Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model

Liang Wang[†], David Suter
Department of Electrical and Computer Systems Engineering
Monash University, Clayton 3800, Victoria, Australia
{lwang, d.suter@}eng.monash.edu.au

Abstract

We describe a probabilistic framework for recognizing human activities in monocular video based on simple silhouette observations in this paper. The methodology combines kernel principal component analysis (KPCA) based feature extraction and factorial conditional random field (FCRF) based motion modeling. Silhouette data is represented more compactly by nonlinear dimensionality reduction that explores the underlying structure of the articulated action space and preserves explicit temporal orders in projection trajectories of motions. FCRF models temporal sequences in multiple interacting ways, thus increasing joint accuracy by information sharing, with the ideal advantages of discriminative models over generative ones (e.g., relaxing independence assumption between observations and the ability to effectively incorporate both overlapping features and long-range dependencies). The experimental results on two recent datasets have shown that the proposed framework can not only accurately recognize human activities with temporal, intra- and inter-person variations, but also is considerably robust to noise and other factors such as partial occlusion and irregularities in motion styles.

1. Introduction

Human motion analysis has recently attracted increasing interest from computer vision researchers [1]. In particular, human activity recognition has a wide range of promising applications, e.g., video surveillance, intelligent interface, and interpretation/retrieval of sport events.

Generally, there are two important questions involved in activity recognition. One is how to extract useful motion information from raw video data, and the other is how to model reference movements, while enabling training and recognition methods to effectively deal with variations at spatial and temporal scales within similar motion classes.

Various cues have been used in the recent literature, e.g., key poses [11,12,13], optical flow [4], local descriptors [5], trajectories or joint angles from tracking [2,6], silhouettes [3,7,12], etc. However, the use of key frames lacks motion information. Image measurements in terms of optical flow or interest points could be unreliable in cases of smooth surfaces, motion singularities and low-quality videos. Feature tracking is not also easy due to the big variability in the appearance and articulation of the human body.

Human activities can be regarded as temporal variations of human silhouettes. Silhouette extraction from video is relatively easier for current imperfect vision techniques, especially in the imaging setting with fixed cameras. So the method that we present here prefers to use (probably imperfect) space-time silhouettes for human activity representation with kernel-induced subspace analysis.

Since human activities evolve dynamically over time, temporal models such as HMMs (Hidden Markov Models) and their variants [2,10] have been widely used to model human motions. However, a strong assumption of independence is usually made in such generative models, which makes them difficult to accommodate multiple overlapping features or long-range dependencies among observations. Conditional random fields (CRFs) [20] proposed by Lafferty avoid the independence assumption between observations, thus having the freedom to incorporate both overlapping features and long-range dependencies into the model. To the best of our knowledge, only two relevant works have tried different forms of conditional approaches for motion [9] or gesture [23] recognition. This paper further explores an alternative conditional model, i.e., factorial CRF [21] that has the joint discriminative learning ability.

The contribution of this paper is to propose an integrated probabilistic framework, as shown in Fig. 1, for the task of activity recognition from simple silhouette observations. The proposed framework consists of two major modules, i.e., feature extraction and description in high-dimensional image space, and activity modelling and recognition in low-dimensional embedded space. We use KPCA [16] to discover the intrinsic structure of the articulated action space, and exploit factorial CRF [21] for activity modeling and recognition (no previous work has investigated FCRF in this context). Experimental results on two datasets have demonstrated both effectiveness and robustness of the proposed framework.

1-4244-1180-7/07/$25.00 ©2007 IEEE

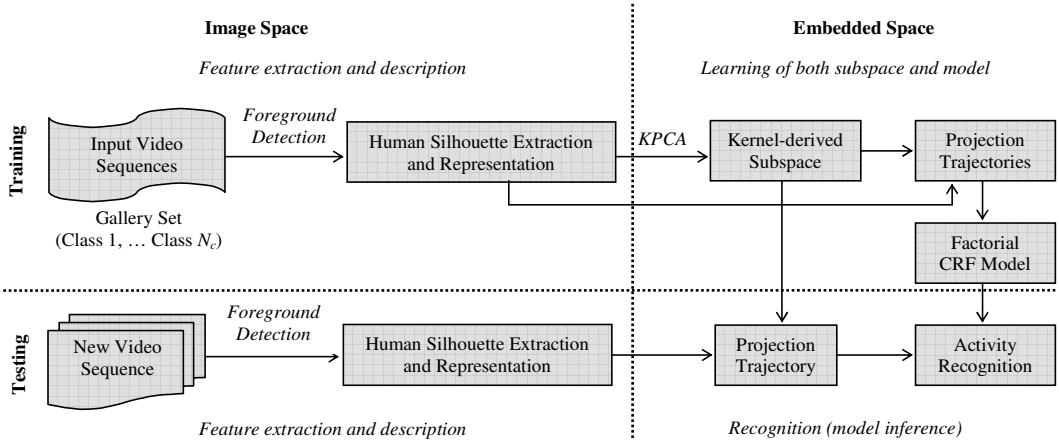

Figure 1. Flowchart of the proposed framework of activity recognition with discriminative conditional graphical model

The remainder of this paper is organized as follows. Section 2 simply reviews related work and Section 3 details feature extraction. Section 4 describes activity modeling and recognition. The results are presented and discussed in Section 5, prior to a summary in Section 6.

