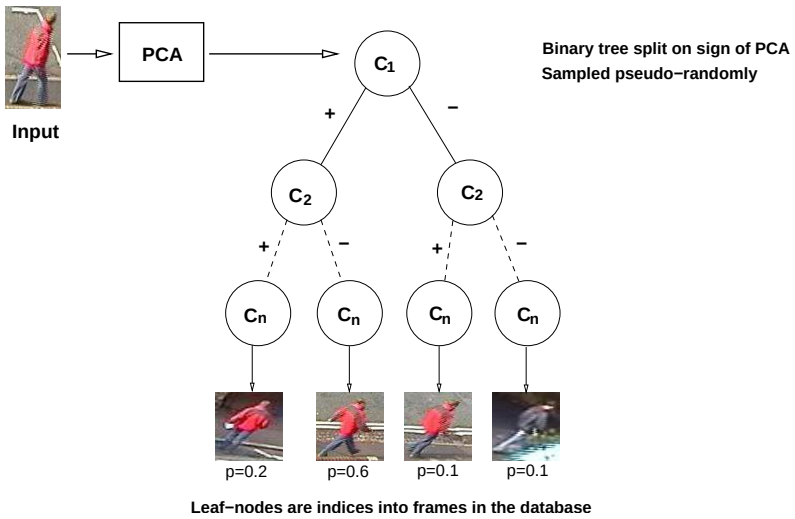


Fig. 5. We sample from the exemplar databases to produce a distribution over the training data given a newly observed descriptor computed from the current sequence. We show here illustrative matches generated for 10 samples with associated probabilities.

match) it is impossible to have a representation of action from every variety of gait. Where a mismatch is eminently possible, instead of computing the ML match and committing to one estimate of activity, it is preferable to compute a probability distribution over the training data. The technique we use for achieving this is described in the following section.

4 Spatio-temporal action recognition

We now discuss the database organisation and search techniques, which relates to the second stage of figure 1. The database search is not trivial for two reasons: (i) the volume of data from the blurry motion descriptors presents a challenge for efficient search: there are 30000 entries in a single local motion feature vector for a 30×50 pixel target (when the temporal extent of the descriptor is 5 frames) - the overall data volumes for the training data is shown in table 1; (ii) for more effective data fusion (and, necessarily, for appropriate use of a Bayes net) we do not simply want one nearest-neighbour (i.e. ML) match, but rather a distribution over possible matches.

In [23] a large database of high-dimensional points is structured as a binary tree via principal component analysis of the data set. The children of each node at level i in the tree are divided into two sets: those whose i^{th} component (relative to the PCA basis) is larger and those whose value is smaller than the mean. In Sidenbladh's application each data point comprised the concatenated joint angles over several frames of human motion capture data. The method,

Sequence	**Total (frames)**	**Example database (frames)**	**Test sequences (frames)**
Urban street	5455	665	2361
Junction surveillance	76040	4491	18445
Tennis	90000	494	3132

Table 1
The data volume for each of the videos used in the analysis of our technique.

Search type	**Detection rate (%)**	**Search time per sample (secs)**
Nearest-neighbour (full data)	83.2	0.461
Nearest-neighbour (PC coeffs)	81.9	0.426
Sampling (per sample)	77.9	0.023

Table 2
Comparison of detection rate for three types best-match search. As expected full comparison of the input descriptor (*row 1*) gives best results.

however, applies equally well to our application of image feature data and the pseudo-random search algorithm is identical to that derived in [23], and we do not repeat the detail in this paper.

This search method is used for two reasons: it is more efficient than a linear nearest-neighbour search and the ability to return multiple neighbours represents a distribution over possible actions i.e. a likelihood. The search time is improved by a factor of 20 and, since we sample many times, the search provides a set of particles which represents a distribution over the exemplar feature vectors into frames of the previously seen examples.

We achieve recognition rates of around 80% (the correct example is chosen as the ML model in almost 8/10 queries) using this pseudo-probabilistic sampling method with 10 samples. For comparison, we show the statistics for a linear search of the complete feature set and for a linear search on the PCA components derived from the features in table 2.

As shown in table 2, using only the Principal Components for brute-force nearest-neighbour search gives very similar results with little improvement in efficiency. This is due to the fact that the order of search is the same i.e. $O(N)$. The sampling method returns a distribution over possible matches and the figures quoted are for the frequency of ML match corresponding to a true match. While detection rate is slightly inferior, the probabilistic information can be exploited and the search is considerably faster. Since we, typically, sample 10 times, the complete search returns a set of particles representing a distribution over matches of the motion-descriptors into frames of the previously seen examples. Illustrative example of the distribution over the exemplar database for an input sequence is shown in figures 5, 6 and 7.

