Combining Models of Pose and Dynamics for Human Motion Recognition

Roman Filipovych and 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
http://www.cs.fit.edu/~eribeiro

Abstract. We present a novel method for human motion recognition. A video sequence is represented with a sparse set of spatial and spatial-temporal features by extracting static and dynamic interest points. Our model learns a set of poses along with the dynamics of the sequence. Pose models and the model of motion dynamics are represented as a constellation of static and dynamic parts, respectively. On top of the layer of individual models we build a higher level model that can be described as "constellation of constellation models". This model encodes the spatial-temporal relationships between the dynamics of the motion and the appearance of individual poses. We test the model on a publicly available action dataset and demonstrate that our new method performs well on the classification tasks. We also perform additional experiments to show how the classification performance can be improved by increasing the number of pose models in our framework.

1 Introduction

Recognizing human actions from videos is of relevance to both the scientific and industrial communities. Humans usually perform actions by means of a number of articulated complex motions. Consequently, creating effective computational models for human motion representation is a crucial but challenging task required to all action recognition algorithms. Despite significant efforts by the computer vision community, action recognition is still an open problem. In general, approaches to human motion recognition work by analyzing the dynamic information of image sequences. Recently, the use of spatial-temporal features has been demonstrated to be an affective tool for motion recognition [18,16]. Additionally, the importance of static information [17] combined with recent advances in probabilistic constellation models [15] have also been demonstrated.

In this paper, we focus ourselves on the problem of learning representational models for human motion recognition. More specifically, we propose a Bayesian probabilistic framework that allows for integrating both static and dynamic information. Here, our main contribution is to present a principled solution to the

G. Bebis et al. (Eds.): ISVC 2007, Part II, LNCS 4842, pp. 21–32, 2007.

human motion recognition problem by combining data of different nature within a single probabilistic integration framework while allowing for computationally efficient learning and inference algorithms. This is accomplished by combining constellation models "tuned" to recognize specific human poses with a constellation model of the motion's spatial-temporal data to form a single human motion model. Our resulting method can be characterized as a "constellation of constellation models" that combines pose recognition and motion dynamics recognition into a single framework.

We demonstrate the effectiveness of the proposed method on a series of motion classification experiments along with a comparison with a recently published motion recognition approach. The results show that our model offers promising classification performance on an established human action dataset.

The remainder of this paper is organized as follows. In Section 2, we comment on the literature related to the problem addressed in this paper. Section 3 describes the details of our action recognition framework. In Section 4, we describe experimental results of our method on an established human action database. Finally, Section 5 presents our conclusions and directions for future investigation.

2 Related Work

Approaches to human motion recognition can be grouped into data-driven and model-based methods. Data-driven approaches operate directly on the data. For example, Dollar *et al.* [7] perform action classification in the space of extracted spatial-temporal features using a support vector machine classifier. Leo *et al.* [13] describe an unsupervised clustering algorithm for motion classification based on histograms of binary silhouette's horizontal and vertical projections. These methods are computationally efficient and achieve good classification performance. However, data-driven methods may be inadequate in most realistic scenarios, primarily because local image features are typically highly ambiguous. On the other hand, model-based approaches explicitly include higher-level knowledge about the data by means of a previously learned model. Despite their computational and mathematical elegance, the performance of model-based approaches strongly depends on both the choice of the model and the availability of prior information about the data at hand. Additionally, in the absence of prior information about the models' structure, the learning task is often intractable. Graphical models represent a suitable solution to this problem as they allow for efficient learning and inference techniques while simultaneously providing a span of models with rich descriptive power. 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 ensemble in a database of spatial-temporal patches. Niebles *et al.* [14] create a generative graphical model where action category labels are present as latent variables.

Recently, there has been considerable development in part-based classification methods that model the spatial arrangement of object parts [9,8,5]. These methods are inspired by the original ideas proposed by Fischler and Elschlager [11]. For example, Fergus *et al.* [9] proposed a fully-connected part-based probabilistic model for object categorization. The approach is based on the constellation model proposed in [4]. Unfortunately, the complexity of the model learning and inference often increases drastically as the number of parts increases. A solution to this problem is to select model structures that allow for both optimal classification performance and tractable learning and inference.