2. Related work

Current studies on human activity recognition have tried a variety of features and classification methods. In [6], an activity was represented by a set of pose and velocity vectors of major body parts and recognition of a sequence of pose vectors was achieved with a method of indexing of multidimensional hash tables. Other approaches exploited local descriptors based on interest points in images or videos. Schuldt *et al*. [5] constructed video representations in terms of local space-time features and integrated such representations with SVM for action recognition. Optical flow has also been widely used. Efros *et al*. [4] proposed a spatiotemporal descriptor based on blurred optical flow measurements to recognize actions. The use of features available from silhouettes is increasingly popular. Bobick and Davis [3] derived temporal templates from background subtracted images for human movement recognition. Blank *et al*. [7] utilized properties of the solution to the Poisson equation to extract features from space-time silhouettes for action recognition, detection and clustering.

The features should be simple, intuitive and reliable to extract without manual labour. As stated before, our work will use silhouettes as cues. Human silhouettes through the activity duration may be considered as points, and these points can be expected to lie on a low-dimensional manifold embedded in the high-dimensional image space. Therefore we are motivated to represent and analyze human motions in a more compact subspace rather than the ambient space.

A few promising methods for nonlinear dimensionality reduction have been recently proposed, e.g., isometric mapping (Isomap) [15], local linear embedding (LLE) [14], KPCA [16], to name just a few. Some researchers have explored these methods' applications, e.g., pose recovery [18] and visual tracking [19] in the area of human motion analysis. However, research on nonlinear manifold learning for complex activity recognition is still quite limited. The works of [18,19] usually obtained the embedded space from the same motion such as walking or running, but here we wish to learn the embedded activity space using all various motion classes.

HMM and its variants have been the dominant tools in human motion modelling [2,10], e.g., Nguyen *et al*. [2] learned and detected activities from movement trajectories using the hierarchical HMMs. Discriminative CRFs [20] were firstly introduced in the natural language processing community. Due to its merits over HMMs, there has recently been increasing interest in using CRFs for vision tasks, e.g., image region labeling [24], object segmentation [22], and gesture recognition [23]. A work closely related to this paper is [9], in which human motions were recognized using a linear-chain CRF based on motion capture data or image descriptors combining both shape context and pairwise edge features. Compared with [9], the features used in this work can be obtained more easily and reliably. In particular, our work further explores a better alternative method for modelling and recognizing human activities conditionally.

Being a joint graphical model with the richer structure, FCRF [21] has been demonstrated to be superior, in the chunking task, to the general linear-chain CRF model that does the individual labelling task sequentially. However, FCRF has not yet been used in the vision community. Our work will extend the original FCRF to model the kernel-induced motion trajectories where the underlying graphical model can capture long-range dependencies across frames.

3. Feature selection

Informative features are critical to the success of the

activity models. We select silhouettes as basic inputs, and perform nonlinear dimensionality reduction for more compact representation.

3.1. Silhouette extraction and representation

Given an action video $\mathcal{V}$ with T frames, i.e., $\mathcal{V} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_T\}$, our basic assumption is that the associated sequence of moving silhouettes $\mathcal{FS} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_T\}$ can be obtained from the original video. The size and the position of the foreground region vary with the distance of object to camera, the size of object and the performed activity. The silhouette images are thus centred and normalized on the basis of keeping the aspect ratio property of the silhouette so that the resulting images $\mathcal{NS} = \{\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_T\}$ contain as much foreground as possible, do not distort the motion shape, and are of equal dimensions $ri \times ci$ for all input frames. Fig. 2 (*top*) shows an example of the normalized silhouette images. Further, if we represent each raw silhouette image $\boldsymbol{R}_i$ as a vector $\boldsymbol{r}_i$ in $\mathcal{R}^{ri \times ci}$ in a row-scan manner, the whole video will be accordingly represented as $\mathcal{Vr} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_T\}$.

