Learning Human Motion Models from Unsegmented Videos

Roman Filipovych Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences, Florida Institute of Technology,
Melbourne, FL 32901, USA
{rfilipov,eribeiro}@fit.edu

Abstract

We present a novel method for learning human motion models from unsegmented videos. We propose a unified framework that encodes spatio-temporal relationships between descriptive motion parts and the appearance of individual poses. Sparse sets of spatial and spatio-temporal features are used. The method automatically learns static pose models and spatio-temporal motion parts. Neither motion cycles nor human figures need to be segmented for learning. We test the model on a publicly available action dataset and demonstrate that our new method performs well on a number of classification tasks. We also show that classification rates are improved by increasing the number of pose models in the framework.

1. Introduction

Human motion recognition is of relevance to both the scientific and industrial communities. Despite significant recent developments, general human motion recognition is still an open problem. Approaches usually analyze dynamic information in image sequences. Successful approaches have focused on probabilistic inference [13, 9]. Spatio-temporal features have been shown to be effective for motion recognition [16, 14]. Additionally, the importance of static information [15] combined with advances in probabilistic constellation models [13] have also been demonstrated. However, most methods still require segmentation of either human figure or motion cycles from videos. This limitation implies the need of substantial data preparation.

In this paper, we propose a human motion model consisting of constellation models "tuned" to recognize specific human poses combined with a constellation model of the motion's spatio-temporal data. This combination of static and spatio-temporal (dynamic) information is the key idea of our method. Our model allows us to develop efficient learning and recognition procedures. The main difference between ours and other approaches [13] is that, in our framework, poses and motion dynamics are modeled explicitly. Also, unlike [9], learning and classification do not require the input of manually extracted motion cycles. We demonstrate the effectiveness of our method on a series of motion classification experiments. A comparison with a recent motion recognition approach is also provided. Our learning algorithm allows for the unsupervised discovery of representative pose models. This contrasts with related approach by Niebles and Fei Fei [13] where spatio-temporal and static features from different frames are combined within a bags-of-words framework. Our unsupervised learning algorithm has significant advantages over existing supervised part-based motion recognition approaches. The enforcement of spatio-temporal constraints between model parts enables us to reduce information loss inherent to some bags-of-words methods. Our results show an improvement in recognition rate (75.3% and 88.9%) over Niebles and Fei Fei's approach (72.8%), and the avoidance of manual dataset preparation (e.g., manual input of motion cycles).

Human motion recognition approaches can be grouped into data-driven and model-based methods. Data-driven approaches operate directly on the data. Dollar *et al*. [7] perform action classification using support vector machines. Leo *et al*. [12] use projection histograms of binary silhouette's for modeling human pose deformation. Unfortunately, high ambiguity of features in videos can be a problem to these methods. On the other hand, model-based approaches explicitly include higher-level knowledge about the data by means of a previously learned model. However, the performance of these approaches strongly depends on both the choice of the model and the availability of prior information. For example, Boiman and Irani [3] propose a graphical Bayesian model for motion anomaly detection. The method describes the motion data using hidden variables that correspond to hidden ensembles in a database of spatio-temporal patches. Niebles and Fei Fei [13] represent

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

actions using a probabilistic constellation model. The performance of model-based approaches strongly depends on the choice of the model and learning procedure. Additionally, in the absence of prior information about the models' structure, the learning task may become intractable.

2. Human Motion Model

Let us begin by considering $\boldsymbol{\mathcal{V}}$ as an N-frame human motion video sequence. $\boldsymbol{\mathcal{V}}$ records a human pose's temporal variation. Let $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ be a set of poses, $K \ll N$, sampled from all possible representative poses of a human motion type. Let $\boldsymbol{\mathcal{M}}$ represent the video's spatio-temporal information. $\boldsymbol{\mathcal{M}}$ describes the video frames' temporal variations obtained from measurements such as optical flow [8] or spatio-temporal features [13]. Also, let $\boldsymbol{\mathcal{X}}$ represent simultaneously a particular spatio-temporal configuration of pose and human motion dynamics. The likelihood of observing a particular video sequence given a human motion's spatio-temporal configuration is $p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}})$. We assume statistical independence of appearances of both pose and dynamics. The likelihood function can then be factorized as follows. From Bayes' rule:

$$\begin{aligned} p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{V}}) &\propto p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}})\; p(\boldsymbol{\mathcal{X}}) \\ &\propto \underbrace{p(\boldsymbol{\mathcal{P}}|\boldsymbol{\mathcal{X}})}_{\substack{\text{poses'} \\ \text{appearance}}} \; \underbrace{p(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{X}})}_{\substack{\text{dynamics'} \\ \text{appearance}}} \; \underbrace{p(\boldsymbol{\mathcal{X}})}_{\substack{\text{spatio-temporal} \\ \text{configuration}}} \end{aligned} \tag{1}$$