In this paper, we propose a method that combines a set of partial motion models into a global model of human motion. Each partial model is a constellation model [4] of a relatively simple yet descriptive structure that describes either the motion dynamics or a specific pose of the motion cycle. The partial models are combined within a Bayesian framework to form a final human motion model that encodes spatial-temporal relationships between the partial models. Here, models are learned from a set of labeled training examples. The key difference between ours and other approaches such as the one described in [15] is that in our framework poses and motion dynamics are modeled explicitly.

3 Our Method

In this section, we present our human motion recognition framework. The goal of our approach is twofold. First, we aim at combining the static information provided by the pose images with the video's spatial-temporal information to obtain an integrated human motion model. Secondly, we will use this integration model for the classification of human motion sequences. We accomplish these goals by approaching the human motion recognition problem as a probabilistic inference task. An overview of our approach is illustrated in Figure 1. Next, we introduce our probabilistic integration framework followed by a description of the learning and classification procedures.

3.1 Integrating Human Pose and Motion Dynamics

We commence by defining the main components of our model. A video sequence $\boldsymbol{\mathcal{V}}$ of human motion can be considered to be the variation of a specific human pose as a function of time. Let $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ represent a discrete set of K poses sampled from the space of all possible poses that are representative of a specific human motion type, where K is a small number, usually much smaller than the number of frames in the video sequence. Let $\boldsymbol{\mathcal{M}}$ represent the spatial-temporal information extracted from the video sequence. This information describes temporal variations in the image frames, and can be obtained from measurements such as optical flow and spatial-temporal features. Additionally, let $\boldsymbol{\mathcal{X}}$ represent simultaneously a particular spatial-temporal configuration of pose and human motion dynamics.

