Recognizing Human Activities from Constituent Actions

S. S. Ravi Kiran, Karteek Alahari, C. V. Jawahar
Centre for Visual Information Technology
International Institute of Information Technology
Gachibowli, Hyderabad 500019. INDIA.
Email: jawahar@iiit.net

Abstract

Many of the human activities such as Jumping, Squatting have a correlated spatiotemporal structure. They are composed of homogeneous units. These units, which we refer to as actions, are often common to more than one activity. Therefore, it is essential to have a representation which can capture these activities effectively. To develop this, we model the frames of activities as a mixture model of actions and employ a probabilistic approach to learn their low-dimensional representation. We present recognition results on seven activities performed by various individuals. The results demonstrate the versatility and the ability of the model to capture the ensemble of human activities.

1. Introduction

The problem of characterizing a set of dynamic tasks, referred to as *activities*, by analyzing their video sequences has received considerable research attention over the past few years [1, 2, 5, 6]. It forms an important problem due to its immediate applicability to surveillance, sign language recognition, Human Computer Interaction etc. [2, 6, 9]. Review of the contemporary methods for modelling human activities can be found in [1, 5] and the references therein. Early methods employ segmentation and tracking of individual moving parts to interpret the dynamic activity in the scene [6, 12]. Modelling of activity using motion-based features is an alternate approach [2]. Of late, there has been a spurt of interest in using probabilistic methods such as time-series models, Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) etc. for analyzing activity videos [1, 8, 10, 11].

In this paper, we present a model to learn a compact representation of human activities. The learnt representation is then used in the recognition framework. The motivation for our model arises from the presence of common atomic spatiotemporal units – actions – among the ensemble of human activities. For example, the activity Jumping has two distinct actions – a standing action and an in-air action. Similarly, the activity Flapping (flapping of hands) has standing and hands stretched out as constituent actions (as shown in Figure 1). Clearly, these activities share the common action 'standing'. The correlation that exists between these and other such activities can be profitably exploited in learning a compact representation of the activities.

A typical frame of the activity, $x^{(t)}$ (at t th time instant) can be generated as follows. The action to which it belongs to, is chosen according to the discrete distribution $P(\omega_j), j = 1 \ldots m$. Depending on the chosen action, a continuous subspace representation $z^{(t)}$ is generated according to the distribution $p(z^{(t)}|\omega_j)$. Having obtained $z^{(t)}$ and action ω_j, we obtain the observed $x^{(t)}$ according to the distribution $p(x^{(t)}|z^{(t)}, \omega_j)$. In other words, $x^{(t)}$ is modelled as a "mixture model of actions", with $\omega_j, j = 1 \ldots m$ denoting the j th action.

Human activities are constrained by the degree of freedom allowed for joints and muscles of the human body and hence, limited to a finite set of actions. The problem of characterizing human activities can, therefore, be modelled as that of identifying the constituent actions and their sequencing. Given a large number of video segments, we employ a probabilistic method to learn these individual actions and their compositional rules for the corresponding activities. These actions, in turn, are represented in a lower dimensional space exploiting the spatial redundancy of the action. We learn the actions from examples using a Mixture of Factor Analyzers model coupled with a transition matrix.

The proposed model differs from some of reported methods in various aspects. Our preprocessing is limited to a simple background subtraction followed by thresholding to retain interesting parts of the individual frames. No explicit feature extraction is performed. The activities are captured by their representative factors and actions. In contrast, some of the methods in the recent past are based on extracting features such as measurement of relative distances [1], motion param-

eter vectors [11], colour and motion densities [2, 10] etc. The extracted features represent the activity directly in many approaches. Fitting probabilistic models such as GMMs and HMMs by assuming the form of data distribution is another popular way of describing features [8, 10, 11]. Our approach is directly related to these methods (in particular, a standard left-to-right HMM). We work in low-dimensional subspace using a single observation model, whereas separate HMMs are trained for modelling each activity typically [11]. Once we model the activities, our task is to use the learnt representation to recognize activities. K-nearest neighbour classifier and its variants are popularly used in many of the methods based on explicit feature extraction to achieve this [2, 9]. The likelihood of the sequence of actions inferred from the observations is maximized in our case.

