Dynamic Events as Mixtures of Spatial and Temporal Features

Karteek Alahari* and C.V. Jawahar

Centre for Visual Information Technology,
International Institute of Information Technology,
Gachibowli, Hyderabad 500032, India
jawahar@iiit.ac.in

Abstract. Dynamic events comprise of spatiotemporal atomic units. In this paper we model them using a mixture model. Events are represented using a framework based on the Mixture of Factor Analyzers (MFA) model. It is to be noted that our framework is generic and is applicable for any mixture modelling scheme. The MFA, used to demonstrate the novelty of our approach, clusters events into spatially coherent mixtures in a low dimensional space. Based the observations that, (i) events comprise of varying degrees of spatial and temporal characteristics, and (ii) the number of mixtures determines the composition of these features, a method that incorporates models with varying number of mixtures is proposed. For a given event, the relative importance of each model component is estimated, thereby choosing the appropriate feature composition. The capabilities of the proposed framework are demonstrated with an application: recognition of events such as hand gestures, activities.

1 Introduction

Characterization of dynamic events, which are spatiotemporal in nature, has been a problem of great interest in the past few years [1,2,3,4,5,6]. Early methods employ segmentation and tracking of individual parts to model the dynamism in events [2,7]. They are based on identifying moving objects – typically referred to as blobs – constrained by their size or shape. Tracked trajectories of these blobs are used to distinguish events. Naturally, these methods are very sensitive to the quality of segmentation and tracking of blobs. A popular approach has been to represent the dynamism in events as image features [1,5,8]. Typically these approaches, of identifying a fixed feature set (or interesting regions), are applicable to a limited set of events. As observed by Sun *et al.* [9], techniques that learn an optimal set of features from the given event set are of much interest for real life applications. In today's scenario, wherein events can be captured as videos under different conditions, there is also a need to model the variations across videos in a probabilistic framework. Models such as Hidden Markov Models (HMMs) are popular to accomplish this [10]. However, these models fail to capture the events in a low dimensional space. Although there have been attempts to use dimen-

* Currently at Oxford Brookes University, UK.

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 540–551, 2006.

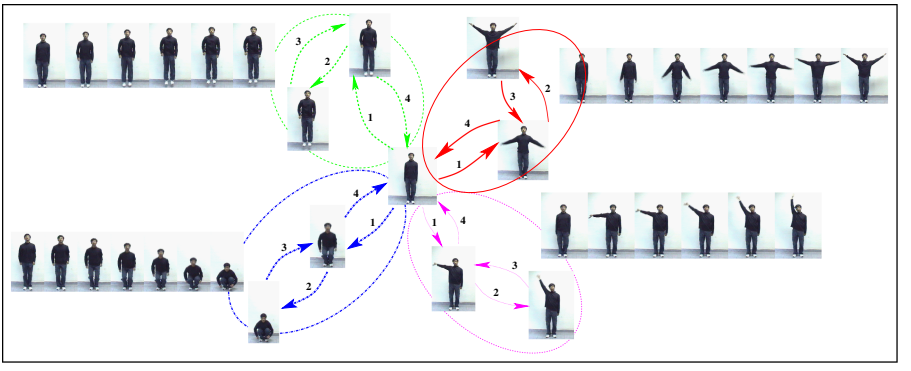

Fig. 1. A sample of events performed by humans (shown as image strips) and action representatives (shown as individual frames). A set of actions constitute an event. Four events and their corresponding actions are shown as distinct groups here (Green (Top Left) - *Jumping*, Red (Top Right) - *Flapping*, Blue (Bottom Left) - *Squatting*, Magenta (Bottom Right) - *Waving*). The arrows denote the temporal transitions between the actions and the number on each arrow denotes the temporal sequencing of the event. Note that the action 'standing' is common to all these events.

sionality reduction methods in combination with these models [9], they fail to be generic. Thus, to characterize events efficiently we need a representation that not only discards the acceptable statistical variability across multiple instances of an event, but also discriminates among different events.