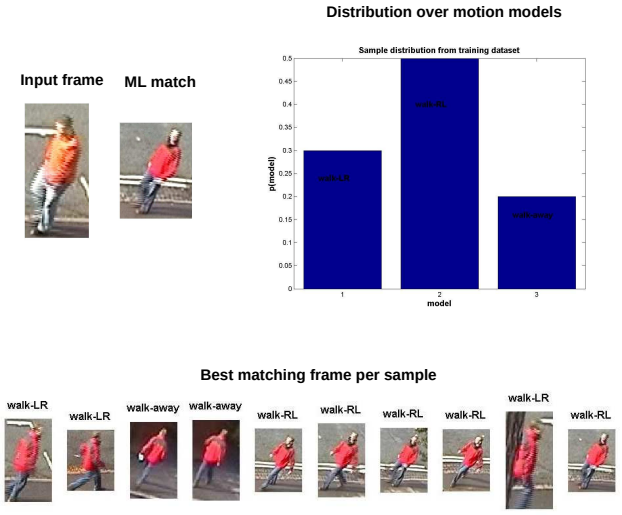


Fig. 6. Pseudo-probabilistic sampling from the exemplar database. The input frame (*top left*) is shown beside the ML frame from 10 samples of the motion-descriptor database. The more complete information is provided by the sampled distribution of matches from the database. These are shown *top right*: the distribution over model-types in the exemplar set, and, *bottom row*: the matching frames for each sample of the database.

Fig. 7. The example set in this second surveillance scene is comprised of 27 different types of spatio-temporal activity with a range of person-centred actions. The action *walking* matched correctly into the exemplar database by taking the ML match from all samples at each frame is shown here (*left*).

4.1 Action likelihood computation

As we noted earlier, Efros *et al.* ignored all positional information. In contrast, we argue that such information can be important in placing an action in its spatial context, particularly in an urban surveillance application.

To that end, we also create additional databases of previously seen trajectories (position and velocity). In each case, the feature vector is the concatenation of a few (typically five) frames worth of position and velocity data, and the database examplars are labelled with qualitative position and qualitative di-

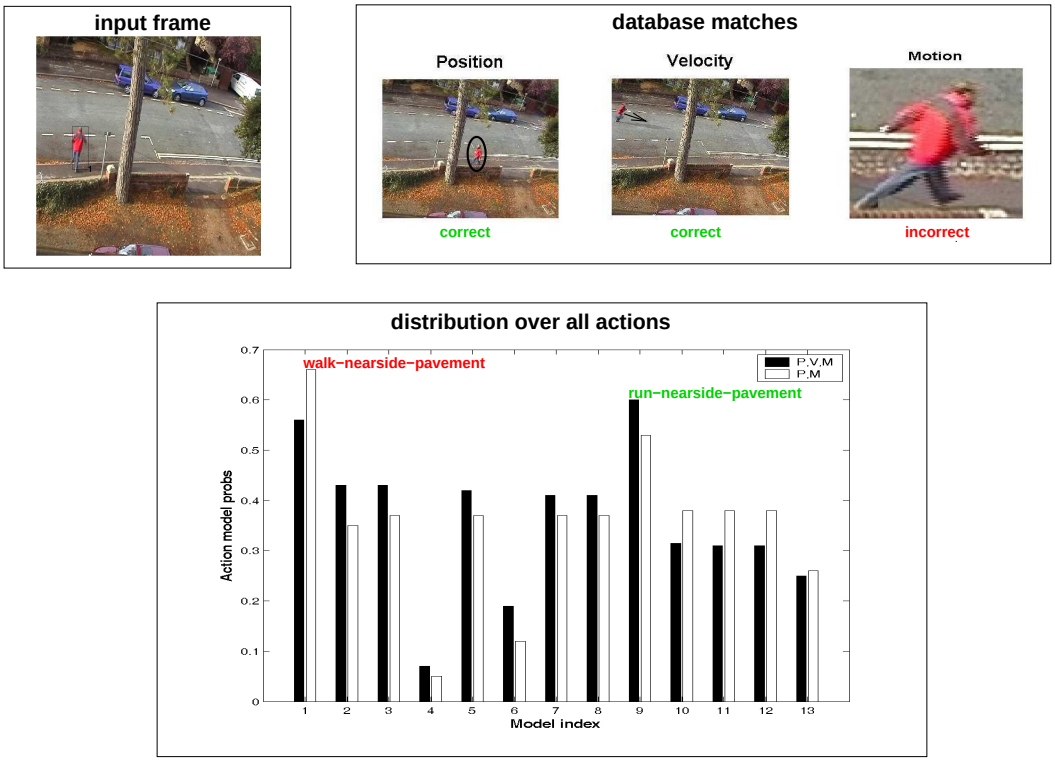


Fig. 8. Velocity, motion-type and position are equally as important for action-recognition.

rection labels e.g. *Pavement* and *Left-to-right* as opposed to a specific image velocity or image coordinate. The databases of position, velocity and local motion are maintained independently, and the set of "normal" actions is the set of combinations of the qualitative labels attached to the exemplars in the feature databases. Matches from the position, velocity and motion-descriptor databases are fused using a Bayes net which is described below.