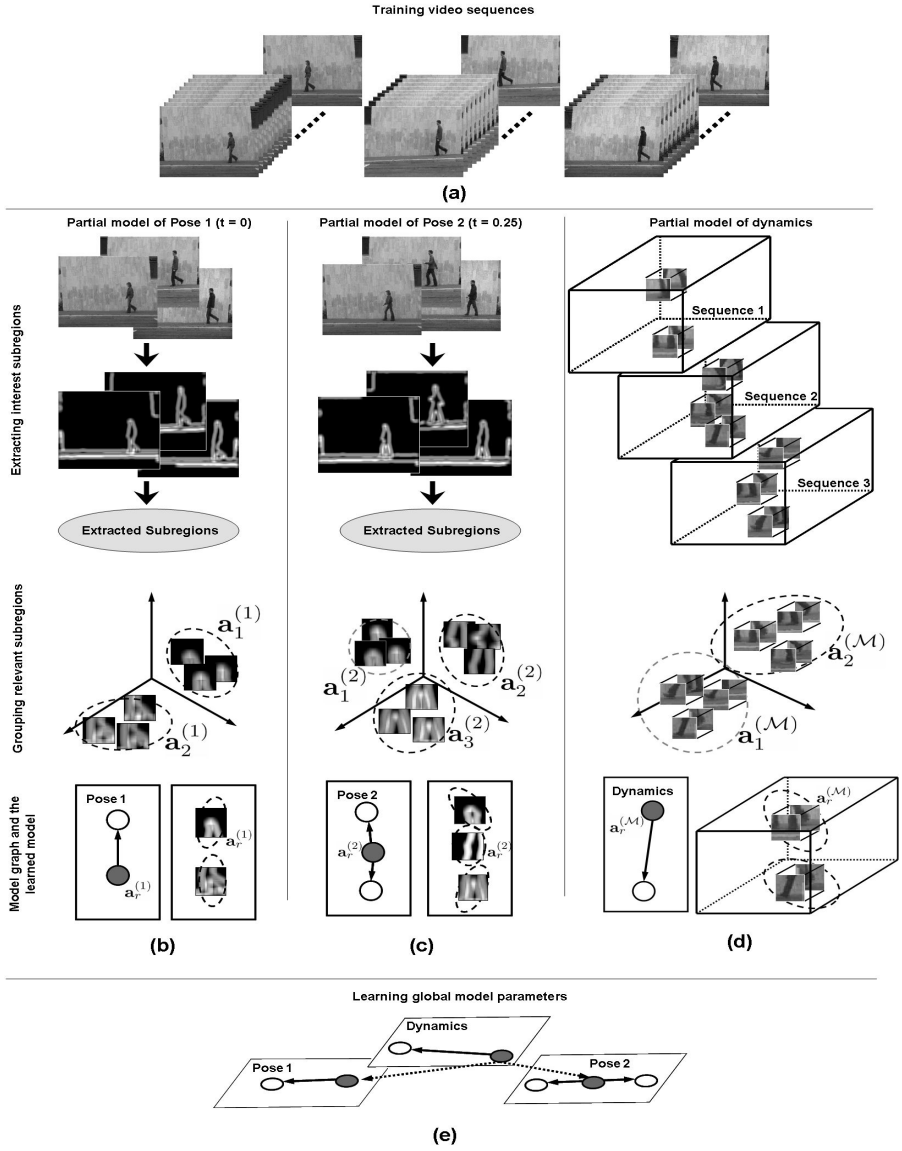


Fig. 1. Diagram of our approach. An illustrative example of "dynamics two-pose" model. Temporally-aligned extracted motion cycles form the training set (a). For each partial pose model a set of frames corresponding to the same time instance are selected (b, c). The images are preprocessed and the interest subregions are extracted. In the case of dynamics partial model, the spatial-temporal features are extracted using the detector from [7] (d). The partial models' parameters are estimated independently. These models are then combined to form a global model (e). *(Note: The extracted and learned subregions displayed in the chart do not present the actual subregions, as in our implementation the dimensionality of the input subregions is reduced using PCA).*

Probabilistically, the likelihood of observing a particular video sequence given that a human motion is at some spatial-temporal location can be represented by the distribution $p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}})$. From the Bayes' theorem, we obtain:

$$p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{V}}) \quad \propto \quad p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}})\, p(\boldsymbol{\mathcal{X}}) \quad \propto \quad \underbrace{p(\boldsymbol{\mathcal{P}}|\boldsymbol{\mathcal{X}})}_{\substack{\text{poses} \\ \text{appearance}}} \; \underbrace{p(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{X}})}_{\substack{\text{dynamics} \\ \text{appearance}}} \; \underbrace{p(\boldsymbol{\mathcal{X}})}_{\substack{\text{spatial-temporal} \\ \text{configuration}}} \tag{1}$$

We further assume that the appearance of both pose and dynamics are statistically independent. This assumption allows us to factorize the likelihood function in Equation 1 into two components. Accordingly, we introduce the variables $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{M}}$ to indicate that the human motion information in the video is represented by a set of static poses and dynamic information, respectively. Our integration model of pose and motion dynamics is inspired by the part-based object factorization suggested by Felzenszwalb and Huttenlocher [8]. The underlying idea in our factorization is that the spatial-temporal arrangement of parts can be encoded into the prior probability distribution while the likelihood distribution encodes the appearance. In this paper, we focus ourselves on the combination of human motion dynamics with both the appearance and the spatial-temporal configuration of the pose models.

Spatial-Temporal Prior Model. The prior distribution in Equation 1 is described as follows. We begin by assuming that each pose $\mathcal{P}_i$ from $\boldsymbol{\mathcal{P}}$ can be subdivided into a number of non-overlapping subregions such that $\mathcal{P}_i = \{(\mathbf{a}_1^{(i)}, \mathbf{x}_1^{(i)}), \ldots, (\mathbf{a}_{N_{\mathcal{P}_i}}^{(i)}, \mathbf{x}_{N_{\mathcal{P}_i}}^{(i)})\}$, where the components of each pair $(\mathbf{a}_j^{(i)}, \mathbf{x}_j^{(i)})$ represent the local appearance $\mathbf{a}$ and the spatial-temporal location $\mathbf{x}$ of the subregion j for the model of pose $\mathcal{P}_i$, respectively. Here, $N_{\mathcal{P}_i}$ is the total number of subregions for the pose $\mathcal{P}_i$. While a pose conveys only two-dimensional spatial information, the temporal position of the pose in the video sequence serves as the temporal coordinate of the parts' locations. Similarly, the dynamic information required by our model can be represented by a sparse set of spatial-temporal features [7,12]. Accordingly, 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 spatial-temporal interest features where $N_\mathcal{M}$ is the number of features in $\mathcal{M}$. The pose models and the dynamics model are the partial models used in our integration framework. The creation of these models is described next.

