Moving Shape Dynamics: A Signal Processing Perspective

Liang Wang[1]*, Xin Geng[2], Christopher Leckie[1], Ramamohanarao Kotagiri[1]
[1]Department of Computer Science & Software Engineering
The University of Melbourne, Vic 3010, Australia
{lwwang,caleckie,rao}@csse.unimelb.edu.au

[2]School of Engineering & Information Technology
Deakin University, Vic 3125, Australia
xge@deakin.edu.au

Abstract

This paper provides a new perspective on human motion analysis, namely regarding human motions in video as general discrete time signals. While this seems an intuitive idea, research on human motion analysis has attracted little attention from the signal processing community. Sophisticated signal processing techniques create important opportunities for new solutions to the problem of human motion analysis. This paper investigates how the deformations of human silhouettes (or shapes) during articulated motion can be used as discriminating features to implicitly capture motion dynamics. In particular, we demonstrate the applicability of two widely used signal transform methods, namely the Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT), for characterization and recognition of human motion sequences. Experimental results show the effectiveness of the proposed method on two state-of-the-art data sets.

1. Introduction

Vision-based human motion analysis has received growing interest, based on a wide range of potential applications such as visual surveillance and human-machine interfaces. Generally, motions fall under two distinct categories, namely *composite* motions that can be divided into different temporal segments and *primitive* motions that cannot be further decomposed. This paper focuses on primitive motion recognition from short videos, which is an important first step towards the problem of action segmentation and profiling in long videos.

Background and motivation: Human motion recognition remains a challenge due to variations in both the *environment* (which determines the quality of visual clues extracted from videos) and the *motion itself* (which exhibits spatial and temporal variations due to subjects with different physical characteristics, motion styles and speeds). Accordingly, *motion characterization and recognition* are central to the interpretation of human motions.

Various visual features have been used to describe motion formation, *e.g.*, *trajectory-based features* [10, 1] from tracking of positions, velocities and joint angles; *intensity-based features* such as local descriptors of interest points [12] and optical flow [5]; and *silhouette-based features* such as motion history images [3] and silhouette (or contour) volumes [2, 18]. Feature tracking is not yet well solved for unconstrained human motions due to the great variability in shape and articulation of the human body. Features based on intensities depend heavily on the imaging conditions. Human motions can be considered as temporal variations of moving silhouettes (or shapes) [3, 2, 16]. In this paper, we use information that can be derived from space-time silhouettes to characterize *moving shape dynamics*. In particular, we embed dynamic silhouette data using a new graph embedding framework, *i.e.*, Kernel Locality Preserving Projections (KLPP) [6], to simultaneously address the high-dimensionality and non-linearity of articulated and deformable moving shapes.

Different strategies have been proposed for motion modeling and recognition, which fall into three major categories: *template-based methods* [12, 1, 3] convert time-varying features into a static pattern (*i.e.*, template) for comparison to pre-stored prototypes during recognition; *direct sequence matching* uses techniques such as Hausdorff distance [8], Dynamic Time Warping (DTW) [15] and spatiotemporal correlation [5] on time-varying features without further feature extraction; and *state-space methods* either use linear models such as ARMA (Autogressive Mov-

*This work is partially supported by the ARC Discovery Project (Grant No. DP0663196)

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

ing Average) models [15] or graphical models such as HMMs (Hidden Markov Models), CRFs (Conditional Random Fields) and their variants [13, 10, 16] to model motions. However, direct sequence matching has high computational cost. The motion is not described by an explicit form, so it is necessary to compute pairwise distances between a test sequence with all stored sequences during recognition, which is impractical for large data sets. State-space models generally involve complex mathematical and statistical computation. Our method falls into the first category. From a general perspective, human motions in video can be regarded as discrete time signals, reflecting temporal variations in observations of interest across image frames. Therefore it is intuitive to use digital signal processing techniques for extracting the signal properties. The DWT (Discrete Wavelet Transform) and DFT (Discrete Fourier Transform) have been widely studied and used in the signal processing community [9]. However, to the best of our knowledge, they have received little or no attention in the context of vision-based human motion recognition. Hence an important open question is how to characterize and recognize human motions by means of available signal transform techniques.

Aim and contributions: The aim of this paper is to investigate how to build signal-transform-based feature extraction for moving shape dynamics that can support characterization and recognition of human motions from videos. To address the nonlinearity and high-dimensionality of dynamic shape data, KLPP is adopted to embed shapes into a low-dimensional subspace, in which a motion sequence is mapped to a multi-dimensional discrete temporal signal. A Fourier Transform (FT) or Wavelet Transform (WT) is then applied to extract the signal properties which provide a compact motion description. Experimental results on two recent data sets demonstrate the effectiveness and efficiency of our method.