We propose a method to learn a compact representation of events preserving their discriminatory characteristics. An event is modelled as a sequence of atomic spatiotemporal units called *actions*. Actions can be interpreted as subsequences from the event sequence. A probabilistic approach is employed to estimate the actions and the compositional rules for the events, in a low dimensional manifold. This is achieved using a Mixture of Factor Analyzers (MFA) model [11] combined with a probability transition matrix, which encodes the transitions among the action mixtures. The mixtures represent the actions while the transitions represent the compositional rules. In other words, the number of mixtures determines the composition of spatial and temporal features in events. Fixing the number of mixtures for the entire event set is not optimal, as the spatiotemporal characteristics vary among events. A unifying framework which incorporates models with varying number of mixtures (which form the model components) is proposed. For a given event, the relative importance of each model component is estimated from an example set.

The model is based on the observation that events comprise of more fundamental units, *actions*. Similar observations were made in the past in different ways [6,7,8,10,12]. Actions were represented as components of PCA [7], the hidden states of HMMs [10], key frames in the event video, canonical poses, *etc.* It has also been a common practice to analyze the event sequences in a window-based fashion [13] to capture the atomic characteristics in events. In

addition to this, we exploit the fact that most of the events have a large degree of overlap among them. This is evident in the form of common actions among various events. An example of this is shown in Figure 1 where the events share the action 'standing'. Furthermore, actions capture the spatial (or the appearance) features in events, while transitions among actions capture the temporal features. The main advantages of the model are: (a) It represents events in a low dimensional manifold retaining their discriminative characteristics, (b) It recognizes events in a real-time fashion, (c) It chooses the appropriate spatial and temporal feature extent by analyzing the event.

Section 2 presents an overview of the event recognition model. It also analyses the dependency of event recognition accuracy on the number of mixtures. Preliminary results on the CMU MoBo database [14] are also presented in this section. The method to combine model components to capture various degrees of appearance and temporal features is described in Section 3. In Section 4 results on human event and Sebastian Marcel Dynamic Hand Posture Database available at [15] are presented along with a discussion. Conclusions are presented in Section 5.

2 Events as Mixture of Actions

Events are represented as a mixture of actions and the transitions among these actions. The representation model consists of an MFA coupled with a probability transition matrix. MFA is essentially a reduced dimension mixture of Gaussians. The model learns action mixtures in a low dimensional space, *i.e.* it accomplishes the task of clustering and estimating a low dimensional representation simultaneously. There are two reasons that argue for action clustering in a subspace representation. Firstly, different actions may be correlated in different ways, and hence the dimensionality reduction metric needs to be different between action mixtures. Secondly, a low dimensional representation may provide better separated mixtures. We choose the MFA model to accomplish this task.

Let the total number of frames from examples of all the events be N and let x_t (of dimension d), $t = 1 \ldots N$, denote the t th frame. Subsequences of x_t form *actions*. For instance, if we consider the event Squatting (which consists of two distinct actions – standing and sitting), the initial few frames represent the action standing and the other frames represent the action sitting (refer Figure 2). The subsequent frames of an action are highly correlated and therefore, for each x_t, a p ($\ll d$) dimensional representation z_t exists. That is, x_t is modelled as $x_t = \Lambda_j z_t + u$, where Λ_j represents the transformation basis for the j th action and u is the associated noise. Multiple such subsequences, occurring across different events, are used to learn Λ_j for each action, and hence the corresponding low dimensional representation.

Consider a generative process for the ensemble of events based on the MFA model. An event, which is captured as a set of frames, is composed of various actions. A typical frame of the event, x_t, can be generated as follows. The action to which it belongs is chosen according to the discrete distribution $P(\omega_j)$,

Fig. 2. A few sample frames showing events performed by humans: Squatting (top row), Flapping (bottom row). Note the presence of a common *action* – Standing – between these events. The initial few frames of the event Squatting represent the action standing while the other frames represent the action sitting. The action standing also occurs in the initial few frames of the event Flapping.

$j = 1 \ldots m$. Depending on the chosen action, a continuous subspace representation z_t is generated according to $p(z_t|\omega_j)$. Having learnt z_t and action ω_j, the observation x_t is obtained according to the distribution $p(x_t|z_t, \omega_j)$, *i.e.* x_t is modelled as a "mixture model of actions" according to $p(x_t) = \sum_{j=1}^{m} \int p(x_t|z_t, \omega_j) p(z_t|\omega_j) P(\omega_j) dz_t$, where ω_j, $j = 1 \ldots m$, denotes the j th action. This is a reduced dimension mixture model where the m mixture components are the individual actions. The probability $p(x_t)$ describes the probability of generating a frame given the action which it belongs to, and its corresponding subspace representation. The generative process is to be inverted to learn the parameters of these distributions from the event sequences. This is achieved using an Expectation Maximization (EM) algorithm. It is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data has missing or unknown values [11]. In this case, the data corresponds to the frames, the unknown values to the low dimensional representations of these frames and the actions to which these frames are associated.