For simplicity, we model both pose and dynamic information using directed acyclic star graphs. This is similar to the part-based object model suggested by Fergus *et al.* [10]. Here, a particular vertex is assigned to be a landmark vertex $(\mathbf{a}_r^{(i)}, \mathbf{x}_r^{(i)})$ for the pose $\mathcal{P}_i$. A similar landmark vertex assignment is done for the dynamics model, $(\mathbf{a}_r^{(\mathcal{M})}, \mathbf{x}_r^{(\mathcal{M})})$. The remaining vertices within each model are conditioned on the corresponding landmark vertex. Figure 1(b) and Figure 1(c) show examples of the partial model graphs for pose while Figure 1(d) shows a graph of the partial model for the motion dynamics. Finally, we build another structural layer on top of the pose models and the motion dynamics model. In this layer, the spatial locations of the partial models are the locations of the

corresponding landmark image subregions. The global structural layer is built by conditioning the landmark vertices of the pose model graphs on the landmark vertex of the dynamics model graph. In this way, we obtain a multi-layered tree-structured model, which is the global model of human motion used in our method. The graph in Figure 1(e) illustrates our partial models' integration concept. Here, the arrows in the graph indicate the conditional dependence between the connected vertices. Accordingly, the joint distribution for the partial models' spatial interaction can be derived from the graphical model shown in Figure 1(e), and is given by:

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

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

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

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

It should be noted that the dependence between the partial models is based solely on their spatial-temporal configuration within the global model. This follows from our assumption that the partial models are statistically independent with respect to their appearance. Next, we describe the appearance likelihood component of Equation 1 for both pose and motion dynamics.

Appearance Model. Under the appearance independence assumption, the appearance likelihood of the pose $\mathcal{P}_i$ can be written as the product of the probabilities of its subregions (i.e., parts):

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

Similarly, in our motion dynamics model, the appearance likelihood 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:

$$p(\boldsymbol{\mathcal{V}}|\boldsymbol{\mathcal{X}}) = p(\boldsymbol{\mathcal{P}}|\boldsymbol{\mathcal{X}})\ p(\boldsymbol{\mathcal{M}}|\boldsymbol{\mathcal{X}}) = \prod_{i}^{K} \prod_{j}^{N_{\mathcal{P}_i}} p(\mathbf{a}_j^{(i)}|\mathbf{x}_j^{(i)}) \times \prod_{j}^{N_{\mathcal{M}_i}} p(\mathbf{a}_j^{(\mathcal{M})}|\mathbf{x}_j^{(\mathcal{M})}) \tag{7}$$

Next, we describe the parameters estimation step (i.e., learning) of our model as well as the motion classification procedure.

3.2 Learning and Classification of Human Motions

Learning. In the learning stage, the parameters of our model are estimated from a set of training video sequences. The factorization in Equation 2 and Equation 7 allows for the learning process to be performed in a modular fashion given a set of training videos $\{\boldsymbol{\mathcal{V}}_1, \ldots, \boldsymbol{\mathcal{V}}_L\}$. We restrict each of the training videos to contain exactly one full motion cycle (e.g., two steps of the walking motion). Additionally, we temporally align motion cycles extracted from the training sequences such that they start and finish with the same pose. To obtain the pose training data, we first normalize the length of the sequences to be within the $[0, 1]$ time interval. Then, we extract the corresponding frames from the normalized sequences for a specific time instant. Consequently, in the case of periodic motion, the frames corresponding to the time instants 0 and 1 will contain the same pose translated in time (and, for some motions, also in 2D space). Figure 2 shows an example of the aligned walking cycles. In the figure, the frames correspond to the normalized time slices $t = 0$, $t = 0.25$, $t = 0.5$, $t = 0.75$, and $t = 1$. The aligned sequences serve as the input to the dynamic model learning algorithm. The learning procedure is divided into two main steps.