Our method's underlying idea is that spatio-temporal arrangement of pose parts and motion dynamics are encoded into the prior probability while the likelihood distributions encode their appearances. Pose and dynamic information are represented by directed acyclic star graphs (*i.e.*, graph vertices are conditioned to a landmark vertex). The factorization is inspired by the part-based object model in [6].

2.1. Spatio-Temporal Prior Model

Part-Based Pose Models. Let us assume that poses $\mathcal{P}_i = \{(\mathbf{a}_1^{(i)}, \mathbf{x}_1^{(i)}), \ldots, (\mathbf{a}_{N_{\mathcal{P}_i}}^{(i)}, \mathbf{x}_{N_{\mathcal{P}_i}}^{(i)})\}$ are subdivided into $N_{\mathcal{P}_i}$ non-overlapping subregions. Each pair $(\mathbf{a}_j^{(i)}, \mathbf{x}_j^{(i)})$ contains the local appearance $\mathbf{a}$ and the spatio-temporal location $\mathbf{x}$ of subregion j of pose model $\mathcal{P}_i$, respectively. The pose's temporal position in the sequence serves as the temporal coordinate of the parts' locations. Pose subregions are assumed to be arranged in a star-graph configuration in which the pose's landmark vertex is $(\mathbf{a}_r^{(i)}, \mathbf{x}_r^{(i)})$. The graphs in Figure 1(a) and Figure 1(b) illustrate the pose models.

Part-Based Motion Dynamics Model. Dynamics information required by our model is given by a sparse set of spatio-temporal features [7, 11]. Let $\mathcal{M} = \{(\mathbf{a}_1^{(\mathcal{M})}, \mathbf{x}_1^{(\mathcal{M})}), \ldots, (\mathbf{a}_{N_\mathcal{M}}^{(\mathcal{M})}, \mathbf{x}_{N_\mathcal{M}}^{(\mathcal{M})})\}$ be a set of $N_\mathcal{M}$ representative spatio-temporal interest features. The set $\mathcal{M}$ is also arranged in a star-graph configuration with $(\mathbf{a}_r^{(\mathcal{M})}, \mathbf{x}_r^{(\mathcal{M})})$ as the graph's landmark vertex. Figure 1(c) shows a dynamics model graph.

Integrated Model of Poses and Motion Dynamics. Finally, a global multi-layered tree-structured model is built by conditioning the landmark vertices of pose model graphs on the landmark vertex of the dynamics model graph as shown in Figure 1(d) (graph arrows indicate conditional dependences between connected vertices). In this layer, the spatio-temporal locations of the partial models are the locations of the corresponding landmark image subregions. The joint distribution for the partial models' spatial configuration can be derived from the graphical model in Figure 1(d):

$$p(\boldsymbol{\mathcal{X}}) = p(\mathbf{x}^{(\mathcal{M})}) \prod_{\mathcal{P}_i \in \boldsymbol{\mathcal{P}}} p(\mathbf{x}^{(i)}|\mathbf{x}^{(\mathcal{M})}) \tag{2}$$

where $\mathbf{x}^{(i)}$ is the spatio-temporal configuration of the pose $\mathcal{P}_i$, and $\mathbf{x}^{(\mathcal{M})}$ is the dynamics model' spatio-temporal configuration. The probability distributions in (2) are:

$$p(\mathbf{x}^{(\mathcal{M})}) = p(\mathbf{x}_r^{(\mathcal{M})}) \prod_{j \neq r} p(\mathbf{x}_j^{(\mathcal{M})}|\mathbf{x}_r^{(\mathcal{M})}) \tag{3}$$

$$p(\mathbf{x}^{(i)}|\mathbf{x}^{(\mathcal{M})}) = p(\mathbf{x}_r^{(i)}|\mathbf{x}_r^{(\mathcal{M})}) \prod_{j \neq r} p(\mathbf{x}_j^{(i)}|\mathbf{x}_r^{(i)}) \tag{4}$$

The above partial models' dependence is based only on their spatio-temporal configuration. This follows from our assumption that the partial models are statistically independent with respect to their appearance.