The EM algorithm alternates between inferring the expected values of hidden variables (subspace representation and actions) using observed data (frames), keeping the parameters fixed; and estimating the parameters underlying the distributions of the variables using the inferred values. All the event videos are represented as a sequence of frames and are used for estimating the parameters. The two phases of the EM algorithm – Inference and Learning – are executed sequentially and repeatedly till convergence. The E-step (Inference) proceeds by computing $E[\omega_j|x_t]$, $E[z_t|\omega_j, x_t]$ and $E[z_t z_t^T|\omega_j, x_t]$ for all frames t and actions ω_j [11]. In the M-step (Learning), the parameters π_j, Λ_j, μ_j and Ψ are computed.

During the E-step the following equations are used.

$$E[\omega_j z_t | x_t] = h_{tj}\beta_j(x_t - \mu_j)$$
$$E[\omega_j z_t z_t^T | x_t] = h_{tj}(I - \beta_j\Lambda_j + \Lambda_j(x_t - \mu_j)(x_t - \mu_j)^T\beta_j^T),$$

where $h_{tj} = E[\omega_j|x_t] = \pi_j\mathcal{N}(x_t - \mu_j, \Lambda_j\Lambda_j^T + \Psi)$, $\beta_j = \Lambda_j^T(\Lambda_j\Lambda_j^T)^{-1}$. The parameters μ_j, Λ_j, $j = 1 \ldots m$, denote the mean and the corresponding subspace bases of the mixture j respectively. The mixing proportions of actions in the event are denoted by π. The noise in the data is modelled as Ψ. The expectation h_{tj} can be interpreted as a measure of the membership of x_t in the j th action. Interested readers may derive the equations for M-step easily from [11].

Although the MFA model captures the spatial features as actions effectively, it does not account for the temporality in events. As shown by Veeraraghavan *et al.* [16] both spatial and temporal features are important for event recognition. This issue is addressed by modelling the dynamism in events as transitions across the learnt actions $\omega_1, \omega_2, \ldots, \omega_m$. The transition probabilities are computed by observing z_ts across the various actions for each event. After the EM algorithm converges, the action transition matrix $T_k = [\tau_{pq}^k]$, for each event k, is formed as follows.

$$\tau_{pq}^k = \sum_{t=1}^{N-1} [c_t = p][c_{t+1} = q] \qquad 1 \le p, q \le m, \tag{1}$$

where c_t denotes the class label of the frame x_t and is given by $c_t = \arg\max_j h_{tj}$; $j = 1 \ldots m$. Normalizing the entries in the transition matrix gives the corresponding probability transition matrix P_k. Thus, a compact representation of the events by automatically learning the m actions in a low dimensional manifold, and the sequencing information are obtained. The structure of the ensemble of events is contained in the parameters of the actions and the probability transition matrix, *i.e.* $\{(\mu_j, \Lambda_j)_{j=1}^m, \pi, \Psi\}, \{P_k\}_{k=1}^K$.

When recognizing events in a new video sequence, the learnt parameters are used to compute the action mixture (cluster) assignment, c_t for each frame x_t. Let $c_1, c_2, \ldots, c_{N_s}$, denote the action assignments for the respective frames of a N_s frame-long event sequence. The probability that the video frames belong to the k th event, S_k, is given by $S_k = \prod_{t=1}^{N_s-1} P_k[c_t][c_{t+1}]$. The video sequence is assigned to be the event k^*, which maximizes S_k.

This model is validated using the CMU MoBo database [14]. The frames of the video sequence are processed minimally before learning the event-set representation using the EM algorithm described above. The available background images are used to obtain the corresponding silhouette images. The silhouette images, represented as vectors, are used to learn the event representation. After the algorithm converges the sequence probabilities of all the events are computed. The transition probability of a new event video is estimated via the inference step of the EM algorithm, and is labelled following a maximum likelihood approach. Even though the four activities in the database (Slow walk, Fast walk, Incline walk, Walking with a ball) had subtle differences, an average accuracy of 85% is achieved. These results compete with, and also outperform, those reported in [16].