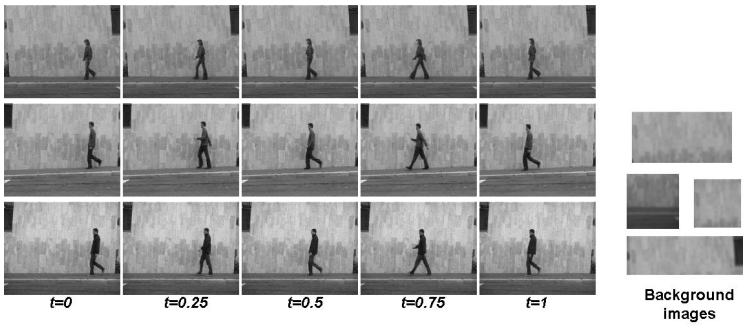


Fig. 2. Examples of segmented and normalized walking sequences, and background images. Frames for normalized time slices: $t = 0, t = 0.25, t = 0.5, t = 0.75$ and $t = 1$.

First, the algorithm estimates the parameters for each of the partial models. Secondly, the parameters representing the spatial-temporal configuration of the global model are determined. The learning steps of our algorithm are detailed next. For simplicity, the probabilities in our model are represented by Gaussian densities.

Learning step 1 - *Learning the parameters of the partial models.* In this step, the parameters of the partial models for pose and motion dynamics are estimated. We begin by modeling the probabilities of subregion locations in Equation 2 using Gaussian joint probability distributions. Fortunately, it can be shown that the conditional distributions relating independent Gaussian distributions are also Gaussian [1]. As a result, the conditional densities in Equations 3 and 4

take a particularly simple form. Further details on Gaussian joint distributions can be found in [1].

For a pose model, we commence by extracting a set of subregions centered at the locations provided by the interest point detector. The method requires two types of input. The first one is a set of positive training images (i.e., images containing the target pose) and a set of negative training images (i.e., background images). We associate a 3D location with every extracted subregion. Here, the locations is represented by the x- and y-coordinates of the subregion in the pose image, and the additional t-coordinate is the frame-position of the pose image in the input sequence. Unlike some other approaches [2], our method does not require the pose to be segmented from the image. We adopt the learning process described by Crandall and Huttenlocher [6]. However, in our work, we consider the spatial-temporal configuration of parts rather than only the spatial configuration. In essence, their method uses a clustering technique to estimate the initial appearances of the possible parts. Then, an initial spatial model is created and the optimal number of parts is determined. An EM-based procedure is used to simultaneously refine the initial estimates of the appearances and the spatial parameters. In our method, we use an E.M. approach to simultaneously estimate the parameters of the distributions $p(\mathbf{x}_j^{(i)}|\mathbf{x}_r^{(i)})$ in Equation 4 and the pose appearance in Equation 5.

We estimate the parameters of the dynamic model in a similar fashion as in the case of the pose models. We proceed by extracting a set of spatial-temporal interest points using the detector described in [7]. We again use the learning method from [6] to estimate the parameters of the distributions $p(\mathbf{x}_j^{(\mathcal{M})}|\mathbf{x}_r^{(\mathcal{M})})$ in Equation 4 and the dynamics appearance in Equation 6.

Learning step 2 - *Estimating the parameters of the global model.* The goal of this step is to estimate the parameters of the distributions that govern the relationships between the partial models. More specifically, we aim at estimating the parameters of the distributions $p(\mathbf{x}_r^{(i)}|\mathbf{x}_r^{(\mathcal{M})})$ in Equation 4. Given the original training data instances for each partial model (i.e., extracted frames for the pose models and aligned sequences for the dynamics model), we compute the most likely location for each data type by maximizing the likelihood of pose models:

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

and the dynamics model:

$$\hat{\mathbf{x}}^{(\mathcal{M})} = \arg\max_{\mathbf{x}} p(\mathbf{x}^{(\mathcal{M})}|\boldsymbol{\mathcal{M}}) \tag{9}$$

Once the maximum likelihood locations evaluated for every partial model and its corresponding data instances are at hand, we can directly estimate the parameters of the distributions $p(\mathbf{x}_r^{(i)}|\mathbf{x}_r^{(\mathcal{M})})$ in Equation 4. These distributions govern the spatial-temporal interaction between the partial models.

Classification. The problem of recognizing a human motion in a video sequence can then be posed as follows. We seek for the spatial-temporal location in the video sequence that maximizes the posterior probability of the location of the motion given a set of partial models as given in (1):

$$\widehat{\boldsymbol{\mathcal{X}}} = \arg\max_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{V}}) \tag{10}$$

It is worth pointing out that, in the case of the tree-structured Bayesian network, the model is equivalent to the Random Markov Fields (RMF) model in which the potential functions are the conditional probability densities. An efficient inference algorithm for such graph structure was studied by Felzenszwalb and Huttenlocher [8]. The algorithm allows to perform exact inference in a reasonable time if the number of partial models is small.