Our main contributions are summarized as follows:

- We investigate the use of shape deformations of human silhouettes as discriminating motion features, and demonstrate the applicability of KLPP for learning a low-dimensional embedding space of dynamic data.
- We exploit the appropriateness of DFT/DWT for human motion characterization and recognition. Moreover, DFT/DWT provide a general framework that can be used to characterize the time evolution of any set of features including the deforming silhouettes here.
- Experimental results validate the feasibility and merits of the proposed method. The combination of dynamic shape manifolds plus DFT/DWT-based signal characterization achieves competitive results on state-of-the-art data sets compared with other algorithms.
- The proposed method has several ideal properties: 1) it is easy to understand and implement, while very effective and efficient; 2) it does not need complex feature tracking; 3) it avoids the use of computationally expensive spatiotemporal matching and state-space models; 4) the use of motion descriptors reduces the original temporal classification problem into a static classification one, enabling the use of any existing efficient static classifier; and 5) as a non-statistical method, large training data are not required. Each example characterizes the motion type, and the discriminative features can be derived from the example itself.

The remainder of this paper is organized as follows. Section 2 introduces the DFT and DWT. Section 3 details the proposed method. Experimental results are provided in Section 4, prior to conclusions in Section 5.

2. DFT and DWT

The transformation of signals to the frequency domain often provides a more effective representation and a more computationally efficient approach to processing of signals compared to time domain processing. To put our method into context, we briefly review two widely-used signal transforms, namely DFT and DWT (discrete versions of FT and WT). Both express the signal as coefficients in a function space spanned by a set of basis functions. For FT, the basis contains only the complex exponential function representing sinusoids, while the basis of WT consists of infinitely many scaled and shifted versions of a mother wavelet function. More details can be found in [7, 9].

DFT: The Fourier transform measures global frequencies of a signal. Let $x(t), t = 0, 1, \cdots, n-1$ be a n-point discrete signal. The DFT of $x(t)$ is defined by

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x(t) e^{\frac{-j2\pi tf}{n}}, \ f = 0, 1, \cdots, n-1 \quad (1)$$

where $j = \sqrt{-1}$. The signal energy is defined as the sum of the energies at each point of the signal. According to Parseval's theorem, $\sum_{t=0}^{n-1} |x(t)|^2 = \sum_{f=0}^{n-1} |X_f|^2$. To obtain real-time computation, a fast Fourier Transform (FFT) is often used which makes use of the symmetry properties of a DFT. If n is an integer of the power of 2, the computation complexity of the FFT is $O(n\log n)$.

Signals are often real-valued in most real applications. Accordingly, except for X_0 representing the DC component of the signal, X_f are complex numbers and represent the *amplitudes* and *phase shifts* of a decomposition of the signal into sinusoid functions, satisfying $X_{n-f} = X_f^*, \ f = 1, \cdots, n-1$, where the asterisk denotes complex conjugation. That is, the FT of a real-valued signal is symmetric $|X_{n-f}| = |X_f|$, *i.e.*, every amplitude at the beginning

except the first one also appears at the end. So the first $\lceil (n+1)/2 \rceil$ DFT coefficients completely encode the signal.

DWT: Wavelet transforms measure frequencies at different time resolutions. A wavelet is a smooth and quickly vanishing oscillating function with good localization in both frequency and time. A wavelet family is a set of functions generated by dilations and translations of a mother wavelet

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k),\ j,k \in \mathbb{Z} \tag{2}$$

The function ψ is the orthogonal wavelet if $\psi_{j,k}$ is the orthogonal basis of $\mathrm{L}^2(\mathbb{R})$, satisfying $\langle \psi_{j,k}, \psi_{l,m} \rangle = \delta_{j,l} \cdot \delta_{k,m}$. The signal $x(t)$ can be represented as

$$x(t) = \sum_{j,k} c_{j,k}\psi_{j,k}(t) \tag{3}$$

where $c_{j,k} = \langle \psi_{j,k}(t), x(t) \rangle$ is called the *wavelet coefficient* of $x(t)$. The WT provides high frequency resolutions at low frequencies and high time resolutions at high frequencies.

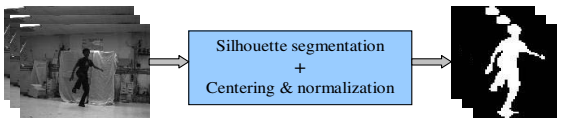
The DWT has been widely used in denoising and compression of images and signals [7, 9]. In the case of the DWT, a time-scale representation is obtained using digital filtering techniques. In the pyramidal algorithm of Mallat [7], the signal is analyzed at different frequency bands by decomposing the signal into coarse *approximation* and *detail* information. The approximation is then further decomposed using the same decomposition step. This is achieved by successive high-pass and low-pass filtering of the signal.The resulting DWT coefficients describe the signal in terms of an approximation of the original signal, plus a set of details that range from coarse to fine. The trend of the signal is preserved in the approximate part, while the localized changes are kept in the detail parts. For the DWT using the Haar wavelet, the computational complexity is $O(n)$.