By fusing the likelihoods of the matches from the position, velocity and motion-descriptor exemplars we compute the probability of a *spatio-temporal action* such as *walking-left-to-right-on-nearside-pavement*. We use a Bayes Net to effect this information fusion: if the spatio-temporal action is denoted a, x is the index into a qualitative position label in the database, similarly v is the index into a qualitative direction label, and m is the index into a person-centred action label, then assuming conditional independence yields

$$p(a, x, v, m) = p(a)p(x|a)p(v|a)p(m|a) \tag{3}$$

The distributions $p(x_{match}|x_{input})$, $p(v_{match}|v_{input})$ and $p(m_{match}|m_{input})$ are estimated by sampling from the databases. We compute the marginal distribution $p(a)$ since, for any given data d (here x, v and m),

$$p(d|a) = \frac{p(a|d)p(d)}{p(a)} \tag{4}$$

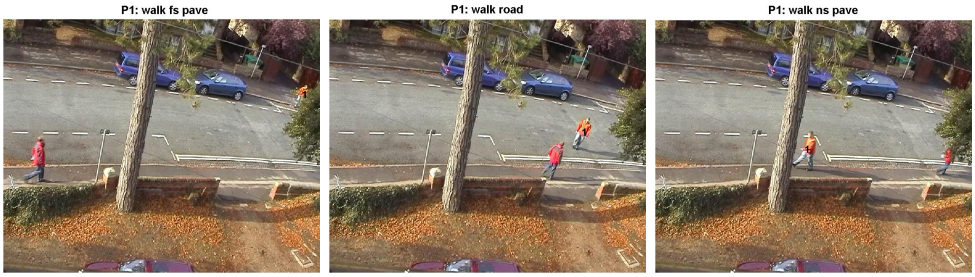
$p(a|d)$ is specified in the conditional probability table for the node a, $p(d)$ is defined from the frequency of occurrence of data d in the training set and $p(a)$ is uniform in most cases. Figure 8 specifically highlights the significance of each feature for successful action classification. In figure 8 the ML motion-type is (incorrectly) classified as *walking*. When the resulting distributions from each of the inputs (i.e. position, velocity and motion-type) are fused the ML estimate is now (correctly) *running-on-nearside-pavement*. The action probability distribution is when velocity is excluded (right-hand distribution) and included (left-hand distribution i.e. shaded bars) are compared.

4.2 ML action commentaries for urban surveillance and tennis matches

In addition to providing a general method for probabilistic human activity recognition in medium to low-resolution video, a useful application of the techniques developed here is the ability to generate text commentary of observed activity by taking the ML estimate of the spatio-temporal action distribution.

At each frame the distribution over all possible spatio-temporal actions is computed using the evidence from the action recognition method described above. To generate a commentary the ML action is chosen and the best description of that spatio-temporal action is used to describe the person's activity. The validity or the accuracy of the description is dependent on: (a) the descriptive language used to label the exemplar sequences in the databases of position, velocity and person-centred action; (b) whether that action has been seen before and how often. The former is an issue for the expert user to ensure that the language used to described the scene is accurate. The dependence on the latter is mitigated by the fact that each activity has a likelihood of occurrence and thus a taking final, hard decision can be avoided too early. The best match is a specific spatio-temporal action but may be ambiguous. If the current action has a number of candidate matches, then the distribution over spatio-temporal actions reflects this uncertainty.

For an urban scene, with a limited set of typical activities, we demonstrate extracting a basic text commentary from the distribution of spatio-temporal actions in figure 9. In figure 10, an example of abnormal activity is captured. Because it is abnormal it is not represented in the exemplar databases of position or person-centred action. The resulting activity therefore has a much lower likelihood of occurrence than the normal example of figure 9 i.e. 0.94 vs. 0.34. This likelihood can be interpreted as a measure of confidence in the accuracy of the commentary.



Frame	Activity	Likelihood
1 - 70	Walking on far-side pavement	0.86
71 - 225	Walking on road	0.94
226 - 450	Walking on near-side pavement	0.94

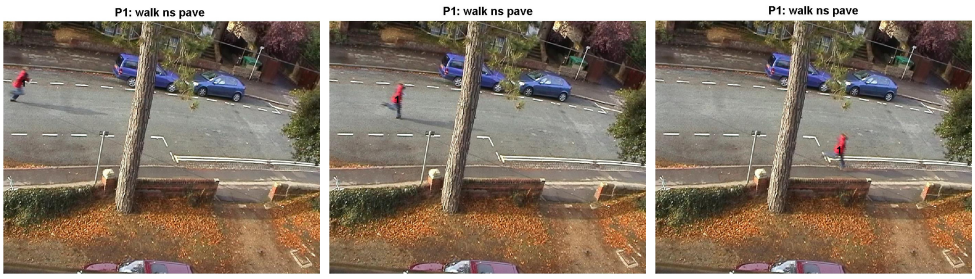
Fig. 9. An accurate commentary is obtained for this urban scene where the person entering from the right is under observation.

We next show the recognition of activity in a more challenging scene, shown in figure 11. The high-level markup by a user with expert domain knowledge is provided for this scene in figure 11. The qualitative positions, actions and directions for pedestrians in this second scene are labelled as shown in the table in figure 11.

