Behaviour understanding in video: a combined method

Neil Robertson [†‡]

[‡]QinetiQ
E206
St Andrews Road
Malvern, WR14 3PS, UK

Ian Reid[†]

[†]Oxford University
Dept Engineering Science
Parks Road
Oxford, OX1 3PJ, 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. A HMM which encodes the rules of the scene is used to smooth sequences of actions. High-level behaviour recognition is achieved by computing the likelihood that a set of predefined Hidden Markov Models explains the current action sequence. Thus, human actions and behaviour are represented using a hierarchy of abstraction: from simple actions, to actions with spatio-temporal context, to action sequences and finally general behaviours. While the upper levels all use (parametric) Bayes 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. In this paper we demonstrate the results chiefly on broadcast tennis sequences for automated video annotation.

1. Introduction

In a system for high-level visual scene understanding, the role played by humans in the scene is almost certainly of paramount importance. In particular, 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 non-parametric learning and classification technique for actions, can be combined with a simple, yet effective, parametric representations of action sequences, which we use to describe behaviours.

The lowest level of our system, for recognising simple actions (e.g. *walking* versus *running*, versus *standing*) is based on the technique described by Efros *et al.* [4] who showed how action recognition can be structured as a search over a comprehensive training database. Though 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. To this end we show how a simple extension to their "blurry motion channel" descriptor can effectively disambiguate between types of action even though the intra-sequence description of each frame of different actions are very similar.

Efros *et al.* deliberately used position independent descriptors, and made no attempt to reason at a higher level about the actions. 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. To this end we consider position and velocity information as additional features; these too are compared against a training database to elicit (respectively) qualitative position and direction labels. In a simple 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 simple 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 a hidden markov model which encodes the known "rules" of the scene then a maximum a posteriori 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 [1][10][7] 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 less training data is required for the HMMs, or indeed they are sufficiently simple that they can be modelled manually using "expert" knowledge.

In summary we make the following contributions:

- Recent results in data-driven human action recognition [4] have been extended: a concatenated local motion descriptor gives more effective discrimination in smaller datasets by improving temporal context,
- By representing position and velocity, in addition to local motion, spatial context is given which is important for higher level reasoning,
- Inspired by Sidenbladh's [13] 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 nearest-neighbour) and additionally by including a stochastic element to the search we can easily obtain a likelihood distribution over possible actions,
- The use of a Bayes net for fusion of non-parametric database search results for action recognition
- Smoothing of action sequences using a 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, with representation and recognition of these is achieved via HMMs. Human level descriptions are achieved by abstracting the actions as a precursor.

The remainder of the paper is structured as follows. We begin with a review of relevant prior art,