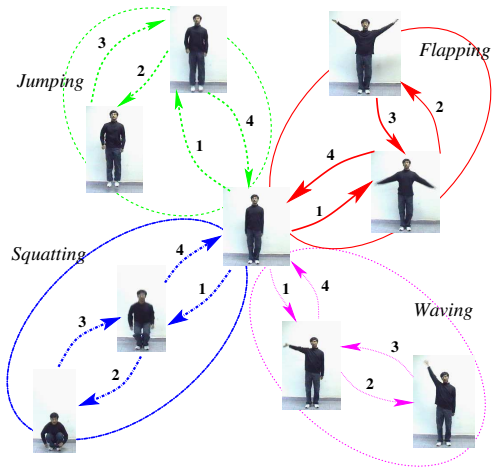


Figure 1: Sample frames showing the representative actions of four activities. The arrows denote the temporal transitions between the actions and the numbers on each arrow denote the temporal sequencing of the activity. In addition, there are self-loops for each action (not shown in the figure). Note that the action 'standing' is common to all of these activities.

The remainder of the paper is organized as follows. We provide the background for Factor Analyzer and Mixture of Factor Analyzers model, following the terminology of [7], in the next section. Section 3 describes the learning algorithm in detail. Experimental results are discussed in Section 4. We summarize the main contributions of the paper and scope for future work in Section 5.

2 Factor Analyzer

A Factor Analyzer (FA) [4] is a statistical model that captures the correlations in data x (of dimension d) to learn a low-dimensional subspace representation z (of dimension p where $p \ll d$). The generative model for this is given by

$$x = \Lambda z + u \tag{1}$$

where Λ is known as the factor loading matrix. The factors z are assumed to be normally distributed ($\mathcal{N}(\mu, 1); \mu = 0$). The d-dimensional random variable u (the associated noise) is distributed as $\mathcal{N}(0, \Psi)$, with Ψ being a diagonal matrix [7]. Given Λ and Ψ, the expected value of the factors can be computed through linear projections as

$$\begin{aligned} E[z|x] &= \beta x, \\ E[zz^T|x] &= I - \beta\Lambda + \beta x x^T \beta^T \end{aligned}$$

where $\beta = \Lambda^T \Sigma^{-1}$ and Σ is the data covariance matrix.

FA has significant advantages over Principal Component Analysis (PCA), another linear dimensionality reduction model. Unlike FA, PCA is not robust to noise in the data [7]. Factor Analyzers provide a statistically appropriate model for data representation. In essence, for each data sample arranged as a vector x, we obtain a low-dimensional representation z by maximizing the expectation $E[z|x]$.

2.1 Mixture of Factor Analyzers

An extension of FA is the Mixture of Factor Analyzers (MFA) model which performs dimensionality reduction along with clustering. Here, we consider a mixture of m factor analyzers (denoted by $\omega_j, j = 1, \ldots, m$) where each factor analyzer has the same number of p factors, but different means (μ_j) and factor loading matrices (Λ_j). The generative model is given by

$$P(x) = \sum_{j=1}^{m} \int P(x|z, \omega_j) P(z|\omega_j) P(\omega_j) dz \tag{2}$$

The parameters in this mixture model are given by $\{(\mu_j, \Lambda_j)_{j=1}^m, \pi, \Psi\}$, where π is the vector of adaptable mixing proportions, $\pi_j = P(\omega_j)$.

The mixture of factor analyzers is essentially a reduced dimensional mixture of Gaussians. Each factor analyzer fits a Gaussian to a portion of the data, weighted by the posterior probabilities h_{ij}. We use MFA to arrive at an efficient representation for human activities. In the following section, we explain the procedure for learning the various actions and the associated activities.

3. Learning Activities

Given multiple instances of the activities, $A_1, \ldots A_K$, our objective is to automatically extract the actions ($\omega_1, \ldots \omega_m$), which constitute these activities and their sequencing

information in order to generate the video segment. Let N be the total number of frames from examples of all the activities. Subsequences of $x^{(t)}$ form actions. These subsequence frames (of an action) are highly correlated and therefore, can be represented in a low-dimensional manifold. That is, $x^{(t)}$ is modelled according to Equation 1 where Λ represents the transformation basis for a particular action. The basis for the j th action is denoted by Λ_j. Multiple such subsequences, occurring across different activities, are used to learn the Λ_js for the actions and the low-dimensional representations ($z^{(t)}$). Now, an activity is modelled as transitions across actions following a specific probabilistic structure. These transitions are learned by observing the $z^{(t)}$s across the various actions for each activity. In the end, we obtain a compact representation of the K activities by automatically learning the m actions and the sequencing information embedded in the example frames of the activities.