3 Combining Mixture of Actions Models

The relationship between the event recognition accuracy and the number of action mixtures is interesting. Varying the number of actions has minimal influence on the accuracy, beyond a certain limit. For instance, when recognising the event Flapping (of hands) it was observed that beyond 5 mixtures, the accuracy varied negligibly. Low accuracy is observed initially, when the number of actions is small, because the temporal characteristics of the event are not modelled. Similar behaviour was observed for all the events, except that the *optimal* number of actions varied with the event in consideration. Also each of these models captures different characteristics of the events. This argues for an integrated model which learns the appropriate number of actions for each event.

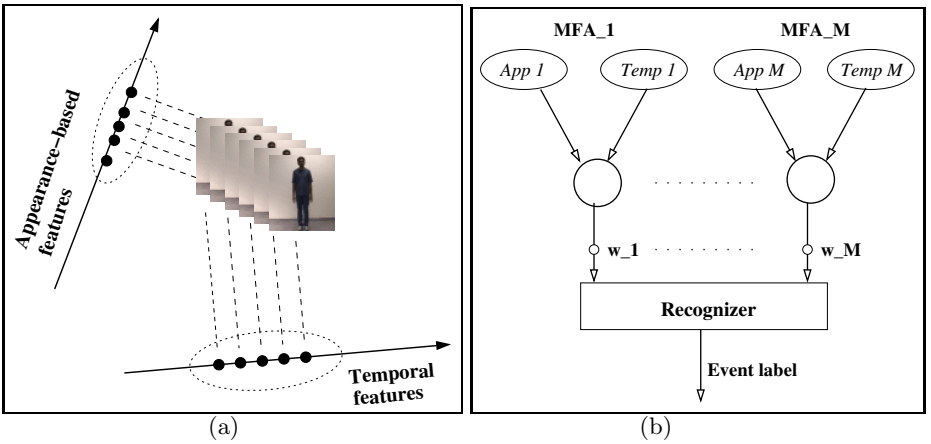

Fig. 3. (a) Event sequences consist of spatial (or appearance) and temporal features. (b) A summary of the proposed appearance and temporal feature integration model: A combination of MFAs (MFA_1 ... MFA_M) is used to have the model choose between appearance (App), temporal (Temp), which are the two extreme cases, and a combination of both features (say, MFA_i) adaptively. The contribution of each of these components in the decision making process is identified by its corresponding weight (w_i).

Varying the number of actions can also be interpreted as varying the appearance and temporal feature content in the event representation. The proposed adaptive scheme chooses the appropriate model component based on the event being recognized. The basic model, *i.e.* mixture of actions model with a transition matrix to capture the temporality in events, is replicated with different number of action mixtures in each of them (see Figure 3). The two ideal extreme cases in this framework are: modelling with (1) a single mixture for each event, and (2) a separate mixture for every frame of an event. In the training phase, the relevance of each component model is also estimated for all the events in the database.

Theoretically, one may define a single mixture for each frame in the event sequence. However, such a scheme is inefficient and impractical due to the possibly large number of transitions between these mixtures. The maximum number of action mixtures is typically decided by the nature of the data set, but is much lower than the total number of frames. Each mixture of actions model, $\mathcal{M}_i, i = 1 \ldots M$, is trained separately with the frames of all the events using an EM algorithm, as described in the previous section. By the end of the mixture model training phase, the parameters of the model – $\{(\mu_j, \Lambda_j)_{j=1}^m, \pi, \Psi\}$, $\{P_k\}_{k=1}^K$ are obtained for each model component.

3.1 Relevance of Each Component

Learning the event representation also involves estimating the relevance of all the component models for any event. This is estimated by optimizing an objective function defined over the training set of N video sequences. The objective function, $J(.)$ is given by

$$J(\Gamma) = \sum_{j=1}^{N} \sum_{i=1}^{M} (\gamma_{ij} d_{ij})^2,$$

where $\Gamma \in \mathbb{R}^{MN}$ is a matrix $[\gamma_{ij}]$. γ_{ij} denotes the contribution of the i th mixture of actions model component for the j th video sequence in the data set, and d_{ij} is the distance metric signifying the cost of recognizing the j th sample with the i th model component. The objective function is minimized over the space of γs. This is done by using Lagrange multipliers with the constraint $\sum_{i=1}^{M} \gamma_{ij} = 1$. The objective function J is formulated so as to minimize the recognition accuracy across all the component models. Given that each component model captures a new composition of temporal and spatial features, this framework provides a unifying scheme to describe events with different compositions of these features.