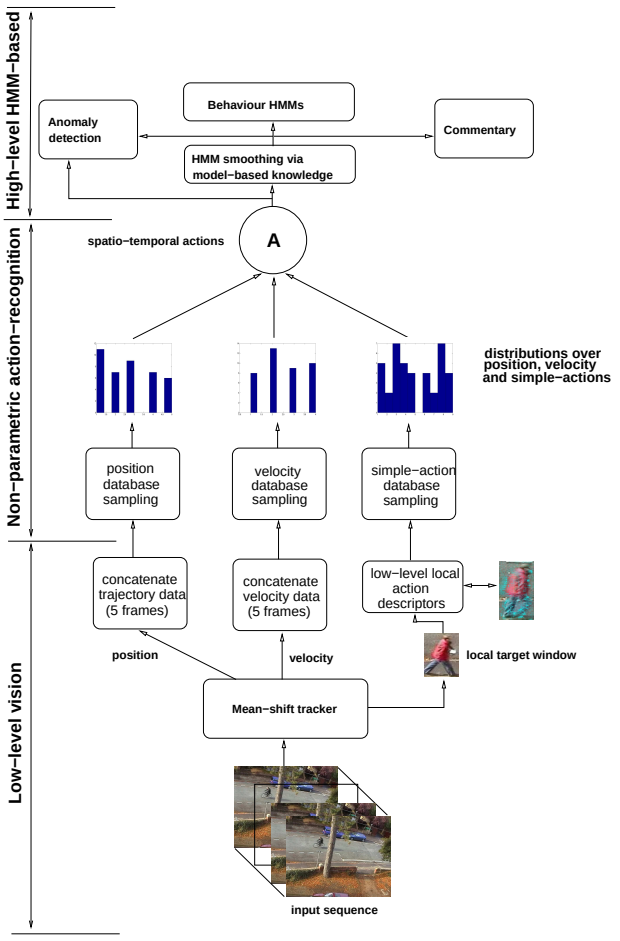


Figure 1. This schematic diagram illustrates the relationship between image features, actions, action sequences and the high-level parameterisation of behaviour. Databases of the position, velocity and motion-descriptor features are prepared in advance and are hand-labelled with qualitative descriptions of place, direction and simple-action. Distributions over each of these features are computed via non-parametric sampling of the databases. These distributions are combined using a simple Bayes Net which produces a distribution over spatio-temporal actions. This provides a text commentary of observed activity. Sequences of actions are also encoded as HMMs allowing higher-level descriptions of overall activity to be inferred. These HMMs are encoded using the spatio-temporal actions and not directly from image data.

then turn to a more detailed description of each of the stages of our algorithm. Section 2.2 deals with the low-level non-parametric action recognition stage, and describes in particular how we have implemented an efficient probabilistic search of an exemplar training database in order to sample from the action (and qualitative position and direction) distribution(s). Sections 2.3-2.4 describe the Bayes Networks that fuse the low-level data, smooth the action sequences and finally infer high-level behaviour. Section three gives experimental results and we conclude in section four. Throughout the paper we use sequences from either a simple urban surveillance scenario or sports footage. In our examples we assume the urban data represents on of a small set of simple actions such as *walking, running, standing, dithering* and a reasonable range of qualitative positions i.e. *nearside-pavement, road, driveway, farside-pavement* and directions i.e. *left-to-right, across* etc. This set of sequences is used to test the simple-action matching and action recognition steps. A richer set of simple-actions is found in tennis. Using our method we show that an intermediate representation of action can provide an automatic commentary. This commentary can be improved by smoothing the action sequences using an HMM which encodes expert knowledge about shot transitions e.g. that a serve starts a point and that a non-shot (e.g. *running*) follows a shot.

1.1. Previous Work

There has been much reported in the recent literature about methods for training recognition systems using large training data sets (e.g. [18]). Recently Zhong *et al* [20] demonstrated detecting unusual activity by classifying motion and colour histograms into prototypes and using the distance from the clusters as a measure of novelty. Sidenbladh and Black have shown that a comprehensive example set of joint angles can be used to aid human body tracking [13]. Also Zelnik-Manor and Irani [19] used a distance metric to identify examples of actions in video. Of most direct relevance is the recent work of Efros *et al* [4] which demonstrated that the general actions of people at medium scale (around 30 pixels high) can be distinguished by representing the action as a set of blurry motion channels derived from the optical 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. The AI Lab at MIT has developed an entirely automated system for visual surveillance and monitoring of an urban site [7] but does not attempt to *explain* observed behaviour.

A number of parametric methods have been formulated for recognising action. Brand and Kettnaker use HMMs for this purpose [1]. Buxton has used Bayesian networks for visual surveillance [2] as has Town [17]. Morellas *et al* [11] show that they can automatically evaluate the threat posed by observed activity using a complete, real-time system deployed in environments such as car parks and oil pipelines. Makris [10] also uses HMMs for detailed modelling of trajectories from learned geometric route data. Porikli and Haga [12] include object-based and frame-based features, parameterised by an HMM. Galata, Johnson and Hogg [5] [8] use Vector Quantisation (VQ) to group and classify trajectory data. ([8] is a notable attempt to introduce the concept of action and behaviour into classification systems.) While the parametric approaches demonstrate 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 sequences of action i.e. behaviour (e.g. *crossing-the-road*).