We achieve an efficient representation for activities using MFA model, which is essentially a reduced dimensionality mixture model where the m mixture components are actions along with the subspace representations of the frames that contribute to the learned model of each action. Our task, then, is to invert the generative process and learn the parameters of the distributions mentioned above from *all* the frames of *all* the activities. We use the Expectation Maximization (EM) algorithm to perform this. EM 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 [3]. In our case, the data corresponds to the frames, the unknown values to the lower-dimensional representations of these frames and the actions to which these frames are associated. EM 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. The procedure is outlined in further detail below.

3.1. EM Framework

The videos of all the activities of the subjects are represented as a sequence of frames and are used for training. The EM algorithm has two phases - Inference and Learning which are executed sequentially and repeatedly till convergence.

Inference - In this phase, the current estimates of parameters are used to compute the *expected* values for various interactions of the subspace representation and the actions, i.e. we compute $E[\omega_j|x^{(t)}]$, $E[z^{(t)}|\omega_j, x^{(t)}]$ and $E[z^{(t)}z^{(t)^T}|\omega_j, x^{(t)}]$ for all frames t and actions ω_j, all of which can be obtained from Equation 2. These quantities, the computation of which is similar to that in [7], are given by

$$\begin{aligned} E[\omega_j z^{(t)}|x^{(t)}] &= h_{tj}\beta_j(x^{(t)} - \mu_j) \\ E[\omega_j z^{(t)}z^{(t)^T}|x^{(t)}] &= h_{tj}(I - \beta_j\Lambda_j + \Lambda_j(x^{(t)} \\ &\quad - \mu_j)(x^{(t)} - \mu_j)^T\beta_j^T) \end{aligned} \tag{3}$$

where

$$\begin{aligned} h_{tj} &= E[\omega_j|x^{(t)}] = \pi_j\mathcal{N}(x^{(t)} - \mu_j, \Lambda_j\Lambda_j^T + \Psi) \\ \beta_j &= \Lambda_j^T(\Lambda_j\Lambda_j^T)^{-1} \end{aligned} \tag{4}$$

Here, each of $\mu_j, j = 1 \ldots m$ denotes the representative appearance for each of the actions while $\Lambda_j, j = 1 \ldots m$ denotes the various subspace bases for the actions. π denotes the mixing proportions of actions in the activity set while Ψ is a measure of noise present in the data. h_{tj} can be interpreted as the membership of frame t to action j – the higher the value of h_{tj}, the more likely that frame t contains a subject performing action j. In this manner, we *infer* the values of the subspace representations of the frames and the actions to which they belong to, in this phase.

Learning - In this phase, the statistics collected during the inference from *all* the training examples are used to obtain better estimates of parameters. We solve a set of linear equations to find π_j, Λ_j, μ_j and Ψ. The interested reader may refer to [7] for more details. Each of the frames $x^{(t)}$ is assigned to an action c_t according to :

$$c_t = \arg\max_j h_{tj} \qquad j = 1 \ldots m \tag{5}$$

Thus, each frame is assigned to the action for which it has the maximum membership.

After the EM algorithm converges, we form the action transition matrix $T_k = [\tau_{pq}^k]$ for each activity A_k as follows.

$$\tau_{pq}^k = \sum_{t=1}^{N-1} [c_t = p][c_{t+1} = q] \tag{6}$$

where $1 \leq p, q \leq m$.

The action transitions for successive frames of the activity A_k are represented by the entries in the transition matrix T_k. This matrix encodes the temporal characteristics of the activity. The corresponding *probability* transition matrix P_k can be easily constructed by normalizing the entries.

Thus, by the end of training phase, we obtain the parameters of the model – $\{(\mu_j, \Lambda_j)_{j=1}^m, \pi, \Psi\}$, $\{P_k\}_{k=1}^K$. The model which now encapsulates the activity structure can be employed for the various tasks such as recognition, which is briefly described below.