Figures 12 and 13 demonstrate activity commentaries by taking the ML estimate of the distribution over all spatio-temporal actions. In figure 13 the spatio-temporal action priors are critical to the estimate of the correct action. The person-centred action *Running* is not represented as often in the example database. Therefore if the priors for each person-centred action are computed on the basis of frequency then the ML spatio-temporal action for this sequence is *road, walking*. If however, the priors are uniform the ML result is as shown here. Note that in both cases (whether uniform or frequency-derived priors are used) the correct activity is still represented in the distribution over spatio-temporal actions. These examples highlight the difficulties in a scene such as this where occlusions occur and the scale of the person leads to increased ambiguity in the person-centred action recognition stage. Moreover, this scene helps understand the limitations of the optic flow technique for action recognition. The mean detection rate over all test sequences is shown in table 4 but the detection-rate for test sequences beyond the north pedestrian crossing drops to 52.0% for the correct ML model and 79.1% for the correct model being found in the distribution. This is due to the fact that the resolution becomes significantly low resulting in ambiguous optic flow channels.

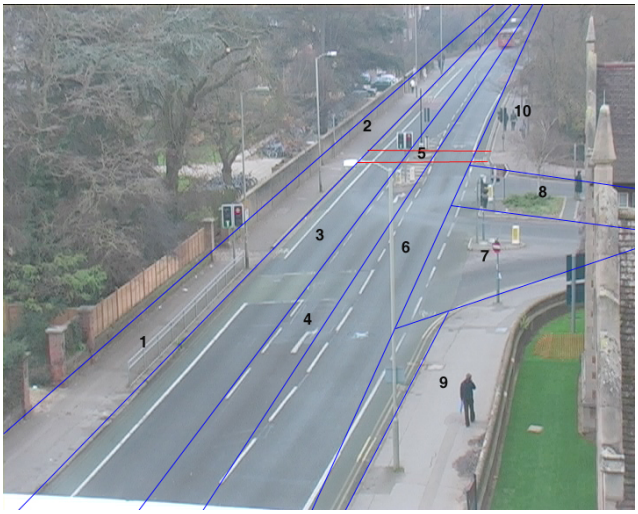
The statistics in table 4 give the mean true detection-rate computed by comparing the ground truth to the estimated ML model and also to how often the true model is found in the action distribution.

We further apply the technique to tennis video in order to classify each players' strokes and producing an automatic text commentary of an entire point. This



Frame	Activity	Likelihood
1 - 75	Walking on near-side pavement	0.34

Fig. 10. Neither the position nor the person-centred action of the person in this sequence is well-represented by the predefined exemplar data. Therefore the probability of the ML activity is significantly lower than that for the commentary of figure 9 i.e. 0.34 vs. 0.94.



Regions	**Person actions**	**Directions**
Northbound Lane (3)	Walking	North
Right Turn Lane (4)	Running	South
Southbound Lane (6)	Stopped	East
Parks Road Westbound (7)		West
Parks Road Eastbound (8)		Stopped
South-East pavement (9)		
South-West pavement (1)		
North-East pavement (10)		
North-West pavement (2)		
North pedestrian crossing (5)		

Fig. 11. This scene is divided into regions and labelled by an expert analyst. The labelled regions, person activities and directions for this scene are detailed in this table.

Frames	Estimated activity
380-427	NW-pavement, walking
428-486	N-ped-crossing, walking
487-511	*N-ped-crossing, stopped*
512-632	N-ped-crossing, walking
633-670	NE-pavement, walking

Fig. 12. The text commentary for a person crossing the road at a set of traffic lights. From frames 487 to 511 the traffic lights obscure the person. Tracking continues because feet are visible but the motion-type is incorrectly estimated.

Frames	Estimated activity
5330-5470	road, running

Fig. 13. The choice of priors on activity is critical when activity occurs which is not abnormal but is less frequently seen (e.g. running).

presents a significant challenge due to the rich set of person-centred actions (in this case, *strokes*) and the ambiguity due to both players. However, since the behaviour in tennis is well-bounded we can reliably extract exemplars of all the expected strokes. A human-readable commentary at the action (stroke) level is now possible since all known activity can be represented in the hand-labelled exemplars.

Following automatic tracking of players in video of 4 different professional tennis matches, we manually segmented the sequences into exemplars of standard tennis strokes and created independent databases of the position, velocity and person-centred action motion descriptors. We extract exemplars for the following strokes: *forehand*, *backhand*, *forehand-volley*, *backhand-volley*, *serve*, *smash*. In addition we provide examples of non-strokes labelled *running*, *walking* and *waiting-for-serve*. stroke example databases are created for each player i.e. facing the camera (farside court) and facing away from the camera (nearside) which significantly reduces ambiguity in the choice of person-centred action (a backhand by a player facing one direction is, motion-wise, very similar to a forehand from the other viewpoint). Taken with the labelled position