2. Action and behaviour recognition

The main components of our behaviour recognition method are (i) action recognition via non-parametric matching of trajectory data and instantaneous motion descriptors, fused via a simple Bayes net; (ii) smoothing of the action recognition sequence using an HMM which encodes known rules for action transitions; (iii) behaviour classification using HMMs.

2.1. Target description

Using a standard mean shift tracking algorithm [3], we extract the following information for each target for each frame: position, velocity and a window around the target (see fig 1). 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*. A simple and effective method to do this was suggested by Efros *et al* [4]. In that work a local motion descriptor based on coarse optic flow is extracted from a target window. 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 match provides an action label for the current data. In our experiments we have found that if the database contains only a small number of examples of a certain action the

risk of the nearest-neighbour being incorrect is greatly increased. In order to add temporal context and mitigate against this type of confusion, we create a richer feature descriptor by concatenating the coarse motion descriptors from a number of consecutive frames, typically 5, to form a motion feature vector at each frame. An example showing the benefits of this enhancement is shown in figures 3 and 4. Efros *et al* deliberately discarded all positional information. In contrast we have argued in section 1 that such information is important in placing an action in its spatial context. 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 (respectively velocity) data, and the database examplars are labelled with qualitative position (respectively, qualitative direction) labels. 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 simple Bayes net described in Section 2.3. Prior to that, we discuss the database organisation and search techniques. This 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; (ii) for more effective data fusion (and necessarily for appropriate use of a Bayes net) we do not simply want a nearest-neighbour (i.e. maximum likelihood) match, but rather a distribution over possible matches.

2.2. Database creation and search

In [13] 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, however, applies equally well to our application of image feature data and the pseudo-random search algorithm is identical to that derived in [13].

Significantly, the first $b = log_2(n)$ (where n is the number of time intervals in the training data) components are organised into a binary tree the nodes of which are split on the basis of the sign of the components $\mathbf{c_i} = [c_{i,1}, \ldots, c_{i,b}]$. The search of the tree is randomised by the inclusion of a random perturbation of the traversal of the tree drawn from a Gaussian distribution. At the leaf nodes a linear search takes place if there is more than one match. The probability of these matches is computed on the basis of how "close" the match in the database is to the input i.e. $p(match|input) = \exp -(\frac{|match-input|}{\sigma})^2$. This search method is used for two reasons: it is more efficient 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 matches of position, velocity and motion-descriptor into frames of the previously seen examples. An example of such a distribution is shown in figure 2. The database was created using 60 minutes of automatically tracked (but hand-labelled) data, and was tested using novel sequences of similar actions.

2.3. Action likelihood computation

A simple-action we define as a target-centred action such as *walking*. This can be estimated by sampling from the motion-descriptor database alone. 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 (trivially) simple Bayes Net to effect this information fusion: if the spatio-temporal action is denoted a, x is the qualitative position, v is the qualitative direction, and m is the simple action, then assuming conditional independence yields $p(a, x, v, m) = p(a)p(x|a)p(v|a)p(m|a)$. The distributions $p(x_m|x_i)$, $p(v_m|v_i)$ and $p(m_m|m_i)$ 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), $p(d|a) = \frac{p(a|d)p(d)}{p(a)}$. $p(a|d)$ 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. Figures 6 and 5 illustrates this process for two different applications. Figure 5 highlights the significance of each input for successful action classification.