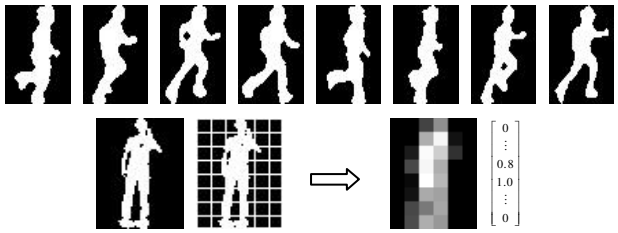

Figure 2. The normalized silhouette sequences of running (*top*) and the illustration of block-based feature representation (*bottom*)

For computational efficiency, we try the block-based features, as illustrated in Fig. 2 (*bottom*). We equidistantly divide each silhouette image into $h \times w$ non-overlapping sub-blocks. Then the normalized value of each sub-block is calculated by $\mathcal{N}_i = b(i)/mv$, $i = 1, 2, \ldots, h \times w$, where $b(i)$ is the number of foreground pixels in the ith sub-block, and mv means the maximum value of all $b(i)$. The resulting silhouette descriptor at frame t is $\boldsymbol{\mathcal{F}}_t = [\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_{h \times w}]^{\mathrm{T}}$ in $\mathcal{R}^{h \times w}$, and the whole video is accordingly represented as $\mathcal{VF} = \{\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \ldots, \boldsymbol{\mathcal{F}}_T\}$. In fact, a raw silhouette representation $\mathcal{Vr}$ may be considered as a special case of the block-based features, i.e., the sub-block size is 1×1, a pixel.

3.2. Nonlinear dimensionality reduction

To obtain compact description and efficient computation, we use the KPCA algorithm [16] to perform nonlinear dimensionality reduction, based on the following two considerations: 1) KPCA provides an efficient subspace learning method to discover the nonlinear structure of the 'action space'. Although it does not obviously consider the local manifold geometry, it can be related to Isomap and LLE in a kernel framework, as discussed in [17], and 2) although nonlinear methods such as Isomap and LLE do yield impressive results on some benchmark datasets, they are defined only on the training data points and how to map new data points remains unclear. In contrast, KPCA may be simply applied to any new data point.

Given a set of training samples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\mathcal{M}}\}$ in $\mathcal{R}^D$ with $\mathcal{M}$ elements, the aim of subspace learning is to find an embedding set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{\mathcal{M}}\}$ in a low-dimensional space $\mathcal{R}^d$ ($d < D$). For KPCA, each vector $\mathbf{x}_i$ is first nonlinearly mapped into the Hilbert space $\mathcal{H}$ by ϕ: $\mathcal{R}^D \rightarrow \mathcal{H}$. PCA is then applied on the mapped data $\mathcal{X}_\phi = \{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_{\mathcal{M}})\}$ in $\mathcal{H}$ [16]. Fortunately, this explicit mapping process is not required at all by the virtue of 'kernel tricks'. Let k be a semi-positive definite kernel function, and it defines a nonlinear relationship between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ by

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) \tag{1}$$

The problem of finding the coefficients of the principal components in $\mathcal{H}$ can be reduced to the diagonalization of the kernel matrix $\mathcal{K}$,

$$\mathcal{M}\lambda\mathbf{e} = \mathcal{K}\mathbf{e} \tag{2}$$

where $\mathcal{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{e} = [e_1, e_2, \ldots, e_{\mathcal{M}}]^{\mathrm{T}}$ so that $\mathcal{Z} = \sum_{i=1}^{\mathcal{M}} e_i \phi(\mathbf{x}_i)$.

The projection of a novel point $\mathbf{x}$ onto the jth principle axis z^j can be expressed implicitly as

$$(\mathcal{Z}^j \cdot \phi(\mathbf{x})) = \sum_{i=1}^{\mathcal{M}} e_i^j (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x})) = \sum_{i=1}^{\mathcal{M}} e_i^j k(\mathbf{x}_i, \mathbf{x}) \tag{3}$$

We use the Gaussian kernel function for our experiments.

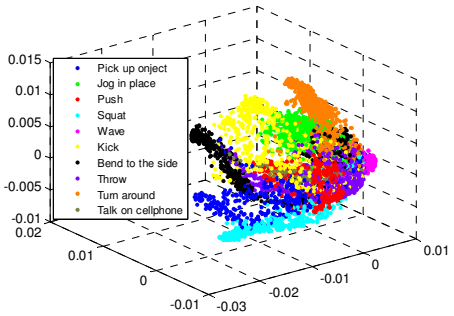

Figure 3. 3D visualization of PTMs in the KPCA-derived subspace, where the points with the same colors come from the same motion class

After obtaining the embedding space including the first d principal components, any one video $\mathcal{V}''$ can be projected into an associated trajectory in d-dimensional feature space $\mathcal{TO} = \{\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T\}$. Fig. 3 shows projection trajectories of motions (PTM) in case of Dataset I, in which temporal orders across frames are not labeled for clarity.

4. Motion modeling and recognition

The discriminative nature and underlying graphical structure of CRFs are very suitable for human motion analysis. Here we explore factorial CRF to label human activity sequences in the embedded space. To make this paper self-contained, we briefly review CRF and FCRF as follows (based on [20, 21]).