2.2. Appearance Models

Under the independence assumption, the appearance likelihood of pose $\mathcal{P}_i$ can be written as:

$$p(\mathcal{P}_i|\boldsymbol{\mathcal{X}}) = \prod_j^{N_{\mathcal{P}_i}} p(\mathbf{a}_j^{(i)}|\mathbf{x}_j^{(i)}) \tag{5}$$

Similarly, the appearance likelihood of the motion dynamics model is given by:

$$p(\mathcal{M}|\boldsymbol{\mathcal{X}}) = \prod_j^{N_\mathcal{M}} p(\mathbf{a}_j^{(\mathcal{M})}|\mathbf{x}_j^{(\mathcal{M})}) \tag{6}$$

As a result, the likelihood term in Equation 1 becomes:

$$\begin{aligned} p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}}) &= p(\boldsymbol{\mathcal{P}}|\boldsymbol{\mathcal{X}})\; p(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{X}}) \\ &= \prod_i^K \prod_j^{N_{\mathcal{P}_i}} p(\mathbf{a}_j^{(i)}|\mathbf{x}_j^{(i)}) \prod_j^{N_\mathcal{M}} p(\mathbf{a}_j^{(\mathcal{M})}|\mathbf{x}_j^{(\mathcal{M})}) \end{aligned} \tag{7}$$