2.4. Action sequences

Since the behaviour in tennis is well-bounded we can reliably extract exemplars of all the expected shots. A commentary at the action (shot) and behaviour (play) level should then be possible since all known activity

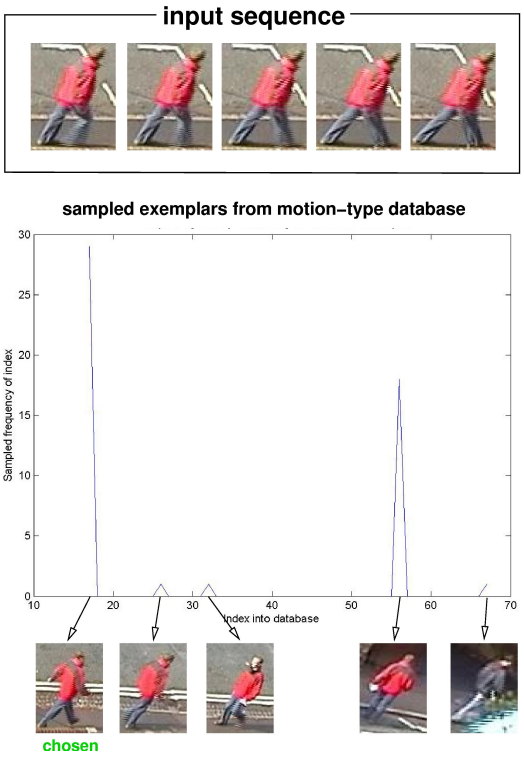


Figure 2. Database searching results. The top row shows the input figures from which the optic-flow motion channels are computed. The sampling of the database (which is represented as a binary tree search of Principal Components) is shown in the graph (*bottom*) with the exemplars at the leaf nodes superimposed. As we expect the most likely simple-action is indeed *walking* (peak on the left) but this is not unambiguous.

Figure 3. Matching optical flow based motion descriptors without large volumes of representative data sets can result in ambiguous matches as shown here. For each pair the input is shown on the left and the ML exemplar from the sampling of the motion-descriptor database is shown on the right (see figure 2).

Figure 4. 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 manifested by the fact that the motion does not reverse in this case (c.f. figure 3).

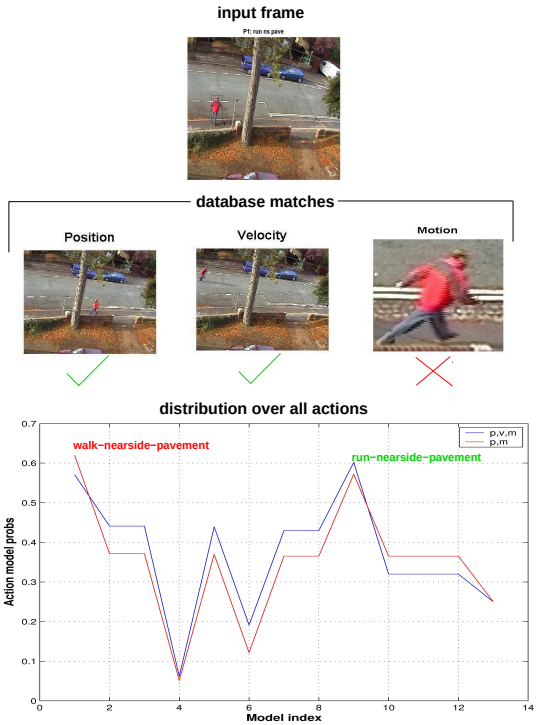


Figure 5. Velocity and motion-type are as important as position for action-recognition. Here the ML motion-type is (incorrectly) classified as *walking*. When the resulting distributions from each of the inputs (i.e. position, velocity and motion-type) are fused the ML estimate is now (correctly) *running-on-nearside-pavement*. The action probability distribution is shown here when velocity is excluded (red) and included (blue).