3. Methodology

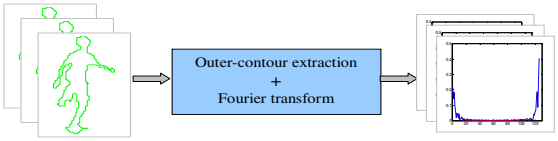
3.1. Extraction of moving shapes

The shape changes captured in individual silhouettes over time naturally provide information about the motion being performed. The use of silhouettes has several advantages: 1) silhouettes are simple and intuitive while containing rich shape information of a moving human; 2) silhouettes are easier to extract than to detect/track body parts; and 3) binary silhouettes are insensitive to texture and color within the foreground region. This step aims to convert motion information in raw video to an associated sequence of shape features, which implicitly reflect motion dynamics.

Given a motion video $\mathcal{V}$ with T frames I, *i.e.*, $\mathcal{V} = [\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_T]$, an associated sequence of moving silhouettes $\mathcal{S} = [\mathrm{S}_1, \mathrm{S}_2, \cdots, \mathrm{S}_T]$ can be obtained by motion segmentation techniques. For simplicity, we use the following two methods to describe each silhouette, as illustrated in Figure 1.



(a) Normalized silhouette images



(b) Shape Fourier descriptors

Figure 1. Extraction and representation of moving shapes

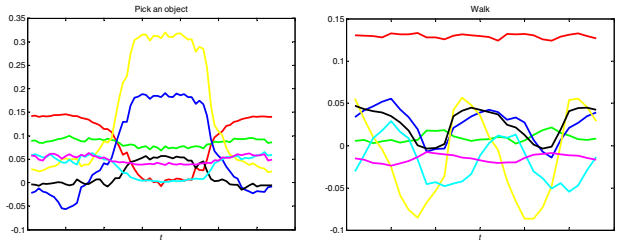
Normalized silhouette image (NSI): The size and position of the foreground region vary with the distance of the human from the camera, the human size and the motion being performed. Silhouette images are thus centered and normalized on the basis of keeping the aspect ratio of the silhouettes so that the resulting images contain as much foreground as possible, do not distort the moving shape, and have the same dimensions of $n_1 \times n_2$. Accordingly, each NSI can be denoted by an $h = n_1 \times n_2$ dimensional vector $\mathbf{f}_{\mathrm{NSI}}$ in a row-scan manner.

Shape Fourier descriptor (SFD): A human shape can be described by N points $\{(\mu_0, \nu_0), \cdots, (\mu_{N-1}, \nu_{N-1}\}$ on the silhouette's outer-contour, denoted in form of centered complex coordinates, *i.e.*, $\omega_i = (\mu_i - \mu_c) + j \cdot (\nu_i - \nu_c)$, where $(\mu_i, \nu_i), i = 0, 1, \cdots, N-1$ are the pixel coordinates of the i-th pixel along the silhouette boundary and (μ_c, ν_c) is the shape centroid, *i.e.*, $\mu_c = \frac{1}{N}\sum_{i=0}^{N-1} \mu_i,\ \nu_c = \frac{1}{N}\sum_{i=0}^{N-1} \nu_i$. The DFT is applied to $[\omega_0, \cdots, \omega_{N-1}]$ to generate Fourier coefficients $\{\Omega_0, \Omega_1, \cdots, \Omega_{N-1}\}$. The resulting SFD is an $h = N - 2$ dimensional vector $\mathbf{f}_{\mathrm{SFD}} = (|\Omega_2|/|\Omega_1|, \cdots, |\Omega_{N-1}|/|\Omega_1|)$, which is invariant to translation, scale and rotation.

3.2. Subspace of shape features

Human motion is inherently highly-nonlinear and high-dimensional. To address the curse of dimensionality and to find a discriminative data representation, we wish to embed the dynamic shape features into a low-dimensional subspace. Spectral methods have recently emerged as a powerful tool for dimensionality reduction and manifold learning. LPP (Locality Preserving Projection) has been demonstrated to be superior to PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) in image-based static face recognition [6]. However it remains unclear how well linear LPP approximates a nonlinear Laplacian Eigenmap (LE). Here we use the kernel extension of LPP, *i.e.*, KLPP [6], to address high-dimensionality and non-linearity of articulated motion data simultaneously.

Given m samples $\{\mathbf{f}_i\}_{i=1}^m \subset \mathbb{R}^h$ from a training set ($\mathbf{f}_i$ can be $\mathbf{f}_{\mathrm{NSI}}$ or $\mathbf{f}_{\mathrm{SFD}}$), dimensionality reduction aims at find-



(a) (b)

Figure 2. Examples of video projections into multi-dimensional time series signals in the embedding space

ing $\{\mathbf{g}_i\}_{i=1}^m \subset \mathbb{R}^l$, $l \ll h$, where $\mathbf{g}_i$ can 'represent' $\mathbf{f}_i$. Assume these m samples are represented by a weighted graph $\mathcal{G}$ with m vertices. Let W be a symmetric $m \times m$ matrix with w_{ij}, the weight of the edge joining vertices i and j. Two kinds of weighting strategies are often used. One is to use a simple 0-1 rule, *i.e.*, $w_{ij} = 1$ if and only if vertices i and j are "close"; otherwise 0. The other is the heat function $w_{ij} = e^{-(\|\mathbf{f}_i - \mathbf{f}_j\|^2/\tau)}, \tau \in \mathbb{R}$. "Close" can be defined by the k-nearest neighbors. LPP aims to find a transform matrix E so as to map $\mathbf{g}_i = E^T\mathbf{f}_i$, based on the generalized eigenvalue problem