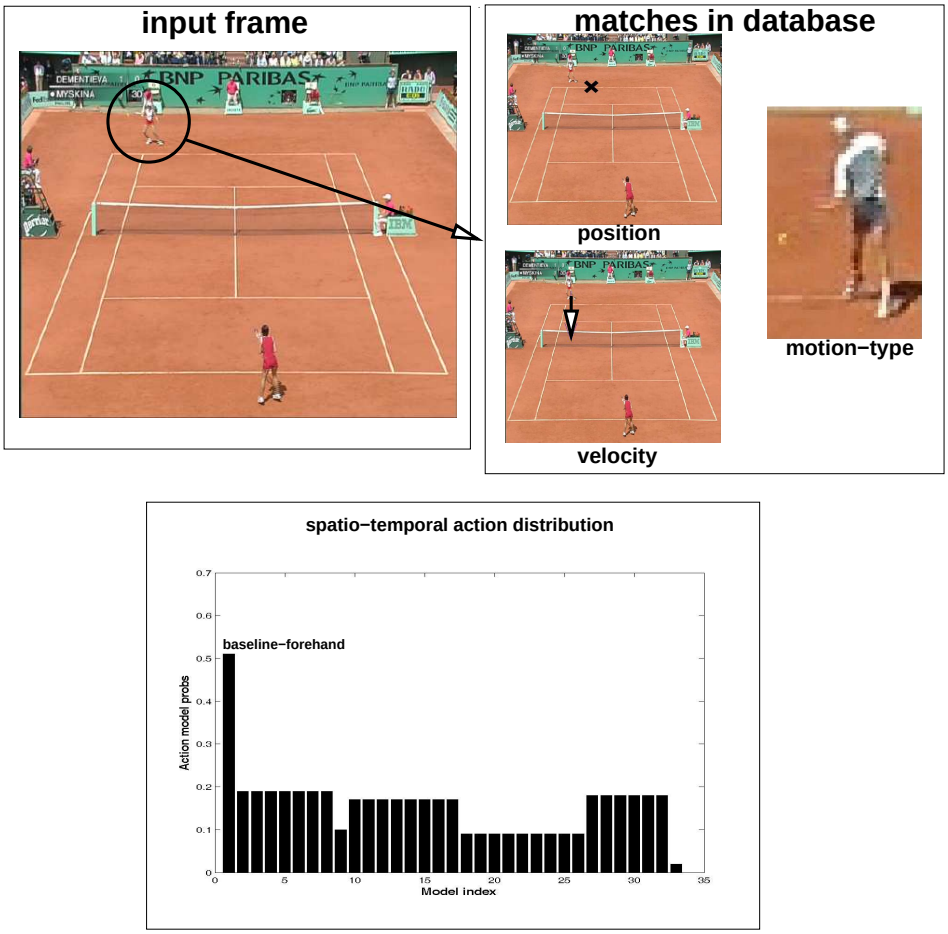
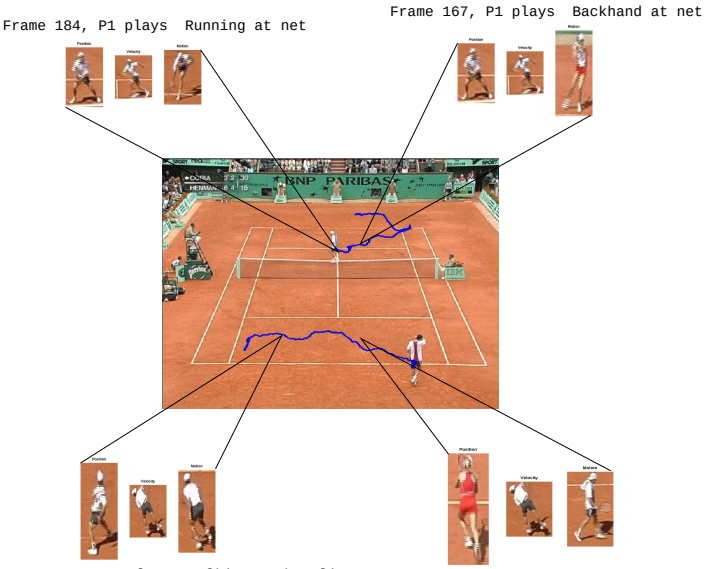


Fig. 14. Action recognition can be reliably achieved in tennis sequences using our method.

examples *baseline*, *mid-court*, *backcourt* and *net*, we have 33 possible actions for each player, including the null hypothesis. Testing is performed using previously unseen footage from a 5th match. Figure 14 shows spatio-temporal action-recognition in tennis video. There are 33 possible strokes resulting from combinations of positions and stroke-types in our exemplar set. The closest ML matches in the databases for this frame are shown next to the still image in the order position, velocity and stroke-type. The distribution over all strokes is shown in the graph. The most likely stroke is computed to be *baseline-forehand* which is correct. A commentary for an entire point is shown in figure 15.

5 Higher-level behaviour parameterisation

We now describe the final stage of our behaviour recognition system which is the encoding of behaviour as a sequence of actions. This section of the text relates to the top (pink) section of figure 1.



Frame	Estimated stroke	Likelihood	Ground truth
32	P2: *Baseline forehand*	0.41	Serve
39	P1: Walking at baseline	0.54	Walking at baseline
51	P1: *Running at baseline*	0.60	Running at mid-court
76	P2: Walking at baseline	0.36	Walking at baseline
99	P2: Walking at mid-court	0.45	Walking at mid-court
132	P2: Backcourt forehand	0.66	Backcourt forehand
148	P1: *Forehand at net*	0.34	attempted forehand

Fig. 15. A text commentary for selected frames of this tennis play. Player 1 ("P1") is at the far-court, Player 2 ("P2") at the near-court. Where the estimated stroke deviates from the ground truth it is marked in italics.