3.2. Applying the model for recognition

Using the parameters obtained in the training phase, we recognize activities in an unlabelled video. Let the activity being recognized have N_s frames. We reduce the dimensionality of the problem by using the factors *learned* from the training data. We also compute the membership of the frames in each of the actions (from Equation 4). Each frame is then assigned a single action label using Equation 5. Let $c_1,\ c_2 \ldots c_{N_s}$ be the action assignments for the respective frames. Then, the sequence probability S_k is computed using $S_k = \prod_{t=1}^{N_s-1} P_k[c_t][c_{t+1}]$. The unlabelled video is assigned to be the activity A_k^*, which maximizes S_k. If the test video has more than one activity, we can obtain each of the activities present by observing the ranges of selected features extracted from the subject performing the activity.

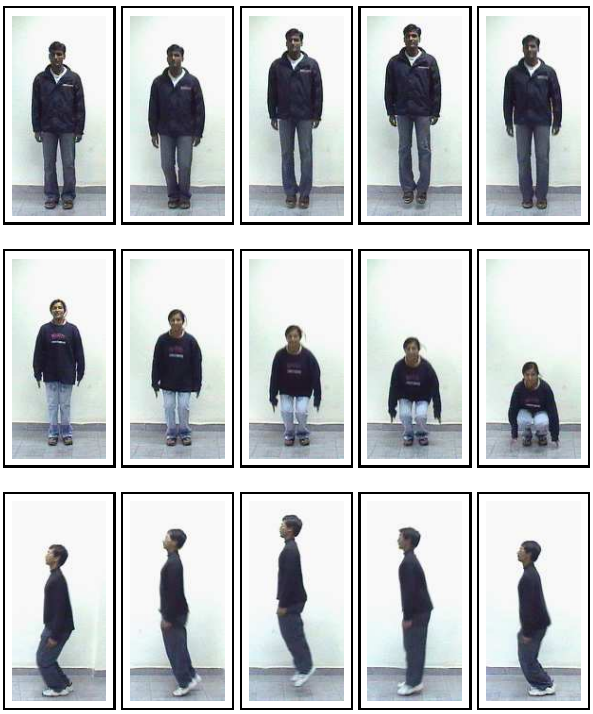
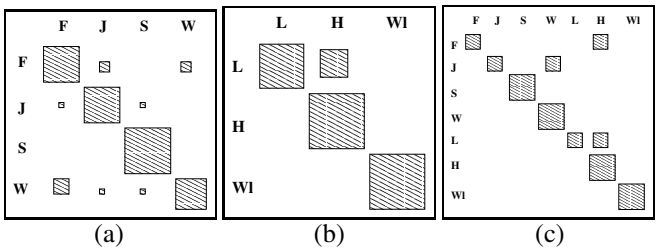


Figure 2: Sample frames of in-place activities - Jumping (top row), Squatting and an activity involving motion - Hopping (bottom row).

4. Results

Recognition of activities involving the whole body finds a plethora of applications in surveillance-based domains. These activities usually occur with the subject stationary or indulging in locomotion. In the former category, we consider activities Flapping, Jumping, Squatting (Figure 2) and Waving, while in the latter category (involving locomotion),



(a) (b) (c)

Figure 3: Confusion Matrices for *in-place* (F - Flapping, J - Jumping, S - Squatting, W - Waving), locomotion (L - Limping, H - Hopping, Wl - Walking) and the entire activity set respectively. The areas of the squares are proportional to the numerical entries of the confusion matrix.

we consider Limping, Walking and Hopping (bottom row of Figure 2). The videos have been captured using a Panasonic Digital Video Camera at 22.4 fps. We use the videos of 7 human subjects performing above mentioned 7 activities, of average duration 8 seconds. A considerable degree of correlation across the ensemble of activities is observed from the videos. For example, the Jumping and Squatting (rows 2 and 3 in Figure 2) have similar kind of frames for most of their activity durations (in this case, standing still for a short duration).

In order to retain only the visually significant information, background subtraction and normalization is performed on all the frames. Motion compensation is performed to center the subject for activities where locomotion is involved. To recognize an unlabelled test activity, the frame sequence transitions are computed via the inference step of EM algorithm and the sequence probability is computed for each activity. The test video is labelled as the activity for which this probability is maximum (Refer to section 3.2).

The ability of the model to accommodate considerable variation in the range and variety of spatial motion is highlighted by the results (Figures 3(a), 3(b) and Figure 3(c) (the entire ensemble of the 7 activities)). The occasional misclassification is present between activities which share spatial coherence to a large degree, for example Jumping and Waving. The accuracy over the various body activities is $87 - 90\%$.