On observing that the relevance (or weights) for each event sequence are independent, the minimization can be done independently in each column. Thus, the Lagrangian is given by

$$\mathcal{J}(\lambda, \gamma_j) = \sum_{i=1}^{M} (\gamma_{ij} d_{ij})^2 - \lambda (\sum_{i=1}^{M} \gamma_{ij} - 1). \tag{2}$$

Differentiating Equation 2 with respect to γ_{pq}, $\gamma_{pq} = \lambda / 2 (d_{pq})^2$. Using this equation and the constraint $\sum_{r=1}^{M} \gamma_{rq} = 1$, γ_{pq} can be computed as

$$\gamma_{pq} = 1 \Big/ (d_{pq})^2 \sum_{r=1}^{M} (d_{rq})^2 \ . \tag{3}$$

Equation 3 provides a method to compute the relevance of component models, given the distance metrics d_{ij}. The distance metrics, in this case, are the

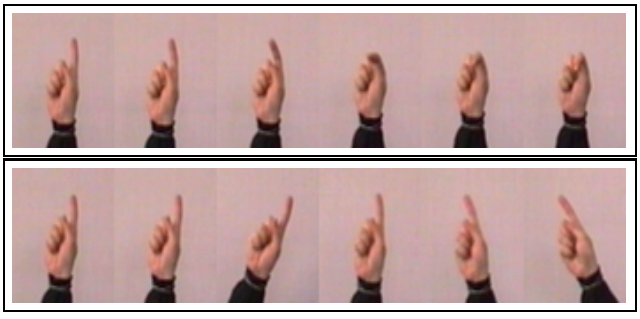

Fig. 4. A few sample frames showing hand gestures [15]: Click (top) and No (bottom)

likelihoods of the mixture of actions model component $\mathcal{M}_i$, which is the probability computed from the corresponding transition matrix. Metrics based on other models such as HMM, SVM, NN, *etc.*, can also be incorporated. Although the framework is generic, we limit the discussion to our mixture of actions model.

3.2 Weighted Measure to Recognize Events

Once the weights $[\gamma_{ij}]$ are identified for all the events, they are used in the recognition framework. Given an un-trained event video sequence, its corresponding low dimensional representation is learnt using each of the model components, $\mathcal{M}_i, i = 1 \ldots M$. The likelihood of the event being recognized as belonging to class j using each of the mixture of actions model components is computed. The decision criteria based on the weighted sum of posterior probabilities (for class j) is given by

$$p_j = \sum_{i=1}^{N} \gamma_{ij} p(j|data, \mathcal{M}_i).$$

The event is labelled as belonging to the class j^*, which maximizes the posterior probability according to $j^* = \arg\max_j p_j$.

4 Recognizing Events

The proposed framework is used to recognise events such as hand gestures and human events. We used hand gesture sequences from Marcel's database [15]. Sample frames of some of the events can be seen in Figures 2 and 4. For the experiment on human events, we used videos of 20 human subjects performing 7 different events for an average duration of 6 seconds. Three samples per subject per event were used. Video sequences of 10 human subjects, *i.e.* $10 \times 7 \times 3$ sequences, and another disjoint set of sequences were used for training and testing respectively. These events occur with the subject either being stationary or indulging in locomotion. In the former category, we consider events Flapping,

Table 1. A comparison of recognition accuracy using a single MFA model (which has a fixed composition of appearance and temporal features) and the combination of MFA models. On an average, 35.35 percentage reduction in error was observed. Sample frames of some of these events can be seen in Figures 2 and 4.