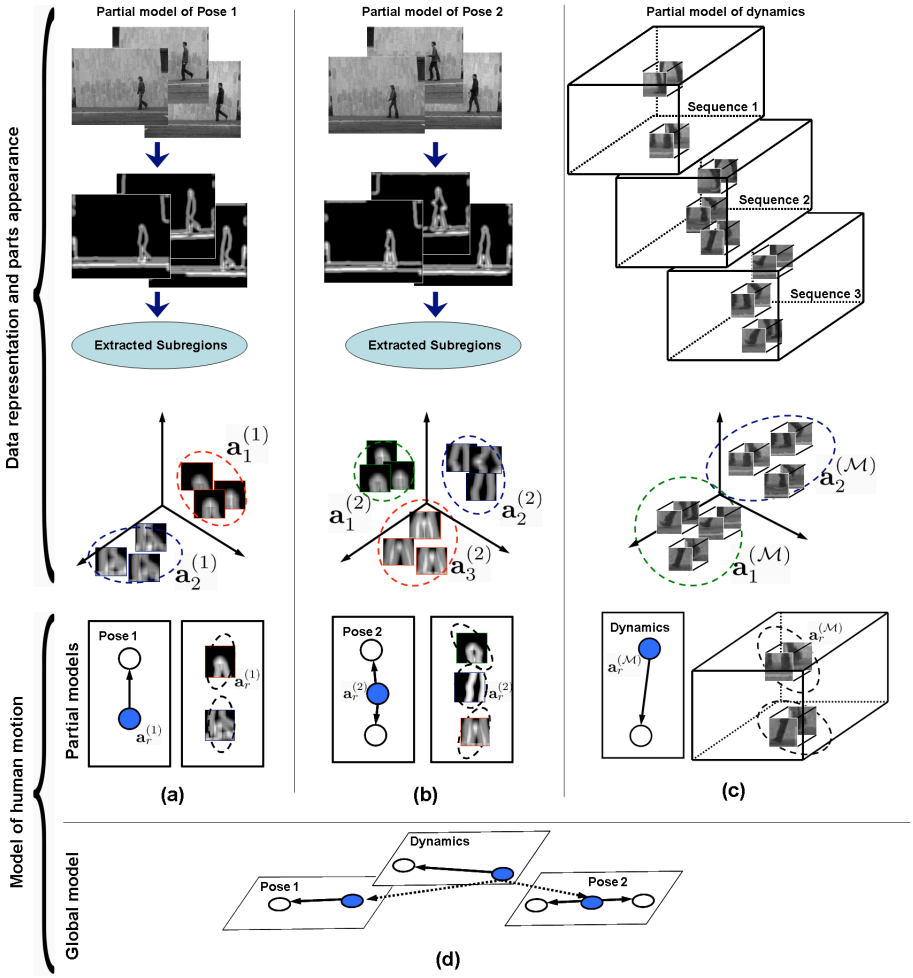


Figure 1. Learning spatio-temporal configurations of human pose and motion dynamics. (a,b) **pose models**: Multiple frames of the same pose are extracted. Interest regions are extracted from the edge maps of frames. Representative pose parts are learned. (c) **dynamics model**: Spatio-temporal features are extracted [7]. Representative spatio-temporal parts are learned. (d) Final integrated model.

3. Model Learning

Model parameters are estimated from a set of unsegmented sequences $\{\boldsymbol{\mathcal{V}}_1, \ldots, \boldsymbol{\mathcal{V}}_L\}$. Our learning method is divided into the following steps (Figure 2).

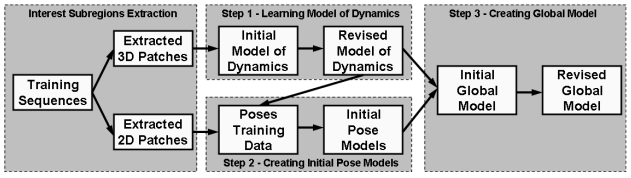


Figure 2. Human motion learning process.

Learning Step 1 - Learning the Dynamics Model. In this step, the initial parameters of the partial model for motion dynamics are estimated. The subregion locations' probabilities in (3) are modeled using Gaussian joint probability distributions. Fortunately, conditional distributions relating independent Gaussian distributions are also Gaussian [1]. As a result, the conditional densities in (3) and (4) take a particularly simple form (*i.e.*, the joint probability distribution can be formed by simply combining the means and covariance matrices of the corresponding independent densities). To proceed, we extract a set of subregions centered at spatio-temporal interest points [7]. We adapted the learning process described by Crandall and Huttenlocher [6] to work with spatio-temporal configurations of parts. The initial spatio-temporal model is created, and the optimal number of parts is determined. An E.M.-based procedure is used to simultaneously estimate the parameters of the distributions $p(\mathbf{x}_j^{(\mathcal{M})}|\mathbf{x}_r^{(\mathcal{M})})$ in (3) and the dynamics appearance in (6). The outcome of this step is a preliminary motion dy-

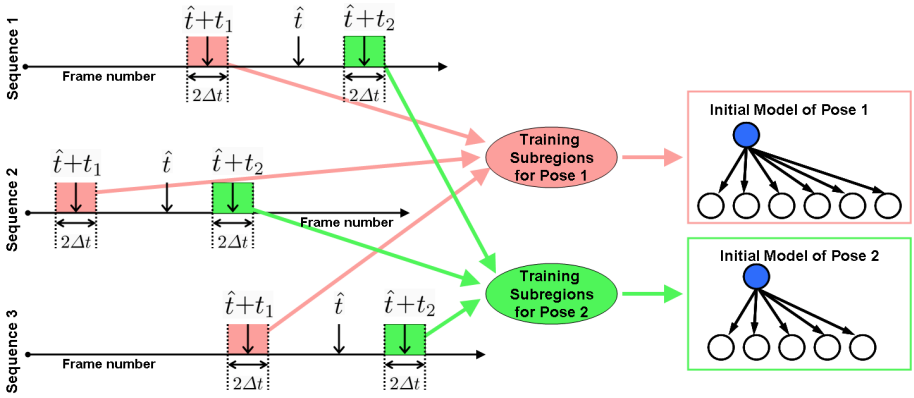


Figure 3. Creating the initial pose models.

namics model. In the next step, this preliminary model will be used to guide the selection of representative poses to be integrated into the final global representation.