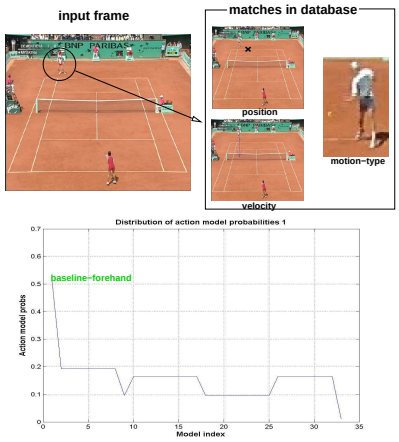


Figure 6. There are 33 possible shots resulting from combinations of positions and shot-types in our exemplar set. The closest ML matches in the databases for this frame are shown next to the still image in the order position, velocity and shot-type. The distribution over all shots is shown in the graph. The most likely shot is computed to be *baseline-forehand* **which is correct.**

is represented in our hand-labelled model. Since the series of expected shot *types* is well-established (e.g. a serve starts a point, a shot is followed by a non-shot period while the opposing player returns etc.) we smooth the shot commentary using a HMM which encodes the rules. Results of shot-matching and the resulting commentary are shown in figures 7.

2.5. Behaviour parameterisation

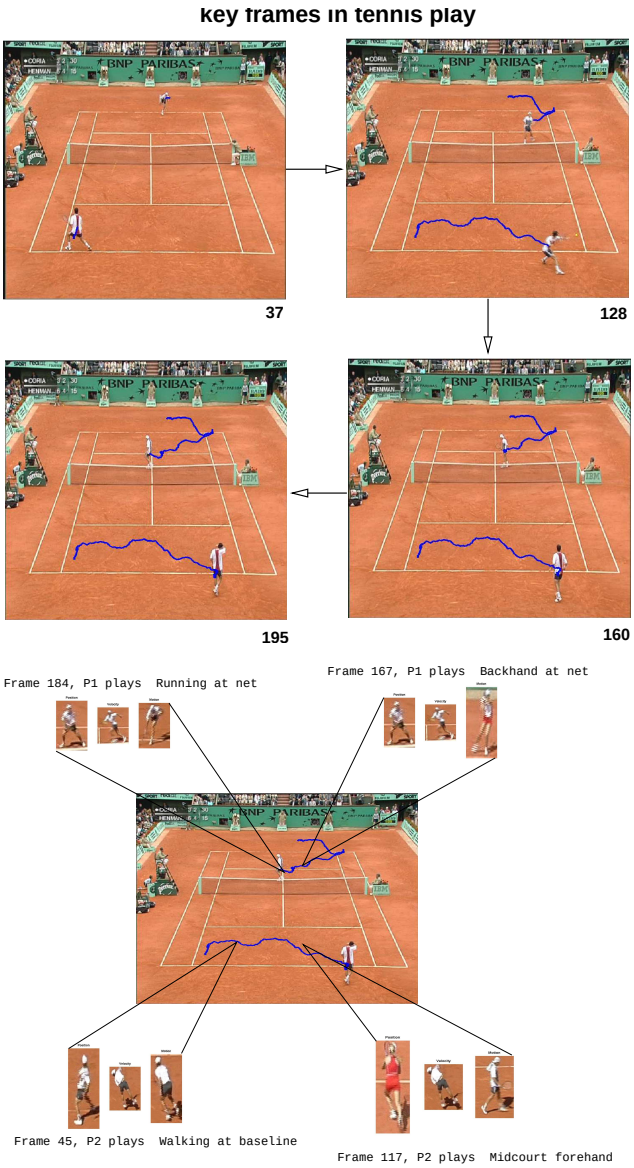
At each time step then we have computed the most likely action. The sequence of actions and their likelihoods over a number of time steps is used to find the most likely behaviour by computing the likelihoods of predefined behaviour HMMs (see [6]) explaining the current action sequence. These HMMs are learned from an "ideal" example which has been automatically tracked and labelled. 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 the result is statistically significant. An example of this high-level classification is shown in figure 7.