$$FLF^T\mathbf{e} = \lambda FDF^T\mathbf{e} \tag{4}$$

where D is a diagonal matrix whose entries are column (or row) sums of W, *i.e.*, $D_{ii} = \sum_j w_{ji}$. $L = D - W$ is the graph Laplacian matrix, and $F = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_m]$. Let the column vectors $\mathbf{e}_0, \cdots, \mathbf{e}_{l-1}$ be the solutions of (4), ordered according to their eigenvalues $\lambda_0 < \cdots < \lambda_{l-1}$. The embedding is represented by

$$\mathbf{g}_i = E^T\mathbf{f}_i,\ E = [\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_{l-1}] \tag{5}$$

To generalize LPP to the nonlinear case, assume that the Euclidean space $\mathbb{R}^h$ is mapped to a Hilbert space $\mathbb{H}$ through a nonlinear mapping function $\phi : \mathbb{R}^h \rightarrow \mathbb{H}$. The eigenvector problem in the Hilbert space can be written as

$$[\phi(F)L\phi^T(F)]\upsilon = \lambda[\phi(F)D\phi^T(F)]\upsilon \tag{6}$$

This can be solved using the following eigenvector problem

$$KLK\alpha = \lambda KDK\alpha \tag{7}$$

where K is the $m \times m$ Gram matrix with elements $K_{ij} = K(\mathbf{f}_i, \mathbf{f}_j) = \langle\phi(\mathbf{f}_i)\cdot\phi(\mathbf{f}_j)\rangle$ and $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_m]^T$. For a new point $\mathbf{f}$, its projection onto the eigenvectors υ^κ is

$$(\upsilon^j \cdot \phi(\mathbf{f})) = \sum_{i=1}^m \alpha_i^j \langle\phi(\mathbf{f}_i)\cdot\phi(\mathbf{f}_j)\rangle = \sum_{i=1}^m \alpha_i^j K(\mathbf{f}, \mathbf{f}_i) \tag{8}$$

where $\alpha_1, \cdots, \alpha_m$ are the solutions of (7), and α_i^j is the i-th element of the vector α^j. Our experiments use the Gaussian kernel function, *i.e.*, $K(\mathbf{f_i}, \mathbf{f_j}) = e^{-(\|\mathbf{f_i}-\mathbf{f_j}\|^2/2\sigma^2)}$ with a scale parameter σ.

After learning the KLPP-based subspace, one shape sequence corresponding to video $\mathcal{V}$ can be projected into a semantically-meaningful trajectory $\mathcal{P}$ in such an embedding space $\mathcal{P} = [P_1, P_2, \cdots, P_T]$, $P_i \in \mathbb{R}^l$, while the temporal order across frames is preserved explicitly. We regard such a trajectory $\mathcal{P}$ as a form of l-dimensional discrete time signal. Figure 2 gives two examples, in which only the first 7 dimensions are shown for clarity, and each dimension is marked using the same color.

3.3. Characterization of shape dynamics

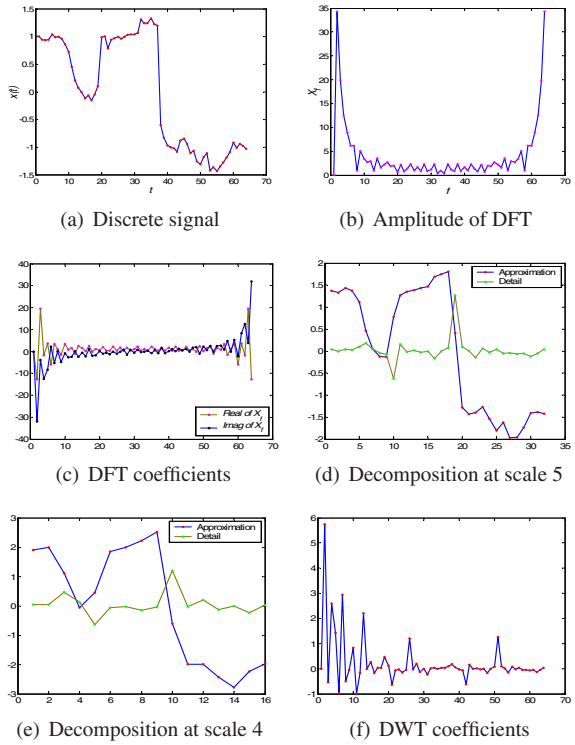
We can represent the i-th dimension of an l-dimensional trajectory $\mathcal{P}$ by $x^i(t), t = 0, 1, \cdots, n-1, i = 1, \cdots, l$. Since the DC component of a signal is useless to represent the signal shape, we normalize the signal to normal form to remove its effect by $\hat{x}^i(t) = (x^i(t) - \bar{x^i})/\varepsilon^i$, where $\bar{x^i}$ and ε^i are respectively the mean and standard deviation of the signal $x^i(t)$. By using an appropriate transformation of a signal, our aim is to identify redundant components of the signal that can be discarded without significant loss of information. We consider several schemes to obtain compact descriptors that can be used to describe motion trajectories.