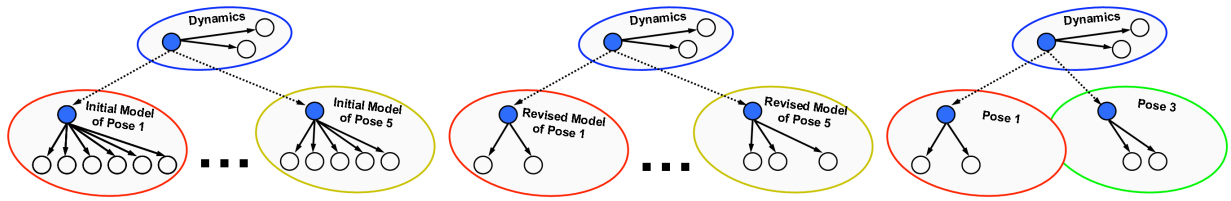
Learning Step 2 - Creating Initial Pose Models. The goal of this step is to estimate the initial candidate pose models. The input to this step consists of a set of interest subregions extracted from all frames of the videos. Each subregion has a spatial image location and a temporal location associated with it. However, direct clustering used in the previous step is not applicable here due to a number of reasons. First, some subregions have similar appearance across a wide range of frames (*e.g.*, the appearance of the head is usually constant during the walking cycle). Hence, clustering would make indistinguishable some similar parts that belong to different poses. Secondly, all parts of a pose model must have zero temporal variance as they naturally belong to the same frame. We address this initialization problem by using the motion dynamics model to constrain the space of input subregions during learning. More specifically, for every training sequence, we obtain the MAP dynamics model locations using the maximization:

$$\hat{\mathbf{x}}^{(\mathcal{M})} = \arg\max_{\mathbf{x}} p(\mathcal{M}|\mathbf{x}^{(\mathcal{M})}) p(\mathbf{x}^{(\mathcal{M})}) \tag{8}$$

For every sequence, the maximization in (8) results in the location of the model's landmark part (*i.e.*, $\hat{\mathbf{x}}^{(\mathcal{M})} = (\hat{x}, \hat{y}, \hat{t})$). Initial samples $\mathbf{t_0} = \{t_1, t_2, ..., t_K\}$ of temporal displacements are generated, where t_i indicate the temporal displacement of a pose $\mathcal{P}_i$ from the landmark node of the dynamics model. For a pose $\mathcal{P}_i$, we select from the training sequences those subregions for which temporal displacements from the localizations of the dynamics model landmark node belong to the interval $(t_i - \Delta t, t_i + \Delta t)$, where Δt is a constant value that defines the span of frames containing the pose. These subregions are subsequently used to obtain candidate parts and initial model parameters for pose $\mathcal{P}_i$. Pose parts selection is performed as in Step 1. Pose subregions are clustered to form initial parts of the underlying pose. Learned pose parts are organized into a star-graph structure with the most descriptive part representing the landmark node. The initial spatio-temporal parameters are estimated from the maximum likelihood locations of parts as follows (Figure 3):

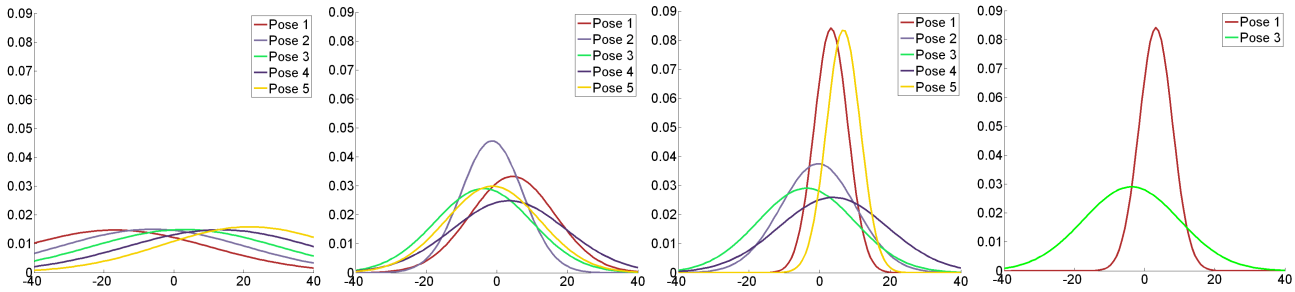
$$\hat{\mathbf{x}}^{(i)} = \arg\max_{\mathbf{x}} p(\mathcal{P}_i|\mathbf{x}^{(i)}) \tag{9}$$