4 Experimental Results

The goal of our experiments is to demonstrate the potential of our method for the classification of human motion. To accomplish this goal, we tested our model on the human action dataset from [2]. This database contains nine action classes performed by nine different subjects. Figure 3 shows a sample of video frames for each motion analyzed in our experiments. Additionally, we compared our results with the results reported by Niebles and Fei Fei in [15]. Finally, we provide some preliminary experimental results on the effect of including additional partial models into the global modeling proposed in this paper.

Video Data Preparation. We begin by pre-processing the video data. Since our method is view-dependent, we reflect frames of some sequences with respect to the y-axis, such that the direction of motion is the same in all sequences (e.g., a subject is always walking from right to left). In our implementation, when learning a pose model, we employ the Harris corner detector to obtain static interest point locations for pose images. We limit the number of interest points to be 20 for every pose image. A Gaussian smoothed edge-map of the pose images is obtained from which we extract square patches centered at the detected locations. The features required to create the dynamics model were obtained by means of the spatial-temporal interest point detector described in [7]. In all cases the dimensionality of the data was reduced using PCA. However, the appearance of the background in the sequences from [2] is very similar. This appearance similarity tends to induce a bias in the learning process. To address this issue, we created background data for the dynamics model from portions of the sequences in which no subject was present. On the other hand, corresponding frames served as the background data for the pose learning module. A sample of the static frames extracted from the background sequences is shown in Figure 2. In the results that follow, the maximum number of learned parts of each partial model is set to four. This is done by selecting only up to four most descriptive parts using the descriptiveness evaluation procedure as described in [6].

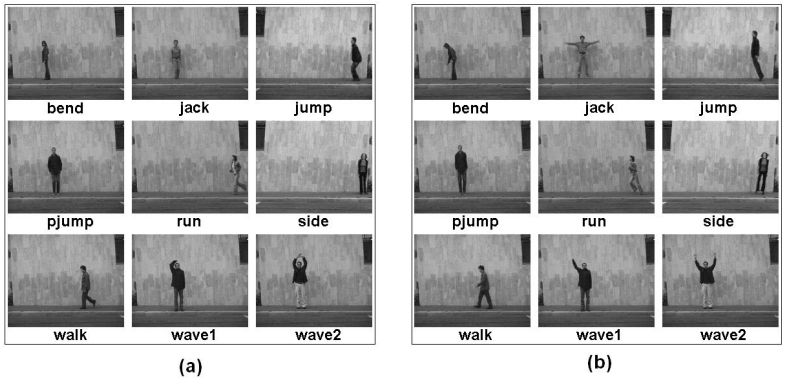
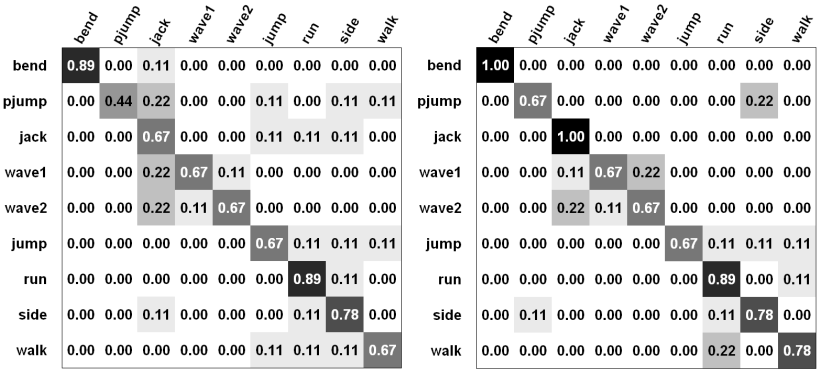


Fig. 3. Human action dataset: Example frames from video sequences in the dataset from [2]. The images correspond to frames extracted at $t = 0$ (a), and $t = 0.25$ (b).