5. Summary

We have presented a framework for learning to represent various kinds of human activities which can be used for recognizing them efficiently. A low-dimensional representation is learned which captures the spatial and temporal aspects of activities ideal for applications involving quick activity recognition. Here, we summarize some of the significant contributions of our work.

- The model frees us from the task of feature extraction. Instead, the features are automatically chosen so as to *best* explain the observed activity in an economical manner. The preprocessing on raw video data is quite minimal. In addition, the model does not incur the computational overhead of subject tracking since such precise spatio-temporal localization is not a primary requirement. The probabilistic framework allows for a coarse localization while leveraging the power of Bayesian inference for learning the actions and subspace representation.

- The framework is independent of scale at which the activity is captured, therefore, it can be applied to the 320×240 frames of whole body activities as well as 64×28 frames of the same activities.

- Since actions can be learned individually from each activity, the training sequences need not be aligned to actions or possess equal length.

- The learned representations are intuitive – they are based on the actions that occur when an activity is performed. This is clearly demonstrated by the representative appearances of actions (shown in Figure 1). Also, the transition matrix in Figure 4 indicates the actions which constitute the activity Squatting. The rows and columns correspond to the actions learned by the model. The areas of the squares indicate the transition probabilities between these actions. Notice that the predominant entries correspond to Standing and Sitting - the main actions present in Squatting.

- The low-dimensional representation makes the model extremely favourable for applications involving real-time recognition. That is, if we consider the one of the activities (shown in Figure 2), we have a representation needing only 40 floating point numbers to explain a 320×240 frame, a reduction of nearly 99.94%.

In conclusion, we look at a couple of possible extensions to the proposed model. We are currently working on developing the low-dimensional representation for multiple views *directly* from the given limited set of views views. We intend to work along these lines to improve the robustness of the model and investigate multi-view motion models in a broader framework. The model is quite suited to popular video applications such as continuous video summarization and representation.

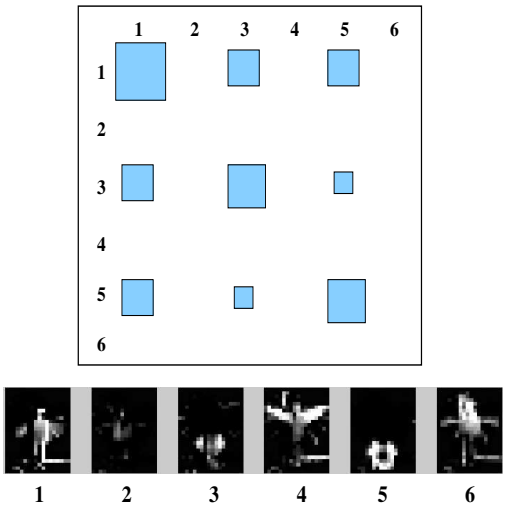


Figure 4: Cluster transition matrix for the activity Squatting, showing the individual cluster means. The rows and columns correspond to the actions learned by the model. The shaded areas are proportional to the numerical probability entries in the transition matrix. Here, squatting is represented by the transitions among clusters $1, 3, 5$. Note the constituent actions - standing and sitting - represented by these cluster means.

References

[1] Proc. of IEEE Workshop on Event Mining: Detection and Recognition of Events in Video. Wisconsin, June 18-20, 2003.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on PAMI*, 23(3):257–267, 2001.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, (39):1–38, 1977.

[4] B. S. Everitt. *An Introduction to latent variable models*. Chapman and Hall, London, 1984.

[5] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[6] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. *Proc. of IEEE Conf. on CVPR*, pages 73–80, 1996.

[7] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. *University of Toronto, Technical Report*, CRG-TR-96-1, 1996.

[8] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Proc. of IEEE Conference on CVPR*, 1:313–320, 2003.

[9] O. Masoud and N. Papanikolopoulos. Recognizing human activities. *IEEE Conf. on AVSS*, pages 157–162, 2003.

[10] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. *Proc. of International Conference on Multimodal Interfaces*, pages 3–8, 2002.

[11] X. Sun, C. C. Chen, and B. S. Manjunath. Probabilistic motion parameter models for human activity recognition. *Proc. of ICPR*, 1:10443–10446, 2002.

[12] Y. Yacoob and M. J. Black. Parameterized modelling and recognition of activities. *CVIU*, 73:232–247, 1999.