4.1. General CRF

The general framework of CRFs [20] is as follows. Let $\mathcal{G}$ be an undirected model over sets of random variables $\mathbf{s}$ and $\mathbf{o}$. Let $\mathbf{s}=\{s_t\}$ and $\mathbf{o}=\{o_t\}$, $t=1,\cdots,T$ so that $\mathbf{s}$ may be thought as a label sequence of an observed sequence $\mathbf{o}$. Let $C=\{\{\mathbf{s}_c,\mathbf{o}_c\}\}$ be the set of cliques in $\mathcal{G}$, then CRFs define the conditional probability of the state (or label) sequence given the observed sequence as

$$p_\theta(\mathbf{s}|\mathbf{o})=\frac{1}{Z(\mathbf{o})}\prod_{c\in C}\Phi(\mathbf{s}_c,\mathbf{o}_c) \tag{4}$$

where $Z(\mathbf{o})=\sum_{\mathbf{s}}\prod_{c\in C}\Phi(\mathbf{s}_c,\mathbf{o}_c)$ is a normalization factor over all state sequences, and Φ is a potential function which factorizes according to a set of features $\{f_n\}$ so that

$$\Phi(\mathbf{s}_c,\mathbf{o}_c)=\exp\left(\sum_{t=1}^{T}\sum_{n}\lambda_n f_n(\mathbf{s}_c,\mathbf{o}_c,t)\right) \tag{5}$$

where the model parameters $\theta=\{\lambda_n\}$ are a set of real weights, one weight per feature.

Previous studies mainly used linear-chain CRFs, as shown in Fig. 4 (*left*), where a first-order Markov assumption is generally made among labels. Accordingly, the cliques of such a conditional model are the nodes and edges, so there are feature functions $f_n(s_{t-1},s_t,\mathbf{o},t)$ for each label transition and $g_n(s_t,\mathbf{o},t)$ for each label.

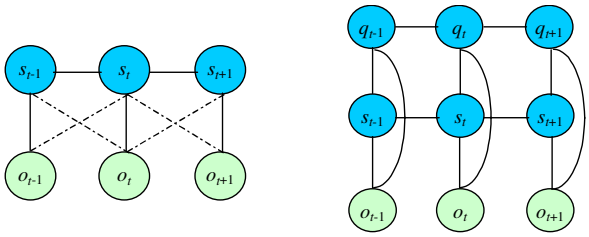

Figure 4. Graphical representation of (*left*) linear-chain CRF (if the dotted lines exit, it will represent a model with a context of 3 observation timesteps, i.e., $\omega=1$), and (*right*) two-chain factorial CRF including links between cotemporal labels, explicitly modeling limited probabilistic dependencies between two different label sequences.

4.2. Factorial CRF

Dynamic CRFs [21] are a generalization of linear-chain CRFs that repeat structure and parameters over a sequence of state vectors – allowing one to represent distributed hidden states and complex interactions among labels, as in a dynamic Bayesian network. As a special case, the factorial CRF has linear chains of labels with connections between cotemporal labels, thus increasing joint accuracy by information sharing.

Considering a FCRF with $\mathcal{L}$ Chains, where $s_{l,t}$ is the variable in chain l at time t. The distribution over hidden state is defined as

$$p(\mathbf{s}|\mathbf{o})=\frac{1}{Z(\mathbf{o})}\left(\prod_{t=1}^{T-1}\prod_{l=1}^{\mathcal{L}}\Phi_l(s_{l,t},s_{l,t+1},\mathbf{o},t)\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{\mathcal{L}-1}\Psi_l(s_{l,t},s_{l+1,t},\mathbf{o},t)\right) \tag{6}$$

where $\{\Phi_l\}$ are the potentials over the within-chain edges, $\{\Psi_l\}$ are the potentials over the between-chain edges [21], and these potentials factorize according to the features $\{f_k\}$ and weights $\{\lambda_k\}$ of $\mathcal{G}$, with the form of

$$\begin{aligned}\Phi_l(.)&=\exp\left(\sum_k\lambda_k f_k(s_{l,t},s_{l,t+1},\mathbf{o},t)\right)\\ \Psi_l(.)&=\exp\left(\sum_k\lambda_k f_k(s_{l,t},s_{l+1,t},\mathbf{o},t)\right)\end{aligned} \tag{7}$$

4.3. Training and inference

Given a set of the training samples $\mathcal{D}=\{\mathbf{o}^{(i)},\mathbf{s}^{(i)}\}_{i=1}^{N}$, the parameters $\theta=\{\lambda_k\}$ can be estimated by optimizing the following conditional log-likelihood function

$$\Omega(\theta)=\sum_i\log p_\theta(\mathbf{s}^{(i)}|\mathbf{o}^{(i)}) \tag{8}$$

The derivative of (8) with respect to λ_k associated with clique index c is

$$\frac{\partial\Omega}{\partial\lambda_k}=\sum_i\sum_t f_k(\mathbf{s}_{t,c}^{(i)},\mathbf{o}^{(i)},t)-\sum_i\sum_t\sum_{c\in C}\sum_{\mathbf{s}_c}p_\theta(\mathbf{s}_c|\mathbf{o}_t^{(i)})f_k(\mathbf{s}_{t,c},\mathbf{o}^{(i)},t) \tag{9}$$

where $\mathbf{s}_{t,c}$ denotes the variables of $\mathbf{s}$ at time step t in clique c of the 2-CRF, and $\mathbf{s}_c$ ranges over assignments to c.