At each time step we have computed the most likely action. The sequence of actions and their likelihoods over a number of time steps is used to find the most likely behaviour by computing the likelihoods of predefined behaviour HMMs explaining the current action sequence. These HMMs are learned from an "ideal" example which has been automatically tracked and labelled at the action-recognition stage. We use a likelihood ratio to manually compare competing behaviour models. The likelihood ratio for comparing two hypotheses H and H' is computed as $LR = 2(\log(p(H)) - \log(p(H')))$, which has a chi-squared distribution parameterised by the difference in the model order. If LR is greater than the 95% confidence value of the chi-squared distribution for $\delta = |O(H) - O(H')|$, the result is statistically significant.

Serve-volley player	Baseliner player
Service	Waiting at baseline
Waiting at backcourt	Backcourt forehand
Backcourt backhand	Running at backcourt
Walking at baseline	Backcourt stroke
Mid-court Backhand	Walking at backcourt
Running at net	Baseline smash
Net volley	Running at backcourt
Running at net	Baseline backhand
Net volley	Waiting at baseline
Running at net	Backcourt smash
Net volley	[END POINT]

Table 3
A simulated play between a baseliner player and a serve-and-volley player using the respective HMMs for the tennis player types is shown in this table.

5.1 Improving the commentary using known player-types

In this scenario, domain knowledge can be used to improve the ML estimates shown in figure 15. The series of expected stroke *types* is well-established: a serve starts a point, a stroke is followed by a non-stroke period while the opposing player returns etc. We can smooth the stroke commentary using an HMM which encodes the "rules", that is, the anticipated ordering of strokes given a certain player-type, which can be defined by an expert.

In our tennis case-study we use an HMM loosely to encode the "rules" of the match: a *serve* starts each point, that a stroke exists for a typical number of frames, that position on the court must go through physically possible transitions (mid-court is *en route* to the net from the baseline) and that a non-stroke always follows a stroke (and *vice versa*). In fact an HMM such as this, built using "expert knowledge" can also form the basis of a generic tennis playing agent. By further training on observations of specific players one could train the HMM to model more specific player characteristics (such as baseliner or serve-volley). This HMM effectively acts as a smoothing prior, ensuring that invalid stroke transitions are penalised and that a maximum *a posteriori* action sequence results. HMMs are also generative models and, in table 3, we show examples of realistic tennis stroke sequences which have been generated.

An example of the results of the smoothing process is shown in figure 16 with the smoothed commentary provided as a text output at the bottom of the figure. Figure 17 shows in detail the improvement for the serving player.

Sequence	% detection, ML model correct	% detection, true model in distribution
Urban street	96.7	100.0
Junction surveillance	74.0	89.5
Tennis (no smoothing prior)	59.4	88.8
Tennis (with smoothing prior)	81.9	-

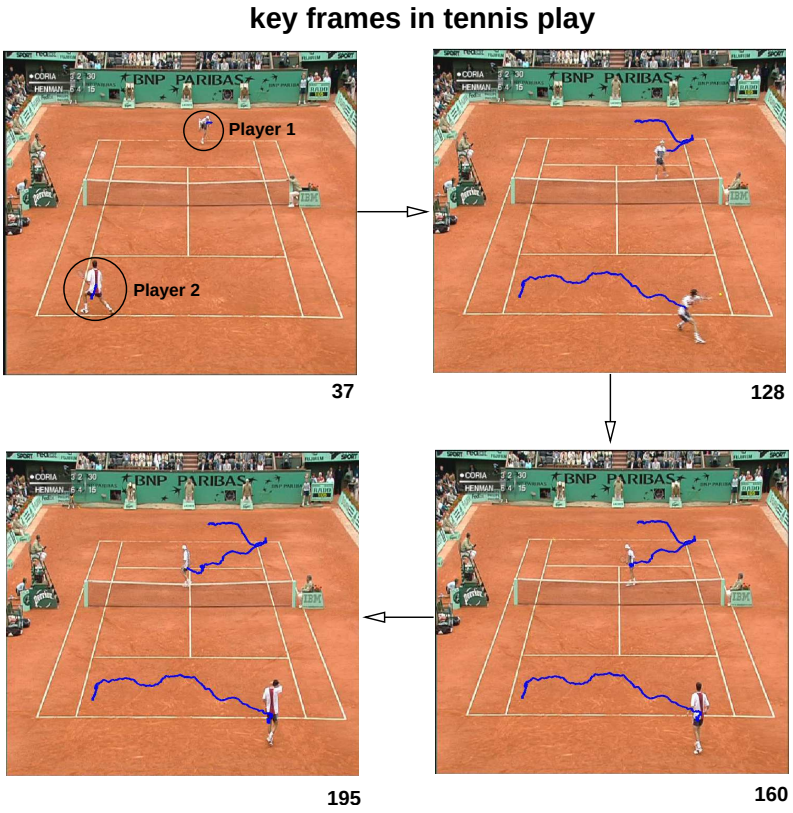
Table 4
The detection rates for the three sequences used in this paper. Note: only the ML sequence is used as the input to the smoothing stage (last row), hence there is no distribution figure.)