DFT-based motion descriptors: For the real-valued signal $\hat{x}^i(t)$, we compute its DFT coefficients $X^i = [X_0^i, X_1^i, \cdots, X_{n-1}^i]$. After signal normalization, the first coefficient $X_0 = 0$. We may use its amplitude spectra $|X^i|$ by selecting the first $r (r \leq \lceil n/2 \rceil)$ coefficients to represent the signal, *i.e.*, $A_i^{(1)} = [|X_1^i|, |X_2^i|, \cdots, |X_r^i|]$. The resulting motion descriptor is an $l \times r$ dimensional vector, *i.e.*, $M^{(1)} = \{A_i^{(1)}\}_{i=1}^l$.

The use of only the amplitude of DFT coefficients makes $A^{(1)}$ shift-invariant according to the shift theorem of DFT. However, the phase information may be useful for describing the signal. We also directly use the DFT coefficients to represent the signal, *i.e.*, $A_i^{(2)} = [X_1^i, X_2^i, \cdots, X_r^i]$, leading to the motion descriptor $M^{(2)} = \{A_i^{(2)}\}_{i=1}^l$.

DWT-based motion descriptors: A commonly used wavelet, *i.e.*, Haar wavelet, is chosen in our experiments for the following reasons: it allows good approximation with a subset of coefficients; it can be quickly computed with the complexity $O(n)$; and the Haar wavelet transform can be seen as a series of *averaging* and *differencing* operations on a discrete time signal, leading to simple coding.

Assume the signal $\hat{x}^i(t) \in \mathbb{R}^n$ is located in the scale $J (J = \log n)$. After decomposing $\hat{x}^i(t)$ at a specific scale $j \in \{0, \cdots, J-1\}$, wavelet coefficients can be represented as $c_j(\hat{x}^i(t)) = \{a_j^i, d_j^i, \cdots, d_{J-1}^i\}$. a_j^i is called the approximation coefficient which is the projection of $\hat{x}^i(t)$ and $d_j^i, \cdots, d_{J-1}^i$ are the wavelet coefficients representing detail information of $\hat{x}^i(t)$. The approximation co-



(a) Discrete signal (b) Amplitude of DFT

(c) DFT coefficients (d) Decomposition at scale 5

(e) Decomposition at scale 4 (f) DWT coefficients

Figure 3. Illustration of DFT and DWT of a 64-point signal $x(t)$

efficients within lower scales correspond to the lower frequency part of the signal. The first few coefficients of $c_0(\hat{x}^i(t))$ (*i.e.*, full decomposition) corresponding to the low frequency part can be viewed as a noise-reduced signal, thus using these coefficients will retain most of the information in the signal. We select the first r coefficients of $c_0(\hat{x}^i(t))$ as the signal feature $A_i^{(3)}$, creating the motion descriptor $M^{(3)} = \{A_i^{(3)}\}_{i=1}^{l}$.

Wavelet-based representation may be performed on different time and frequency resolutions. It is thus possible to measure the signal at different resolutions concurrently. We also keep all the approximation coefficients within a specific scale j as the signal features, *i.e.*, $A_i^{(4)} = [a_j^i]$, which retain the entire information of $\hat{x}^i(t)$ at a particular level of granularity. The task of choosing the first few wavelet coefficients in case of $A_i^{(3)}$ is thus circumvented by choosing a particular scale j in this case, and the time dimension of the signal is reduced to $r = 2^j$. The resulting motion descriptor is $M^{(4)} = \{A_i^{(4)}\}_{i=1}^{l}$.

Figure 3 shows examples of the DFT and DWT of a 64-point signal in its normal form (a), from which we can see that the symmetric energy spread of the DFT coefficients suggests that most energy is preserved in the low-frequency and high-frequency coefficients but not in the mid-frequency ((b) and (c)), and the hierarchical energy of the DWT coefficients suggests that most energy is preserved in the low resolution coefficients (f). This suggests that it is reasonable to keep the first few coefficients with most energy to represent the signal sketch. Such motion descriptors have several advantages: 1) most of the energy of the signal can be represented by only a few coefficients, which greatly reduces time dimensionality; 2) approximate coefficient representations of the signal in case of DWT preserve time order information to some degree ((d) and (e)); 3) discarding partial detail parts of DWT coefficients or relatively middle-frequency DFT coefficients acts as denoising; and 4) DFT and DWT have fast computation algorithms, and the availability of motion descriptors converts the original multi-dimensional time series classification problem to a relatively easier static classification problem.

3.4. Recognition of shape dynamics

Although many methods exist for static classification problems, we choose the simplest Nearest-Neighbor (NN) classifier because our main concern is to examine the discriminatory powers of our motion descriptors.