Events	% Accuracy	
	Single MFA	Comb. of MFAs
Hand gestures:		
Click	89	94
No	88	93
StopGraspOk	90	92
Rotate	86	90
Human Activities:		
Flapping	83	88
Jumping	80	86
Squatting	83	90
Waving	82	86
Limping	85	92
Walking	87	93
Hopping	84	90
CMU MoBo database:		
Slow walk	84	92
Fast walk	85	94
Incline walk	86	93
Walk with Ball	85	93

Jumping, Squatting and Waving, while in the latter category (involving locomotion), we consider Limping, Walking and Hopping. All the videos were captured with a Panasonic Digital Video Camera at 24 fps. The model is also validated on the MoBo Database [14] available from the Robotics Institute, Carnegie Mellon University. The database consists of 25 subjects performing 4 different walking activities on a treadmill. Each sequence is 11 seconds long recorded at 30 fps. Data corresponding to one of the view angles (vr03_7 of [14]) is used for experimentation. The training and testing data sequences were disjoint in all the three validations.

Minimal preprocessing is done on the video sequences. In order to retain the visually significant information, background subtraction and normalization is performed on all the frames. The intensity values obtained are used in the process henceforth. For the events involving locomotion, the frames are motion compensated to centre the subject performing the event. Using a set of example videos as the training set, the appropriate composition of appearance and temporal features is learnt, and the parameters that describe them for all the events (refer Section 3). Same training sequences are used in all the component models. To recognize an unlabelled test event, the frame sequence transitions are computed via the inference step of EM algorithm. This results in a set of sequence probabilities computed for each event. The test video is then labelled

as the event whose corresponding weighted probability measure is maximum (refer Section 3.2). The recognition accuracy results obtained using the proposed model and an MFA model are presented in Table 1. When compared to the single MFA model, we achieved 35.35 percentage reduction in error on average.

4.1 Discussion

We performed a quantitative analysis of the subspace by reconstructing the original sequences from the learnt representations. Using Λ_j and the low dimensional representation, z_t, the original frames, x_t, $\forall t$, are recovered, thereby generating the entire sequence. The reconstruction error is found to be 0.5%. A comparison of some of the original and recovered frames is shown in Figure 5.

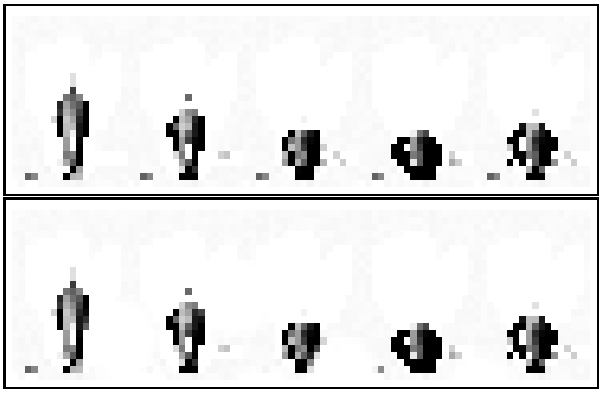

Fig. 5. A comparison of the original (top) and the reconstructed (bottom) frames of the activity Squatting. Even though we achieve 99.94% reduction in size, the reconstruction error is negligible (0.5%).

The recognition process over frames is displayed in Figure 6, as a plot of the log likelihood for each possible activity. The correct activity Squatting – the topmost plot in the figure – is clearly disambiguated within the first few frames (around 5), which shows the ability of the model to obtain all the aspects of the activity quickly and accurately.

The proposed approach differs from various time-series models in many aspects. Our techniques for preprocessing, feature extraction and representation have considerable advantages, as described below.

– In comparison with a standard left-to-right HMM based on [9], the mixture model provides superior recognition. For example, HMM results in 88% accuracy for the hand gesture Click, while the mixture model provides 94% accuracy. Similar improvement (of $6 - 8\%$) is observed in the case of other events.