Generally, a penalized likelihood function is used in training the parameters in order to reduce over-fitting, i.e., $\log p(\theta|\mathcal{D})=\Omega(\theta)+\log p(\theta)$ where $p(\theta)$ is a Gaussian prior over parameters ($p(\theta)\propto\exp\left(\frac{1}{2\varepsilon^2}\|\theta\|^2\right)$), so that the gradient becomes [21]

$$\frac{\partial p(\theta|\mathcal{D})}{\partial\lambda_k}=\frac{\partial\Omega}{\partial\lambda_k}-\frac{\lambda_k}{\varepsilon^2} \tag{10}$$

This convex function can be optimized by a number of techniques such as Quasi-Newton optimization methods.

Typically two inference problems need to be solved, i.e., computing the marginal $p(\mathbf{s}_{t,c}|\mathbf{o})$ over all cliques $\mathbf{s}_{t,c}$

and the Viterbi decoding $\tilde{\mathbf{s}} = \arg\max_{\mathbf{s}} p(\mathbf{s}|\mathbf{o})$. The former is used for parameter estimation, and the latter is used to label a new sequence.

4.4. Problem adaptation

A set of key poses can represent an action, and the sets of key poses show differences from action to action, though with possible partial overlap among activities. Key poses have been used to describe actions [11-13], e.g., Dedeoglu *et al*. [13] established the template pose dataset to contain key poses for all actions, and actions were represented as a histogram of key poses it matches. But such methods ignore temporal information between poses (as an intuitive example, they cannot tell reverse pairs of actions such as sitting down and standing up).

Assume that we wish to simultaneously perform both key-pose classification and activity classification. This problem can be solved by jointly representing these two tasks in a single graphical model: a two-chain FCRF, as shown in Fig. 4 (*right*), with one chain modelling the key-pose temporal process, and the other modelling the activity label, both representing their dependencies explicitly and preserving uncertainty between them. Accordingly, the problem is to learn a mapping of observations $\mathbf{o}$ to two different types of class labels, i.e., $s_1 \in \mathcal{S}_1 = \{1, 2, \cdots, N_c\}$ for N_c human activities and $s_2 \in \mathcal{S}_2 = \{1, 2, \cdots, K\}$ for K key poses, where $\mathbf{o}$ is a sequence of local observations (i.e., projection trajectory of a motion video here).

The basic point in creating the key pose dataset is to include as much key frames as possible for a specific action and at the same time to pay attention to make the distance between inter-key frames of different actions as much as possible [13]. We use the MDL (Minimum Description Length) rule to determine the number K of key poses in the whole dataset, and use K-means clustering to obtain these key poses $\mathcal{KP}=\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K\}$ for training process.

To incorporate long-range dependencies, we modify the potential function in (7) to include a window parameter ω that defines the amount of past and future history to be used when predicting the state at time t, i.e.,

$$o_t(t = t-\omega, \ldots, t+\omega) \mapsto s_t \qquad (11)$$

Following previous work [21], we adopt limited-memory BFGS optimizer to learn parameters with the variance of the prior $\varepsilon^2 = 10$, and loopy belief propagation is used for approximate inference. We factorize pair-wise features as $f_k(s_{t,c}, \mathbf{o}, t) = \mu_k(s_{t,c})\varphi_k(\mathbf{o}, t)$ where the former is a binary function on the assignment, and the latter is a function solely of the input features.

5. Experimental results

Extensive experiments have been carried out to evaluate the proposed framework. Note that the classification accuracy reported here is in terms of the percentage of the correctly recognized action sequences among all tests.

5.1. Activity datasets

There is no common evaluation database in the domain of human activity recognition. Here we use two recent databases reported in [7] and [8], respectively. These two databases are appreciably sized (among current databases publicly available), in terms of the number of subjects, actions and videos.

Dataset I: Different instances of the same activity may consist of varying relative speeds. Dataset I [8] consists of 10 different activities performed by one subject, i.e., pick up object, jog in place, push, squash, wave, kick, bend to the side, throw, turn around, and talk on cell phone, and 10 different instances for each activity. These activities were captured using two synchronized cameras that were about 45 degrees apart. The example images are shown in Fig. 5 (*top*). This dataset is used to systematically examine the effect of the temporal rate of execution (alone) on activity recognition (but also including slightly different intra-person motion styles among different instances).