Let $y^i(t), i = 1, \cdots, l$ represent each dimension trajectory of one test sequence in the embedding space, $z_q^i(t)$ is the corresponding trajectory signal of the q-th reference motion sequence, and their motion descriptors are respectively M_y and M_{z^q}, where M can be any one of $M^{(1)}$, $M^{(2)}$, $M^{(3)}$, and $M^{(4)}$. We measure their dissimilarity by the Euclidean distance, *i.e.*, $s_q = \|M_y - M_{z^q}\|$. To incorporate time shifting, especially since motion video alignment has not been solved (*e.g.*, in the case D-II in our experiments), we modify the similarity measure to

$$\hat{s_q} = \min_b \|M_y^b - M_{z_q}\|, \; b = 1, \cdots, n-1 \qquad (9)$$

where M_y^b represents the corresponding motion descriptor of the circularly shifted signal $y_\circ^i(t-b)$ of $y^i(t)$. Note that for $M^{(1)}$, this process is not needed according to the shift theorem of the DFT. The test sequence is classified as the class label c_q of the q-th reference motion sequence with the lowest dissimilarity value, *i.e.*, $c_{test} = \arg\min_{c_q} \hat{s_q}$.

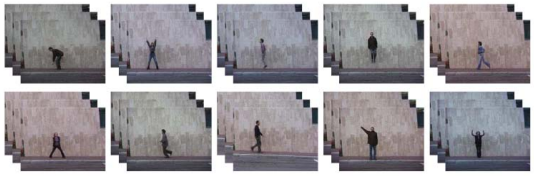
4. Experiments

Standard evaluation databases are unavailable in the area of human motion analysis. We select two recent databases to evaluate the proposed method, which are appreciably sized in terms of the number of persons, motions and video sequences, as shown in Figure 4.

4.1. Evaluation video databases

Maryland dataset (D-I) [14] includes 10 actions performed by one person, 10 instances per action, thus 100 sequences in total. These actions are respectively pick-up-object, jog-in-place, push, squash, wave, kick, bend-to-the-side, throw, turn-around, and talk-on-cell-phone. This data

(a) D-I



(b) D-II

Figure 4. Examples of motion videos in the data sets used

set is used to examine the effect of the temporal rate of execution on motion recognition, as discussed in [14].

Weizmann dataset (D-II) [2] is initially composed of 81 low-resolution videos, including 9 actions performed by 9 different people. These actions are respectively run, walk, jump-jack, jump-forward-on-two-legs, jump-in-place-on-two-legs, gallop-sideways, wave-two-hands, wave-one-hand. Together with a later added action (*i.e.*, skip), there are 10 actions per subject, in total 90 sequences. Compared to D-I, D-II is more realistic and challenging, including inter-person variations due to different physical structures, motion styles and performing speeds.

4.2. Experimental procedure and data processing

To obtain an unbiased estimation of accuracy on small data sets, we perform a series of leave-one-out recognition experiments. For each data set, each time we leave one sequence out for testing, and use the remaining for training. For D-I, the start and end of each motion are basically consistent in each sequence, so we do not consider time shifting. In D-II, people perform each motion multiple times in a continuously repetitive manner (except for bend). Each video generally includes $2 \sim 4$ full cycles of atomic motions, which allows us to compute each motion's duration by periodicity analysis. The real video length is selected as the duration for bend, while for other actions, we select two complete cycles from the middle part of each original video for our experiments. When selecting cycles, we do not, and need not, temporally assign an onset and ending for each class of action. Accordingly we need to consider time shifting when computing the sequence similarity.

These two data sets are provided with silhouette masks. The quality of these masks is generally good, but many defects are also present, *e.g.*, shadows, partially missing body parts, *etc*. We center and normalize silhouette images into 64×48 resolution, leading to $h = 3072$ for the NSI representation. We sample $N = 128$ points along the silhouette's outer-contour, leading to $h = 126$ for the SFD representation. Given a training set to learn KLPP, we use the k-nearest neighbors ($k = 15$) to construct the affinity graph and the 0-1 weighting rule to generate the weight matrix W. Computationally efficient spectral regression [4] is adopted to solve KLPP, in which the Gaussian kernel function's argument σ is selected as the mean of pairwise distances of training samples excluding self-dissimilarities. Each sequence is projected into an l-dimensional trajectory signal in the embedding space, from which we extract its $r \times l$-dimensional motion descriptors. To enable fast computation of DFT and DWT, we temporally scaled the trajectory length to $n = 64$ for D-I and $n = 32$ for D-II (Actually the length of each sequence is 80 frames in D-I, and an average of about 38 frames in D-II). We implement the nearest-neighbor classifier as a baseline classification. Correct Classification Rates (CCR) are measured with respect to both the reduced space dimension $l \in \{1 \sim 30\}$ and the reduced time dimension $r \in \{1 \sim n/2\}$ simultaneously.