Classification. We compared our results with the results reported by Niebles and Fei Fei [15]. Similarly, we adopted a leave-one-out scheme for evaluation, by taking videos of one subject as testing data, and using sequences of the remaining subjects for training. The segmentation of sequences is not required by our method. Additionally, only the best match for each model is considered when making the labeling decision. For a given motion type we selected one motion cycle from every original training sequences. The set of the segmented motion cycles is the training set of our method. The partial model learning algorithm is EM-based and strongly depends on the correct initialization. To reduce the effect of incorrect initialization, we removed one of the training sequences from the training set and assigned it to be a validation sequence. The learning algorithm was repeated five times and the model for which the posterior probability calculated on the validation sequence was the highest was retained.

In our experiments, we also investigated the effect of including additional partial models into the global model within our framework. First, we built a "dynamics one-pose" model that combines the motion dynamics model and a single pose model. The classification results obtained with this model were compared to results produced by a "dynamics two-pose" model that combines motion dynamics model and two pose models. For the "dynamics one-pose" model the pose was extracted at $t = 0$. For the "dynamics two-pose" model, the poses were extracted at $t = 0$ and $t = 0.25$, respectively. Figure 3 shows a sample of pose images that correspond to these time instants. The confusion tables generated by our classification results are shown in Figure 4. When classifying sequences, the "dynamics one-pose" model allows to correctly classify 70.4% of the testing videos. The method mostly misclassified those sequences for which the pose is similar at the given time instant. This is the case for pose images for the "pjump", "jack", and "side" actions (Figure 3(a)). On the other hand, with the "dynamics two-pose" model our system was able to correctly classify 79.0% of the test sequences. This is superior to the 72.8% classification rate in [15].



(a) (b)

Fig. 4. Confusion matrices for the "dynamics one-pose" model (a) and "dynamics two-pose" model (b). The "dynamics one-pose" model correctly classifies 70.4% of the test sequences. The "dynamics-two poses" model correctly classifies 79.0% of the sequences.

5 Conclusions

In this paper, we presented a novel principled solution to the problem of recognizing human motion in videos. Our method works by combining data of different nature within a single probabilistic integration framework. 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 the proposed 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. Future directions of investigation include a study of the possibility to automatically select poses that would lead to optimal recognition performance. A possible way to follow is to use boosting to improve the selection of optimal poses and dynamics information.

Acknowledgments. This research was supported by U.S. Office of Naval Research under contract: N00014-05-1-0764.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Secaucus, NJ, USA. Springer, Heidelberg (2006)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Int. Conference on Computer Vision, 1395–1402 (2005)
3. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: Conf. on Computer Vision and Pattern Recognition, pp. 462–469 (2005)

4. Burl, M.C., Weber, M., Perona, P.: A probabilistic approach to object recognition using local photometry and global geometry. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 628–641. Springer, Heidelberg (1998)
5. Carneiro, G., Lowe, D.: Sparse flexible models of local features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 29–43. Springer, Heidelberg (2006)
6. Crandall, D.J., Huttenlocher, D.P.: Weakly supervised learning of part-based spatial models for visual object recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 16–29. Springer, Heidelberg (2006)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (October 2005)
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision 61(1), 55–79 (2005)
9. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. Int. J. Comput. Vision 71(3), 273–303 (2007)
10. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: CVPR 2005. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2005)
11. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions - Computers 22, 67–92 (1977)
12. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE Int. Conf. on Computer Vision, Nice, France (October 2003)
13. Leo, M., D'Orazio, T., Gnoni, I., Spagnolo, P., Distante, A.: Complex human activity recognition for monitoring wide outdoor environments. In: ICPR 2004. Proceedings of the Pattern Recognition, 17th International Conference, vol. 4, pp. 913–916. IEEE Computer Society Press, Los Alamitos (2004)
14. Niebles, J., Wang, H., Wang, H., Fei Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC 2006. British Machine Vision Conference, p. 1249 (2006)
15. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA (July 2007)
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR 2004. Proceedings of the Pattern Recognition, 17th International Conference, vol. 3, pp. 32–36. IEEE Computer Society Press, Los Alamitos (2004)
17. Wang, Y., Jiang, H., Drew, M.S., Li, Z.-N., Mori, G.: Unsupervised discovery of action classes. In: CVPR 2006. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1654–1661. IEEE Computer Society Press, Los Alamitos (2006)
18. Wong, S.-F., Kim, T.-K., Cipolla, R.: Learning motion categories using both semantic and structural information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA (June 2007)