Dataset II: In addition to the temporal rates, there exists inter-person difference between the same activities since different people have different physical sizes and perform activities differently in motion styles (and speeds). Dataset II [7] consists of 81 low-resolution videos (180×144, 25fps) from 9 different people, each performing 9 activities, i.e., bend, jump jack (jack), jump-forward-on-two-legs (jump), jump-in-place-on-two-legs (pjump), run, gallop-sideways (side), walk, wave-one-hand (wave1), and wave-two-hands (wave2). Together with one more recently added activity of skip, this dataset in total includes 10 activities and 90 videos. The sample images are shown in Fig. 5 (*bottom*). This dataset provides more realistic data for the test of the method's versatility with respect to variations at both temporal and spatial scales.

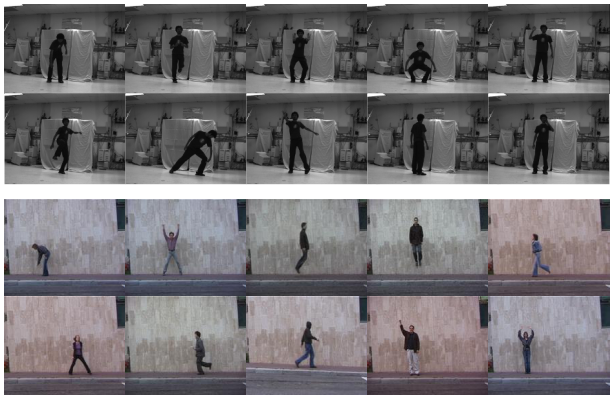

Figure 5. Example images from the activity datasets: from top to bottom and from left to right: Dataset I (*top*) - pick up object, jog

in place, push, squash, wave, kick, bend to the side, throw, turn around, talk on cell phone, respectively, and Dataset II (*bottom*) - bend, jack, jump, pjump, run, side, skip, walk, wave 1 and wave 2, respectively

5.2. Experimental procedure

We wish to compute an overall unbiased estimate of the recognition accuracy using the leaving-one-out validation method. We perform the round-robin activity recognition experiments. For Dataset I, we partition the dataset into 10 disjoint sets, each containing 1 instance of every activity. Each time we leave one set out for the test, and use the remaining nine sets to learn both subspace and model parameters. Similarly, for Dataset II, we divide the dataset into 9 sets, each set including all activities from one subject. To perform the recognition of each left-out set each time, we learn both subspace and model parameters from the remaining eight sets. Thus, if one video in the left-out test set is classified correctly, it must have a high similarity to a video from a different person performing the same activity.

Foreground detection is not our main concern in this work. We directly use the silhouette masks obtained in [7,8] for our experiments, though these silhouette images are not very satisfactory, including leaks and intrusions due to imperfect subtraction and shadows. We center and normalize all silhouette images into the same dimension (i.e., 64×48 pixels), and represent them as the block-based features with different sub-block sizes (e.g., 8×8, 4×4, 1×1). We learn FCRFs that model various-degree long-range dependencies between observations (e.g., ω=0 or 1). We empirically adjust the parameters of the reduced dimension d and the kernel width σ of KPCA under the supervision of the recognition rates.

5.3. Results and analysis

The activity recognition results on the above two datasets are clearly summarized in Table 1, where CCR means correct classification rates. From Table 1, we can basically draw the following conclusions: 1) Dynamic silhouette variations are indeed informative for analyzing human activities; 2) The proposed framework can effectively recognize human activities performed by different people with different body builds and different motion styles and speeds; 3) The recognition accuracy generally decreases as the sub-block size increases, especially quickly in case of 8×8 (note that the original silhouette image size is only 64×48 here); 4) Raw silhouette representation (i.e., the sub-block size is 1×1) performs best, though it is a little computationally intensive. This is because it keeps full information while other block-based features with bigger sizes considerably lose silhouette shape information; Although the block-based features might introduce some discretization errors, it gives an insight on how to select a good tradeoff between accuracy and computational cost in real applications; and 5) The introduction of long-range observations in the FCRF models generally improves the recognition accuracy (an exception, as bolded in table, may be due to over-fitting of parameter training).

Table 1. Accuracy of activity classification using FCRF

CCR / Size (D)	Dataset I (100 tests)		Dataset II (90 tests)	
	ω=0 (%)	ω=1 (%)	ω=0 (%)	ω=1 (%)
1×1 (3072)	100.0	100.0	94.44	97.78
4×4 (192)	**98.00**	**93.00**	87.78	92.22
8×8 (48)	71.00	78.00	73.33	77.78

5.4. Comparison

Currently, activity recognition is performed by either template matching or state-space approaches [1]. Here, we select a few schemes, namely one template matching method based on the Hausdorff distance metric and two state-space methods using HMM and linear-chain CRF respectively, to replace the FCRF model in the embedded space for the purpose of comparing their performance.