5.2 Overall play type estimation in tennis matches

A play is represented by a sequence of strokes from both players. Two HMMs are created to represent types of play, *baseline-rally* and *serve-and-volley*, from ideal, hand-selected action sequences. As the play unfolds in a new video sequence the HMM play model which best explains the sequence of strokes is automatically chosen. An example of this overall play estimate is shown at the bottom of figure 16.

6 Conclusions

In this paper a general method for action recognition from medium-resolution video is reported. The particular features we have chosen to use to construct a feature-level description are easy to obtain and photometrically invariant, but one is certainly not limited to these features. The inclusion of a description of local motion raised three issues: 1. searching a large database effectively; 2. ensuring temporal consistency of model choice when the example data is sparse; 3. combining independent descriptions of action in a principled way to describe action and behaviour. We combined, extended and improved disparate ideas from the literature for each of these problems in a novel way and the results demonstrated the efficacy of these solutions. We showed that by creating a framework for the propagation of uncertain information in a principled fashion coupled with a method for incorporating expert domain knowledge it is possible to classify human action non-parametrically and deal with ambiguity. Where the goal is to explain, at a high level, human behaviour in video, the use of compact behaviour HMMs which model behaviour as a sequence of actions allows for a rich description of behaviour which could be a significant component of a system for high-level reasoning.



Frame	stroke
1 - 49	Player 1 Service
1 - 18	Player 2 Waiting at backcourt
19 - 41	Player 2 Baseline backhand
50 - 70	Player 1 Walking at net
81 - 113	Player 1 Backhand at net
42 - 91	Player 2 Walking at baseline
92 - 134	Player 2 *Baseline backhand*
114 - 122	Player 1 Walking at net
123 - 140	Player 1 Backhand at net
135 - 200	Player 2 Waiting at backcourt
141 - 146	Player 1 Walking at net
147 - 155	Player 1 Backhand at net
Overall play	Serve-and-volley

Fig. 16. The estimated stroke sequence is smoothed using an HMM which encodes expert knowledge about tennis stroke sequences. In the commentary the misclassified strokes are shown in italics.

7 Further work

Although we have demonstrated the system with application to a video annotation system, we could equally apply the techniques to abnormality detection.

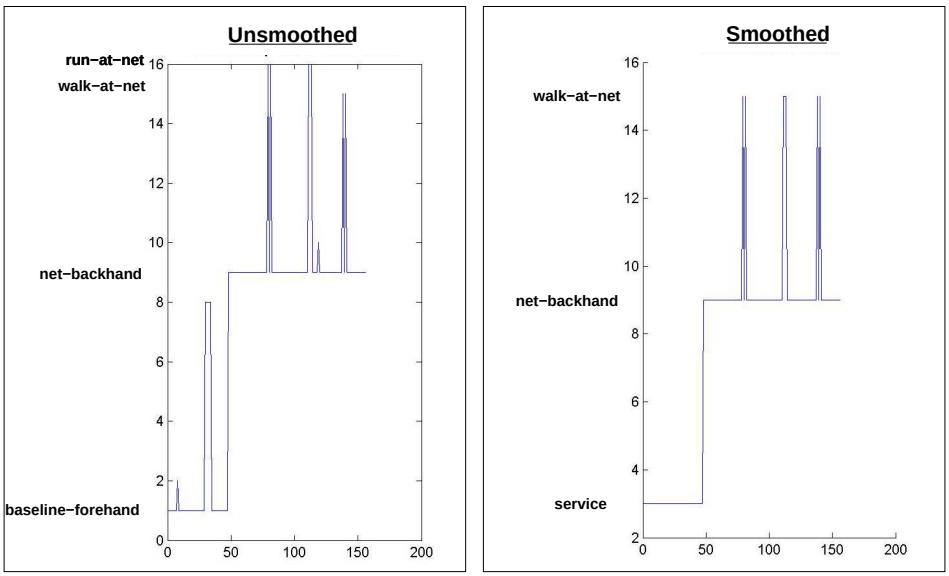


Fig. 17. Player 1 in figure 16 is known to be serving and the HMM for a serving player is used to smooth the stroke sequence. The improvements can be seen by comparing the unsmoothed (*left*) and smoothed (*right*) sequences; in particular the serve is no longer omitted and the expected stroke to non-stroke transition is observed.

Video annotation and/or novelty detection are simply means to a grander goal of developing a system which can *explain* what is being observed, not simply *detect* what has been previously observed. A future area of research which could yield significant results for intelligent surveillance is to extend the information extracted from video to include, for example, gaze direction [21,22] and to develop reasoning engines to explain human activity in surveillance and sports video.

The set of descriptions which were specified in each of the application domains on which we demonstrated our methods, were defined by a person with detailed, but not professional, knowledge of the scene. A researcher's language for describing human activity will not necessarily seem realistic to a true expert in the domain. In the military or law-enforcement context the researcher's descriptive language may well be misunderstood. This could create significant problems within the chain-of-command leading to operational failure. Therefore, it is critical that researchers seeking to develop systems involve the user in this phase of development to avoid such ambiguity.

Moreover, an interesting research topic will be to explore the possibility of defining a robust surveillance ontology for general use in urban environments.