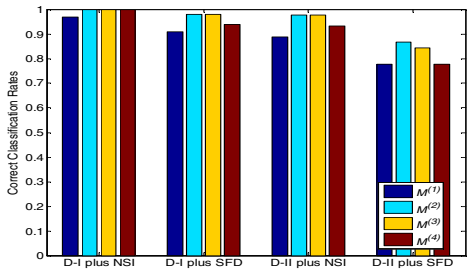
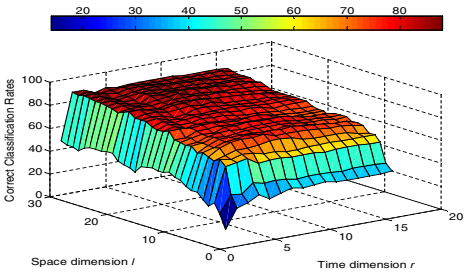


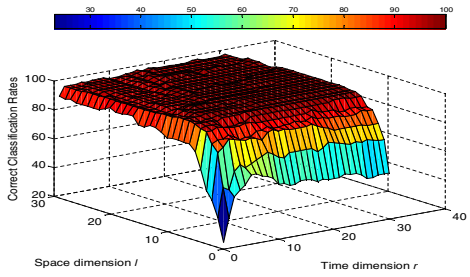
Figure 5. CCRs with respect to different configurations

4.3. Classification results and analysis

We report the best results of our method obtained within the ranges of l and r in Figure 5. The CCR reported here is in terms of the percentage of the correctly recognized motion sequences among all test sequences. For $M^{(4)}$, we use the 3^{rd} decomposition scale, *i.e.*, $r = 8$. From Figure 5, we can conclude that: 1) Space-time shapes are very informative and rich, as demonstrated by the relatively high classification rates achieved. 2) D-I is more easily classified. This is probably because the same motions, even from different instances, are performed by the same person, thus there are comparatively fewer shape changes among silhouette sequences. 3) D-II is relatively harder to classify because these motions are performed by different people with different body builds and motion styles; 4) Overall, the NSI representation is better than SFD. This may be due to unreliability in shape-based representation, especially when we only describe outer-contours of imperfect silhouettes. 5) Overall, DWT and DFT perform similarly. This seems to be consistent with the conclusion of Wu *et al*. [17] that DFT and Haar DWT are comparable in energy preservation. 6) The representations of $M^{(4)}$ and $M^{(1)}$ are a little worse in precision than the corresponding $M^{(3)}$ and $M^{(2)}$.



(a) D-II plus NSI plus DFT plus $M^{(2)}$



(b) D-I plus NSI plus DWT plus $M^{(3)}$

Figure 6. CCRs vs. reduced space and time dimensions

In summary, our method can effectively recognize the motion sequences with inter-person and intra-person variations in both temporal and spatial scales.

The recognition rates are dependent on the reduced space dimension l and the reduced time dimension r. We show two examples reflecting CCRs with respect to l and r in Figure 6, from which we can see that: 1) Generally when l is fixed, DFT and DWT perform better as r increases; but after reaching a peak, CCRs will slightly decrease as r further increases. This is probably because when r is too big, though the preserved energy is somewhat increased, it also introduces considerable noise. 2) When r is fixed, as l increases, DFT and DWT perform better until CCRs finally reach a relatively steady curve. 3) Our method does not need a high space or time dimension to obtain satisfactory performance, thus leading to low computational cost.

4.4. Comparison of different subspace methods

We compare several unsupervised methods, *i.e.*, PCA, Kernel PCA [11] and linear LPP [6], to examine which method is better for learning dynamic shape manifolds. We report their best results using NSI representations within $l \in \{1 \sim 30\}$ and $r \in \{1 \sim n/2\}$. From Table 1, we can see that: 1) Overall, spectral methods (LPP and KLPP) outperform PCA and KPCA, and KLPP performs best. LPP and KLPP have relatively more compact visual clustering effects in the embedding space than PCA and KPCA. This is probably because they implicitly emphasize natural clusters in the data by trying to preserve the neighborhood structure. Although their advantages on these data sets are not apparent, it could be expected that they would perform well for larger data sets. 2) Although motion measurements are

Methods	CCR on D-I (%)		CCR on D-II (%)	
	$M^{(2)}$	$M^{(3)}$	$M^{(2)}$	$M^{(3)}$
PCA	99.0	99.0	92.2	92.2
KPCA	99.0	99.0	94.4	92.2
LPP	100	100	94.4	95.6
KLPP	100	100	97.8	97.8

Table 1. Comparison of different subspace learning methods

inherently nonlinear, linear PCA provides good discrimination rates, just slightly lower than KPCA and LPP. This is probably because, although in subspace learning, we treat each frame shape as an independent sample regardless of temporal information, in the process of recognition, the action is not considered as one single entity but a sequence of entities, thus the re-introduction of the temporal relation increases the discriminating power. 3) We do not show a significant improvement in accuracy using nonlinear KPCA and KLPP over their corresponding linear PCA and LPP, but this is not entirely surprising and can be explained. Temporal information is more important than individual shapes. These subspace learning methods ignore temporal relations across frames. However, the preservation of temporal order information in sequence projection and recognition seems to compensate for such limitations to some degree.