Hausdorff-based matching: The projection trajectory of a sequence can be simply considered as a point set. We adopt the symmetric mean Hausdorrf distance to measure the similarity between a test and all reference activities (i.e., templates). The test is recognized as the class of the reference template with the minimum dissimilarity value.

HMM model: We train an ergodic HMM model for each class of activity. Each model has five states and uses single Gaussian emission models. Test sequences are passed through each of these trained models, and the model with the highest likelihood is chosen as the recognized activity.

CRF model: We train a linear-chain CRF model where each class had a corresponding state. During evaluation, we perform the Viterbi decoding and assign the sequence label based on the most frequently occurring activity label per frame. We also carry out experiments that incorporate different long-range dependencies in the same way described in the FCRF experiments.

Table 2. Activity classification using different schemes

Models	Accuracy (%)
Hausdorff-metric based matching	82.00
Hidden Markov Model	89.00
CRF (ω=0)	92.00
CRF (ω=1)	95.00
Factorial CRF (ω=0)	100.0
Factorial CRF (ω=1)	100.0

Table 2 summarizes activity classification accuracies using different schemes on Dataset I with raw silhouette representation, from which it can be seen that: 1) Template matching performs worst. This may be due to its sensitivity to noisy features and the inability to explicitly capture temporal transition; 2) State-space methods generally

outperform template matching-based method, though they are computationally expensive; 3) Both CRF and FCRF have better performance than HMM, which shows that discriminative models are generally superior to generative models; 4) FCRF performs better than CRF, even when long-range dependencies are not considered, which demonstrates the advantage of jointly discriminative learning by information sharing between different label sequences; and 5) Performance of both CRF and FCRF is improved with increased window sizes, which shows that incorporating long-range dependencies is useful.

5.5. Robustness test

We construct two experiments for robustness test with respect to silhouette quality and other challenging factors. Note that the results reported below are in terms of raw silhouette representation and FCRF with $\omega = 1$.

Noise-corrupted silhouettes: Though being imperfect, the silhouette masks used for the above experiments are relatively smooth. A simple method to check sensitivity to silhouette quality is to add various amounts of synthetic noise to silhouette images to simulate corrupted silhouettes. Since the silhouette image is binary, we use 'salt & pepper' noise. A parameter, the noise density, is used to represent the percentage of the affected pixels in the whole image, as shown in Fig. 6 (*top*). We use original (uncorrupted) silhouette sequences for training, and the noise-corrupted silhouette sequences for testing. The results are shown in Fig. 6 (*bottom*), from which we can see that the proposed framework can tolerate a considerable amount of noise (e.g., 25%). This is probably because the statistical nature of FCRF renders overall robustness to both representation and recognition.

Noise Density	CCRs (%)
0.05	100.0
0.10	100.0
0.15	100.0
0.20	100.0
0.25	97.00
0.30	90.00
0.35	72.00
0.40	63.00
0.45	42.00
0.50	27.00

Figure 6. Silhouette images with different degrees of synthetic noise (*top*): From left to right: noise densities are respectively 0, 0.1, 0.2, 0.3, 0.4, and 0.5, and the accuracies of activity classification with respect to different noise densities (*bottom*)

Other factors: We also consider the robustness of our method with respect to other factors such as different clothes, occlusion and motion styles. The walking action is one of the most common motions in real life. Here we test 10 walking sequences captured in different scenarios [7] on Dataset II. Some example images and the associated silhouettes are shown in Fig. 7. In contrast to synthetic noise-corrupted silhouettes, the silhouettes here exhibit deformations of human shapes produced by realistic variations, compared to normal walking pattern.

Figure 7. From left to right and from top to bottom: diagonal walk, walk with a dog, walk and swing a bag, walk in a skirt, walk with the legs occluded partially, sleepwalk, limp, walk with knees up, and walk when carrying a briefcase, respectively

Table 3 summarizes the test results including the first and second best matches, from which it can be seen that, except for four sequences, all other test sequences are correctly classified as the 'walk' action. This shows that the proposed method has relatively low sensitivity to considerable changes in scale, clothes, partial occlusion, and irregularities in walking forms.

Table 3. Robustness evaluation with respect to other factors

Test sequences	Varying conditions	Results
Diagonal walk	Scale and Viewpoint	**Pjump** (walk)
Walk with a dog	Non-rigid deformation	**Run** (skip)
Walk and swing bag	Rigid deformation	**Skip** (walk)
Walk in a skirt	Clothes	Walk (side)
Walk with occluded legs	Partial occlusion	Walk (jump)
Sleepwalk	Walking style	**Side** (skip)
Limp	Walking style	Walk (jump)
Walk with knees up	Walking style	Walk (jump)
Walk/Carry briefcase	Carried object	Walk (skip)
Normal walk	Background	Walk (skip)

5.6. Discussion and future work

Although we could currently not provide a theoretical explanation for why the additional complexity of FCRF is favored for human activity recognition, there are marked

improvements in recognition accuracy in our experiments. Further performance evaluation is still needed on larger and more realistic datasets, in order to be conclusive.