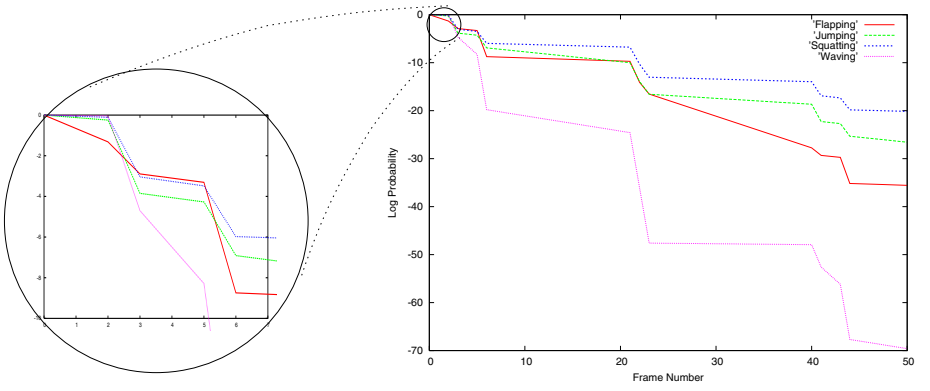

Fig. 6. Cumulative sequence probabilities for the activity Squatting. The horizontal axis represents the frame number and the vertical axis represents the logarithm of sequence probability. The topmost plot (blue dotted line) corresponds to Squatting. A closer view of the graph (shown in inset) indicates that the activity is recognized after observing a few frames – 5 in this case. *Best viewed in pdf*

- The proposed method is related to a standard left-to-right HMM. However, we work at a lower dimension, which is simultaneously obtained while modelling the event structure. Furthermore, a single observation model is used to train all the events in the ensemble unlike HMMs where each event is modelled separately [9].
- Events have been modelled, in the past, using a variety of features [1,7,9]. Most of these methods involve large amount of preprocessing. In contrast, we perform minimal preprocessing and avoid any explicit feature extraction. It is limited to background subtraction and binarization of the individual frames.

5 Conclusion

The mixture model presented in this paper adapts based on the set of events being considered. It learns an optimal combination of various mixture of actions model components. It can also be interpreted as a unifying framework for combining appearance and temporal features in events. The composition of the feature content is controlled by the number of mixtures in the model. The applicability of this framework has been demonstrated using the Mixture of Factor Analyzers model. However, it can easily incorporated in other mixture modelling schemes such as Gaussian Mixture Models. Other video (or event) analysis problems which require a higher level of semantic understanding are yet to explored. Incorporating a discriminant based scheme into this framework is another interesting direction.

Acknowledgments. Karteek Alahari thanks the GE Foundation for financial support through the GE Foundation Scholar-Leaders Program 2003-2005.

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. on PAMI **23** (2001) 257–267
2. Gavrila, D.M.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding **73** (1999) 82–98
3. Greenspan, H., Goldberger, J., Mayer, A.: A probabilistic framework for spatio-temporal video representation and indexing. In: ECCV. Volume IV. (2002) 461–475
4. Veeraraghavan, A., Chellappa, R., Roy-Chowdhury, A.: The Function Space of an Activity. In: CVPR. Volume 1. (2006) 959–968
5. Wong, S.F., Cipolla, R.: Real-time Interpretation of Hand Motions using a Sparse Bayesian Classifier on Motion Gradient Orientation Images. In: BMVC. Volume 1. (2005) 379–388
6. Sheikh, Y., Sheikh, M., Shah, M.: Exploring the Space of a Human Action. In: ICCV. Volume 1. (2005) 144–149
7. Yacoob, Y., Black, M.J.: Parameterized Modeling and Recognition of Activities. CVIU **73** (1999) 232–247
8. Yilmaz, A., Shah, M.: Actions Sketch: A Novel Action Representation. In: CVPR. Volume 1. (2005) 984–989
9. Sun, X., Chen, C.C., Manjunath, B.S.: Probabilistic Motion Parameter Models for Human Activity Recognition. In: ICPR. Volume 1. (2002) 443–446
10. Brand, M., Kettnaker, V.: Discovery and Segmentation of Activities in Video. IEEE Trans. on PAMI **22** (2000) 844–851
11. Ghahramani, Z., Hinton, G.E.: The EM Algorithm for Mixtures of Factor Analyzers. Technical Report CRG-TR-96-1, University of Toronto, Canada (1996)
12. Robertson, N., Reid, I.: Behaviour understanding in video: a combined method. In: ICCV. Volume 1. (2005) 808–815
13. Zelnik-Manor, L., Irani, M.: Event-Based Analysis of Video. In: CVPR. Volume II. (2001) 123–130
14. Gross, R., Shi, J.: The CMU Motion of Body MoBo Database. Technical Report CMU-RI-TR-01-18, Robotics Institute, CMU, Pittsburgh, PA (2001)
15. Marcel, S.: (Dynamic Hand Posture Database: http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/dhp_marcel.tar.gz)
16. Veeraraghavan, A., Roy-Chowdhury, A., Chellappa, R.: Role of shape and kinematics in human movement analysis. In: CVPR. Volume 1. (2004) 730–737