Learning Step 3 - Creating the Global Model. We commence by obtaining an initial estimate of the parameters of the conditional distributions $p(\mathbf{x}_r^{(i)}|\mathbf{x}_r^{(\mathcal{M})})$ in (4) from the MAP locations of the dynamics model and initial maximum likelihood location of pose models. However, not only the initial pose models contain noisy parts, but the parameters of the conditional distributions $p(\mathbf{x}_r^{(i)}|\mathbf{x}_r^{(\mathcal{M})})$ in (4) are very inaccurate. We use E.M. algorithm to revise the global model parameters. Figure 5(a) shows cross-sections of the conditional distributions along the time axis. An example of the corresponding cross-sections of the revised conditional distributions is shown on Figure 5(b). Our MAP estimation algorithm considers only model configurations in which pose parts belong to the same frame. This algorithm is presented in Section 4. While the revised conditional distributions become better defined, the updated model still contains a number of overlapping parts. We remove overlapping parts and re-estimate model parameters with the E.M. algorithm once again (Figure 4(b)). Figure 5(c) shows examples of the cross-section of the resulting conditional distributions along the time axis. Finally, only poses that are temporally well-defined are retained. This is accomplished by pruning pose model subgraphs whose landmark nodes have the largest temporal variances when conditioned on the dynamics model. Moreover, when two intervals $(t_i - \Delta t, t_i + \Delta t)$ and $(t_j - \Delta t, t_j + \Delta t)$ overlap, several instances of a same pose may be incorrectly



(a) (b) (c)

Figure 4. Creating the global model. (a) Initial pose models are combined with the model of dynamics; (b) Nodes corresponding to overlapping parts are pruned; (c) Finally, only temporally well defined poses are retained (Pose 1 and Pose 3).



(a) (b) (c) (d)

Figure 5. Temporal cross-sections of partial models conditional distributions (five-pose models). (a) Initial pose models are created; (b) Initial global model parameters are revised; (c) Overlapping parts removed and model parameters revised. (d) Two final remaining pose models. Note: three pose subgraphs were pruned to enforce the minimum temporal distance between poses to be four-frames.

learned. To address this problem, we greedily select a subset of pose models such that, when conditioned on the dynamics model's landmark node, the mean temporal locations of their landmark nodes are separated by a pre-defined distance (*e.g.*, one frame)(Figure 4(c)).

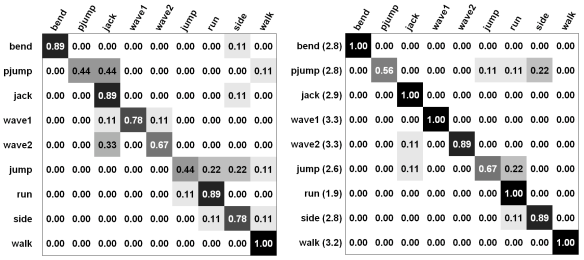
4. Classification

For recognizing a human motion in a video sequence, we seek for the video's spatio-temporal location that maximizes the posterior probability in (1). In the case of the tree-structured Bayesian network, the proposed representation is equivalent to the Random Markov Fields (RMF) model in which the potential functions are conditional probability densities. As mentioned earlier in this paper, we expect same-pose parts to belong to the same frame in the video sequence. This observation allows us to significantly reduce the global model's MAP search space. The steps of our exact inference algorithm are: (a) Consider all $\mathbf{x}_r^{(\mathcal{M})}$; (b) Consider all $\mathbf{x}_j^{(\mathcal{M})}$; (c) Consider all $\mathbf{x}_r^{(i)}$; (d) For every $\mathbf{x}_r^{(i)}$, consider only those $\mathbf{x}_j^{(i)}$ that have the same temporal coordinate as $\mathbf{x}_r^{(i)}$. (e) Calculate the configuration's posterior probability.

5. Experimental Results

We tested our model on the human action dataset from [2]. This database contains nine action classes performed by nine different subjects. Since our method is view-dependent, the direction of the motion in all input videos is the same in all sequences (*e.g.*, a subject is always walking from right to left). We extracted square patches centered at the image locations detected by a Harris operator. Gaussian smoothed edge-maps are generated from the extracted square subregions. The edge maps served as input to build static-pose models. Features required to create the dynamics model were obtained using the spatio-temporal interest point detector described in [7]. Principal component analysis (PCA) was used to reduce the dimensionality of all features. However, the similar background appearance in the sequences from [2] induced bias in the learning process. To address this issue, we synthesized background data for the dynamics model training step from portions of sequences containing no subjects. Corresponding frames served as background data for the pose learning module.