Acknowledgements

We thank Mike Brady for his contribution to this work. Neil Robertson is an Industrial Fellow of the Royal Commission for the Exhibition of 1851 (`www.royalcommission1851.org.uk`).

Thanks also to the reviewers for their constructive comments on the draft of this paper.

References

[1] J.L. Barron, D.J. Fleet, S.S. Beauchemin *Performance of Optical Flow Techniques* International Journal of Computer Vision 12:1 pp. 43-77, 1994

[2] J. Black, D. Makris, T.J. Ellis *Validation of Blind Region Learning and Tracking* Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, Beijing, China, October, 2005

[3] O. Boiman and M. Irani *Detecting Irregularities in Images and in Video* IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005

[4] M. Brand and V. Kettnaker *Discovery and Segmentation of Actions in Video* IEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, August 2000

[5] C. Bregler *Learning and Recognizing Human Dynamics in Video Sequences* Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Juan, Puerto Rico, June 1997

[6] H. Buxton *Learning and Understanding Dynamic Scene Activity* ECCV Generative Model Based Vision Workshop, Copenhagen, Denmark, 2002

[7] D. Comaniciu and P. Meer *Mean Shift Analysis and Applications* Proceedings of the International Conference on Computer Vision-Volume 2, p.1197, September 20-25, 1999

[8] A.A. Efros, A. Berg, G. Mori and J. Malik *Recognising Action at a Distance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003

[9] A. Galata, N. Johnson, D. Hogg *Learning Behaviour Models of Human Activities* British Machine Vision Conference, 1999

[10] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. *Using Adaptive Tracking to Classify and Monitor Activities in a Site* Computer Vision and Pattern Recognition, June 23-25, 1998, Santa Barbara, CA, USA

[11] N. Johnson and D. Hogg. *Learning the Distribution of Object Trajectories for Event Recognition* Proc. British Machine Vision Conference, volume 2, pages 583-592, September 1995

[12] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. McFarland, and B. Temelkuran *Omnibase: Uniform Access to Heterogeneous Data for Question Answering* Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems, June 2002, Stockholm, Sweden

[13] B. Katz, J. Lin, C. Stauffer, E. Grimson *Answering Questions about Moving Objects in Surveillance Videos* AAAI Spring Symposium on New Directions in Question Answering, March 2003

[14] Y. Ke, R. Sukthankar and M. Hebert *Efficient Visual Event Detection using Volumetric Features* Proc. IEEE International Conf. on Computer Vision

[15] B.D. Lucas and T. Kanade *An Iterative Image Registration Technique with Application to Stereo Vision* DARPA Image Understanding Workshop, 1981.

[16] D. Makris and T.Ellis *Spatial and Probabilistic Modelling of Pedestrian Behaviour* British Machine Vision Conference 2002, vol.2, pp. 557-566, Cardiff, UK, September 2-5, 2002

[17] D. Makris, T.J. Ellis *Learning Semantic Scene Models from Observing Activity in Visual Surveillanc* IEEE Transactions on Systems Man and Cybernetics - Part B 35(3) June, pp. 397-408. ISBN/ISSN 1083-4419, 2005

[18] F. Porikli and T. Haga *Event Detection by Eigenvector Decomposition Using Object and Frame Features* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004

[19] J.R. Renno, D. Makris, T.J. Ellis, G.A. Jones *Application and Evaluation of Colour Constancy in Visual Surveillance* Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, VS-PETS Beijing, China, October, 2005

[20] M. Rigolli and M. Brady *Towards a Behavioural Traffic Monitoring System* Autonomous Agents and Multi-agent Systems, Utrecht, Netherlands, July 2005

[21] N.M. Robertson, I.D. Reid and J.M. Brady *What are you looking at? Gaze recognition in medium-scale images* Proc. Human Activity Recognition and Modelling , British Machine Vision Conference (BMVC), Oxford, UK, September 2005

[22] N.M. Robertson and I.D. Reid *Estimating Gaze Direction from Low-Resolution Faces in Video* Proc. 9th European Conference on Computer Vision (ECCV), Graz, Austria, May 2006

[23] H. Sidenbladh M. Black, L. Sigal. *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking* European Conference on Computer Vision, Copenhagen, Denmark, June 2002

[24] L. Tarassenko, A. Nairac, N. Townsend, I. Buston and P. Cowley. *Novelty Detection for the Identification of Abnormalities* International Journal of Systems Science, 11, 1427-1439 (2000)

[25] C.P. Town *Ontology-driven Bayesian Networks for Dynamic Scene Understanding* Proc. International Workshop on Detection and Recognition of Events in Video (at CVPR04), 2004

[26] P. Viola, M. Jones, D. Snow *Detecting Pedestrians using Patterns of Motion and Appearance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003

[27] T. Xiang and S. Gong *Video behaviour profiling and abnormality detection without manual labelling* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005

[28] T. Xiang and S. Gong *Visual learning given sparse data of unknown complexity* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005

[29] L. Zelnik-Manor and M. Irani *Event-Based Video Analysis* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), December 2001

[30] H. Zhong, J. Shi and M. Visontai *Detecting Unusual Activity in Video* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004