We have explored the KPCA for discovering the action space, but its performance is somewhat dependent on the selection of the width of the kernel function, as well as the kernel function itself. How to automatically set the optimal parameters involved will be investigated.

Different cues have various discriminative abilities, e.g., silhouettes, shapes, trajectories, optical flow, etc. Here we have only tried the simple silhouette cue. It is conceivable that fusion of multiple cues is ideal for improving the algorithm's effectiveness and robustness.

It would be interesting to systematically investigate how long-range observations should be considered for optimal recognition. We have used a model with simple first-order state dependency, but it would also be interesting to study longer-range state dependencies, e.g., trigrams. In addition, generative and discriminative models have their own advantages in modelling temporal sequences. The effective combination of both models is also a part of future work.

6. Conclusion

This paper has described an effective probabilistic framework for human activity recognition in monocular video. The novelty of the method is two-fold: a) in feature extraction and representation, we selected simple but easy-to-extract space-time silhouettes as inputs, and embedded them into a low-dimensional kernel-derived space; and b) in activity modeling and recognition, we presented the first use of FCRF in the vision community, and demonstrated its superiority to both HMM and general CRF. The proposed framework is not dependent on the features used, so we believe that it can be easily extended to other types of temporal data analysis.

Acknowledgment

The authors would like to thank Moshe Blank and Ashok Veeraraghavan for providing the datasets. Special thanks also go to Charles Sutton, Yang Wang, Robin Li, Sy Bor Wang, and Asela Gunawardana for their valuable discussion on the theory and implementation of CRF and FCRF.

References

[1] J. K. Aggarwal and Q. Cai, Human motion analysis: A review, *CVIU*, 73 (3) (1999): 428-440.

[2] N. Nguyen, D. Phung, S. Venkatesh, and H. Bui, Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models, *CVPR*, 2005.

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

[4] A. Efros, A. Berg, G. Mori, and J. Malik, Recognizing action at a distance, *ICCV*, 2003.

[5] C. Schuldt, I. Laptev, and B. Caputo, Recognizing human actions: a local SVM approach, *ICPR*, 3 (2004): 32–36.

[6] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, Human activity recognition using multidimensional indexing, *TPAMI*, 24(8) (2002): 1091-1104.

[7] M. Blank *et al*., Action as space-time shapes, *ICCV*, 2 (2005): 1395-1402.

[8] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury, The function space of an activity, *CVPR*, 2006.

[9] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, Conditional models for contextual human motion recognition, *ICCV*, 2 (2005): 1808-1815.

[10] M. Brand, N. Oliver and A. Pentland, Coupled hidden Markov models for complex action recognition, *CVPR*, 1996.

[11] S. Carlsson and J. Sullivan, Action recognition by shape matching to key frames, *Workshop on Models versus Exemplars in Computer Vision*, 2001.

[12] R. Collins, R. Gross, and J. Shi, Silhouette-based human identification from body shape and gait, *AFG*, 2002.

[13] Y. Dedeoglu, B. Toreyin, U. Gudukbay, and A. Cetin, Silhouette-based method for object classification and human action recognition in video, *HCI*, 2006, pp. 64-77.

[14] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, 290 (2000): 2323–2326.

[15] J.B. Tenenbaum, V. de Silva, and J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, 290 (2000): 2319-2323.

[16] B. Scholkopf, A. Smola, and K. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10 (1998): 1299-1319.

[17] J. Ham, D. Lee, S. Mika, and B. Scholkopf, A kernel view of dimensionality reduction of manifolds, *ICML*, 2004.

[18] A. Elgammal and C-S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, *CVPR*, 2 (2004): 681-688.

[19] C. Sminchisescu and A. Jepson, Generative modelling for continuous non-linearly embedded visual inference, *ICML*, 2004, pp. 140-147.

[20] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, *ICML*, 2001, pp. 282–289.

[21] C. Sutton, K. Rohanimanesh, and A. McCallum, Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data, *ICML*, 2004.

[22] Y. Wang and Q. Ji, A dynamic conditional random field model for object segmentation in image sequences, *CVPR*, 2005.

[23] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, Hidden conditional random fields for gesture recognition, *CVPR*, 2006.

[24] S. Kumar and M. Herbert, Discriminative random fields: a framework for contextual interaction in classification, *ICCV*, 2003.

---

† Liang Wang is currently with the Department of Computer Science and Software Engineering, The University of Melbourne, Australia. This work is supported by the ARC Centre of Perceptive and Intelligent Machines in Complex Environments.