We compared our results with the results reported by Niebles and Fei Fei [13]. Similarly, we adopted a leave-one-out scheme for evaluation. Labeling decision is based only on the best match for each model. Results suggest that the dynamics model alone is not sufficient to perform accurate classification. In our experiments, human motion mod-



(a) "dynamics one-pose" (b) "dynamics several-pose"

Figure 6. Classification confusion matrices (a) 75.3% correct classification; (b) 88.9% correct classification.

els using no pose models achieved classification rate of only 35.8%. In order to obtain the initial pose models, pose models' relative temporal distances to the dynamics model were manually set to $\mathbf{t_0} = \{-20, -10, 0, 10, 20\}$. In our experiments, we set $\Delta t = 5$ frames.

In the first experiment, we retained only the temporally best defined pose. This model's classification confusion matrix is shown in Figure 6(a), and presents a 75.3% overall recognition rate. This rate is superior to the 72.8% classification rate reported in [13]. In the second experiment, we increased the number of pose models. In our implementation, the number of pose models is indirectly controlled by an inter-pose temporal distance threshold. We set this threshold value to four frames. This model's classification confusion matrix is presented in Figure 6(b). The figure also displays (in parentheses along side the motion types) the average number of poses retained in the global model for a specific motion. The overall recognition rate was 88.9%. The method mostly misclassified those actions where the pose does not change significantly during the motion (*e.g.*, "jump" and "pjump" actions). Figure 7 provides qualitative results for action models from the latter experiment. More specifically, it displays several motion models superimposed on the test sequences at the detected locations. Plots at the top of corresponding actions represent cross-sections of "dynamics-pose" conditional distributions along the time axis. They provide the answer to the question of at what temporal displacement from dynamics a specific pose is expected to be located. White borders indicate static-pose parts. Grayed rectangles represent slices of the landmark's spatio-temporal subregion in corresponding frames. The results allow us to make the following conclusions: (1) The inclusion of additional pose models helps remove overall classification ambiguity. (2) Final global model's conditional distributions may significantly differ from the initial ones. See, for example, the plot in Figure 7, fourth column, where all pose models have positive temporal displacements.

6. Conclusions

We presented a novel method for learning human motion models from unsegmented sequences. More specifically, we demonstrated how partial models of individual static poses can be combined with partial models of the video's motion dynamics to achieve motion classification. We demonstrated the effectiveness of our method on a series of motion classification experiments using a well-known motion database. We also provided a comparison with a recently published motion recognition approach. Our results demonstrate that our method offers promising classification performance. We understand that limitations of the benchmark dataset does not allow us to show the approaches' full potential in presence of complex actions. For this reason, we intend to further investigate the performance of our method on private datasets. Future directions of investigation include a study of the possibility of using alternative appearance models, and the development of a new framework for classifying human-object interactions.

Acknowledgments. Research supported by U.S. Office of Naval Research's contract: N00014-05-1-0764.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., NJ, USA, 2006.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *CVPR*, pages I: 462–469, 2005.

[4] M. C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV-II*, pages 628–641, 1998.

[5] G. Carneiro and D. Lowe. Sparse flexible models of local features. In *ECCV-III*, pages 29–43, 2006.

[6] D. J. Crandall and D. P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV-I*, volume 3951 of *LNCS*, pages 16–29. Springer, 2006.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[8] R. Fablet and M. J. Black. Automatic detection and tracking of human motion with a view-based representation. In *ECCV-I*, pages 476–491, 2002.

[9] R. Filipovych and E. Ribeiro. Combining models of pose and dynamics for human motion recognition. In *ISVC*, Lake Tahoe, USA, November 2007.

[10] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions - Computers*, 22:67–92, 1977.

[11] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, Nice, France, October 2003.