4.5. Comparison of different recognition algorithms

We compare several recent motion recognition algorithms, especially those using the same evaluation data sets. Blank *et al*. [2] utilized the solution to the Poisson equation of space-time silhouette volume to extract various shape properties for action classification. Veeraraghavan *et al*. [14] learned a function space of time warping for each activity. Wang and Suter [16] used factorial CRF to model and recognize actions. Ali *et al*. [1] used chaotic invariants of motion trajectories for action recognition. Unlike the sequence-based evaluation methods in [1, 16], the method in [2] used a sliding window with a fixed size to extract space-time cubes for tests. We cite the results reported by these algorithms in Table 2, in which D-II* means a subset of D-II without the skipping action (*i.e.*, 81 videos). It can be seen that our method performs better than that of [1] and achieves comparable results to [16, 14, 2]. The advantages of our method are the simplicity and reliability of extraction and characterization of motion features, avoiding explicit tracking for feature extraction (and, hence, their complexities and brittleness) and the use of computationally complex state-space models. In particular, the computational efficiency of the signal transforms used is very competitive.

4.6. Discussion

We have conducted additional experiments to assess the effects of parameter settings. For $n_1 \times n_2$, we found that

Methods	Data sets	Accuracies
Veeraraghavan06 [14]	D-I	100%
Wang07 [16]	D-I	100%
Our method	D-I	100%
Blank05 [2]	D-II*	99.6%
Ali07 [1]	D-II*	92.6%
Wang07 [16]	D-II	97.8%
Our method	D-II	97.8%

Table 2. Comparison of different recognition algorithms

128×96, 96×72 and 64×48 made little difference. For N, 64, 128 and 256 also gave similar results. When using k-nearest neighbors to construct the affinity graph in KLPP, we found that k can be reliably selected in the range of $10 \sim 25$. How to set an optimal σ in the Gaussian kernel function is an open question. We empirically set it to be the mean value of the pairwise distances and obtained satisfactory results. For n, we found that 16, 32, 64, and 128 perform similarly on current data. These observations suggest that our method is insensitive to parameter settings. The Discrete Cosine Transform (DCT) is another form of signal transform technique, which is a Fourier-related transform similar to DFT, but using only real numbers. Experiments using DCT gave similar results to DFT. Although the time-frequency localization property of DWT suggests wavelet representation of signals may bear more information than that of DFT, experimental results on current data sets show little difference. We also tried several other orthogonal wavelets and observed that Haar wavelets far outperform Daubechies and Coiflet wavelets in accuracy, as well being computationally less expensive. This means that not all wavelets are suitable for analyzing motion data derived from short shape sequences since the effectiveness of the power concentration of a particular transform depends greatly on the nature of time signals.

Current work only focuses on short video analysis including atomic motions. We plan to extend this work to automatic action segmentation and detection in long videos containing composite motions. How to detect transient changes between temporal segments and how to perform subsequence matching are issues for future research.

5. Conclusion

This paper describes an effective approach to silhouette-based motion recognition. The method investigates human silhouette deformations during the articulated motion as discriminating features to implicitly capture moving shape dynamics. In particular, the method exploits the applicability of DFT and DWT for characterization and recognition of human motion sequences. Experimental results based on state-of-the-art data sets have validated our method.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proc. Intl. Conf. Computer Vision*, 2007. 1, 7, 8

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. Intl. Conf. Computer Vision*, pages 1395–1402, 2005. 1, 6, 7, 8

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 1

[4] D. Cai, X. He, and J. Han. Spectral regression for dimensionality reduction. Technical Report UIUCDCS-R-2007-2856, University of Illinois at Urbana-Champaign, May 2007. 6

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. Intl. Conf. Computer Vision*, 2003. 1

[6] X. He and P. Niyogi. Locality preserving projections. In *Proc. Advances in Neural Information Processing Systems 16*, 2003. 1, 3, 7

[7] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1999. 2, 3

[8] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, 2003. 1

[9] G. J. Miao and M. A. Clements. *Digital Signal Processing and Statistical Classification*. Artech House, 2002. 2, 3

[10] N. Nguyen, D. Phung, S. Venkatesh, and H. Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. In *Proc. Intl. Conf. Computer Vision and Pattern Recognition*, 2005. 1, 2

[11] B. Scholkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, (10):1299–1319, 1998. 7

[12] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. Intl. Conf. Pattern Recognition*, volume 3, pages 32–36, 2004. 1

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *Proc. Intl. Conf. Computer Vision*, volume 2, pages 1808–1815, 2005. 2

[14] A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury. The function space of an activity. In *Proc. Intl. Conf. Computer Vision and Pattern Recognition*, 2006. 5, 6, 7, 8

[15] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(12):1896–1909, 2005. 1, 2

[16] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graph model. In *Proc. Intl. Conf. Computer Vision and Pattern Recognition*, 2007. 1, 2, 7, 8

[17] Y. Wu, D. Agrawal, and A. Abbadi. A comparison of dft and dwt based similarity search in time-series databases. In *Proc. ACM Intl. Conf. Information and Knowledge Management*, pages 488–495, 2000. 6

[18] A. Yilmaz and M. Shah. Action sketch: A novel action representation. In *Proc. Intl. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 984–989, 2005. 1