3. Experiments and results

We apply the technique to tennis video in order to classify each players' shots and producing an automatic text commentary. This presents a significant challenge due to the rich set of simple actions and the ambiguity due to both players. Following automatic tracking of players in video of 4 different professional tennis matches, we manually segmented the sequences into a exemplars of standard tennis shots and created independent databases of the position, velocity and simple-action motion descriptors. The shots we extract exemplars for are labelled with the following qualitative descriptions: *forehand*, *backhand*, *forehand-volley*, *backhand-volley*, *serve*, *smash*. In addition we provide examples of non-shots labelled *running*, *walking* and *waiting-for-serve*. Shot 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 simple-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 examples *baseline*, *midcourt*, *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 involving two previously unused players. Figure 6 shows an example of the spatio-temporal action selection performed by the first two levels of our system. Note that although the figure shows the maximum likelihood estimate, the system in fact retains a distribution over possible spatio-temporal actions.

3.1. Tennis commentary

A simple commentary can be obtained from the first two levels of our system by simply selecting the ML action at each instant. This however neglects that in many scenarios domain knowledge can be used to improve these estimates. In our tennis case-study we use a hidden markov model loosely to encode the "rules" of engagement: a *serve* starts each point, that a shot exists for a typical number of frames, that position on the court must go through physically possible transitions (midcourt is *en route* to the net from the baseline) and that a non-shot always follows a shot (and vice-versa). This HMM effectively acts as a smoothing prior, ensuring that invalid shot transitions are penalised and that a maximum a posteriori action sequence results. An example of this process is shown in figure 7 with the smoothed commentary provided as a text output at the bottom of the figure. A play is represented by a



Frame	Shot
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

Figure 7. The text commentary automatically produced from this tennis play is shown below the figures and the matches from the position, velocity and motion databases for critical points in the sequence. The estimated shot sequence is smoothed using an HMM which encodes expert knowledge about tennis shot sequences. In the commentary the misclassified shots are shown in italics.

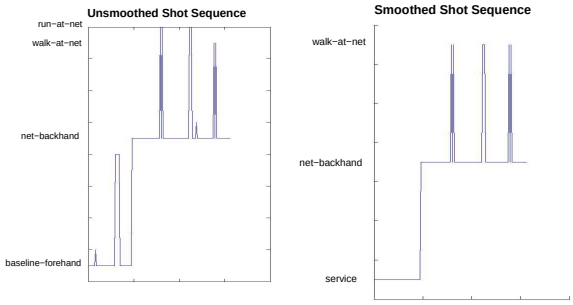


Figure 8. Smoothing the shot sequence which arises from the spatio-temporal action-recognition phase (see figure 1) provides consistency across the shot choice and allows important expert knowledge to refine the shot selection. In this example here the player (which is player 1 in figure 7) is known to be serving and HMM for a serving player is used to smooth the shot 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 shot to non-shot transition is observed.

sequence of shots 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 we choose the HMM play model which best explains the sequence of shots.

4. Conclusions

In this paper a method for action recognition 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 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. Though we have demonstrated the system with application to video annotation system, we could equally apply the techniques to abnormality detection. 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.

Acknowledgments

We thank Mike Brady for his contribution to this work. Neil Robertson is supported by an Industrial Fellowship from the Royal Commission for the Exhibition for 1851.

References

[1] 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

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

[3] 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

[4] 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

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

[6] Z. Ghahramani *Learning Dynamic Bayesian Networks* In C.L. Giles and M. Gori (eds.), Adaptive Processing of Sequences and Data Structures . Lecture Notes in Artificial Intelligence, 168-197. Berlin: Springer-Verlag

[7] 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

[8] 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

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

[10] 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

[11] V.Morellas, I.Pavlidis, P.Tsiamyrtzis *DETER: Detection of Events for Threat Evaluation and Recognition* Machine Vision and Applications, 15(1):29-46, October 2003

[12] 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

[13] 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

[14] N.T. Siebel *Fusion of Multiple Tracking Algorithms for Robust People Tracking* European Conference on Computer Vision, May 2002, Copenhagen, Denmark.

[15] C. Stauffer, E. Grimson *Learning Patterns of Activity Using Real-Time Tracking* Pattern Analysis and Machine Intelligence, 22(8):747-757, 2000

[16] 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)

[17] 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

[18] 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

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

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