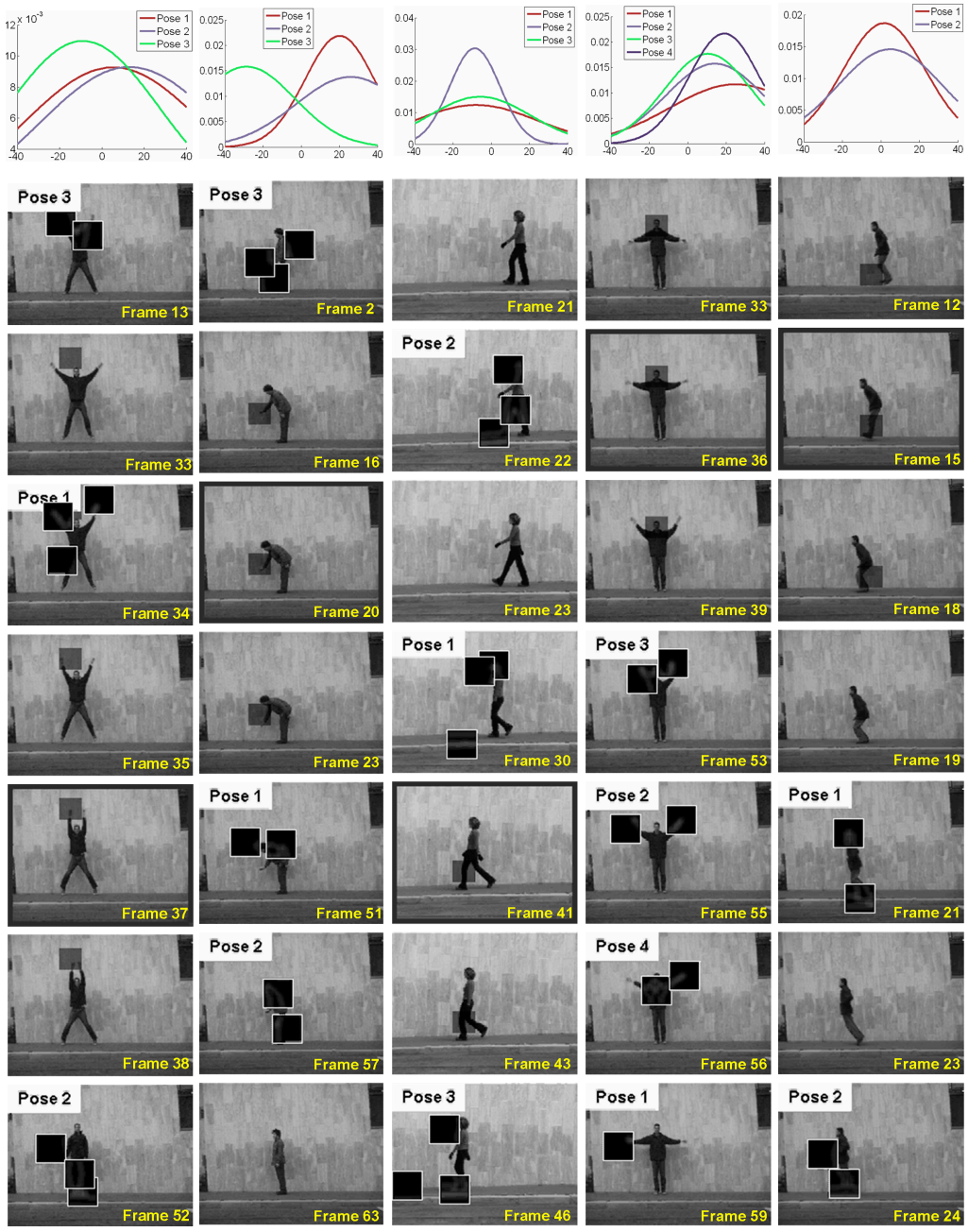


Figure 7. Learned motion models superimposed on the test sequences at the detected locations. White borders indicate static-pose parts. Grayed rectangles represent 2D central slice of the landmark spatio-temporal subregion in corresponding frames. Dark border highlights the temporal coordinate of the spatio-temporal feature corresponding to the landmark node of dynamics.

[12] M. Leo, T. D'Orazio, I. Gnoni, P. Spagnolo, and A. Distante. Complex human activity recognition for monitoring wide outdoor environments. In *ICPR*, Vol.4, pages 913–916, 2004.

[13] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, Minneapolis, USA, June 2007.

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, Vol. 3, pages 32–36, 2004.

[15] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, pages 1654–1661, 2006.

[16] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR*, Minneapolis, USA, June 2007.