LEARNING SILHOUETTE DYNAMICS FOR HUMAN ACTION RECOGNITION

Guan Luo and Weiming Hu

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China

ABSTRACT

In this paper, we address the problem of recognizing human actions with motion dynamics alone. For this purpose, we propose to use silhouette sequences to represent the human actions by discarding the appearance information, and then model the sequences with linear dynamical systems (LDSs). Recognition is achieved by directly comparing the distance between LDSs, rather than resorting to complex Bayesian learning and inference. In particular, we introduce an efficient optimization method to learn robust LDSs, and develop a shift invariant distance metric to measure the similarity on the LDSs space. We evaluate our approach on the human action data set and achieve comparable results.

Index Terms— Action recognition, linear dynamical system, silhouette, similarity measurement

1. INTRODUCTION

Analysis of human activities has always been an active research area in computer vision. Over the past couple of decades, a large amount of algorithms have been proposed for this task. In terms of the action representation, previous work can be roughly classified into two aspects: appearance-based approaches and motion-based approaches. The former usually characterizes the motion sequence with various local [1, 2] or global [3] visual features extracted from raw video data. The major problem in these approaches is that they discard the temporal information inherent to actions and thus fail to capture the temporal dynamics of human activities. The latter generally models the motion sequence with state-space models [4, 5] by viewing the human action recognition as a temporal classification problem. These approaches are of comparatively high complexity and require detailed statistical modeling and parameter learning.

It is known that appearance and dynamics are two important cues for human action recognition. Since much of the previous work has focused on appearance cues, our goal in this paper is to consider motion dynamics alone. For this purpose, we propose to use silhouette images which are insensitive to the subject appearance to represent the human actions. Thus we can focus on how to learn the intrinsic dynamics of silhouette sequences.

In recent years, system theoretic methods to recognition of human actions [3, 6] have attracted much interest, inspired by the work in dynamic texture literature [7]. By modeling the motion temporal variation with dynamical systems, system theoretic methods specifically consider the global dynamics of activities. Among all the methods, linear dynamical systems (LDSs) are used broadly for the simplicity and efficiency. Once each action sequence is characterized by a LDS model, the similarity between two LDSs is directly measured with a distance or kernel metric defined on the LDS space. After all pairwise similarities are evaluated on the training data, classifiers such as nearest neighbors or SVM can be used to categorize the testing video sequences.

In this paper, we propose to learn robust LDSs to describe the dynamics of silhouette sequences. We emphasize that stability is a crucial property for LDSs, while it is commonly omitted by most of previous work [3, 7]. We also develop a shift invariant distance metric based on the subspace angles distance, which is insensitive to the starting frame of motion sequences. We show that our method achieves encouraging results for the task of human action recognition.

The remainder of the paper is structured as follows: Section 2 first briefly introduces LDS and its parameter learning methods. Then a robust LDS learning algorithm is introduced and analyzed. Finally a shift invariant distance metric is proposed to measure the similarity between LDSs. We conduct experiments to evaluate the performance of our method in Section 3 and give our conclusions in Section 4.

2. RECOGNITION WITH ROBUST LDS

Dynamical system methods have been studied extensively in fields ranging from control engineering to visual process. For instance, dynamic texture represent the texture's temporal variation as a LDS [7]. In graphical model's perspective, LDS is indeed a generative state-space model with Gaussian observations and Markov states. For human action sequence, many inherent nonlinearities such as phase transition, turbulence and delay can be eliminated by choosing proper coordinates or mapping into high-dimensional spaces [6]. Therefore in this paper, we will focus on how to model the human motion dynamics with LDSs.

This work is partly supported by NSFC (Grant No. 61272330, 60935002), the Beijing Natural Science Foundation (4121003), and the Guangdong Natural Science Foundation (Grant No. S2012020011081).

2.1. Linear Dynamical System

Let $A \in \mathbb{R}^{n\times n}$ denote the system dynamic matrix, and $C \in \mathbb{R}^{p\times n}$ denote the subspace mapping matrix. Here p and n are the dimensions of the observation space and the state space, respectively. Then a stationary LDS can be represented by the tuple parameter $\mathbf{M} \doteq (A, C)$ and evolves in time according to the following equations

$$\begin{cases} x_{t+1} = Ax_t + v_t \\ y_t = Cx_t + w_t \end{cases} \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state or latent variable, $y_t \in \mathbb{R}^p$ is the observed random variable or feature, v_t and w_t are the system noise and observation noise, respectively. If we assume the noises are zero-mean i.i.d Gaussian processes, then we have $v_t \sim \mathcal{N}(0, Q)$ and $w_t \sim \mathcal{N}(0, R)$. Here Q and R are covariant matrices of multivariate Gaussian.

Given a video sequence $y_{1:\tau}$, learning the intrinsic dynamics amounts to identifying the model parameter $\mathbf{M}$. This is a typical system identification problem and there are normally two ways to solve it: maximum likelihood estimation and least squares estimation.

Let $Y_{1:\tau} = [y_1, y_2, ..., y_\tau]$ and $X_{1:\tau} = [x_1, x_2, ..., x_\tau]$ represent the original observation sequence and state sequence, respectively. For the least squares estimation method, the model parameter $\mathbf{M}$ is computed by firstly decomposing observation matrix $Y_{1:\tau} \approx U\Sigma V^T$ with SVD to obtain an estimate of the underlying state sequence

$$\hat{C} = U \quad \hat{X}_{1:\tau} = \Sigma V^T \tag{2}$$

Then the least squares estimation of A is

$$\hat{A} = \arg\min_A \|AX_{1:\tau-1} - X_{2:\tau}\|_F^2 = X_{2:\tau}X_{1:\tau-1}^{+} \tag{3}$$

According to (2), LDS implicitly models $Y_{1:\tau}$ with a set of subspaces matrix C and its corresponding coefficients $X_{1:\tau}$. In action recognition task, the subspaces matrix C describes the action appearance, while matrix A derived from the state sequence $X_{1:\tau}$ represents the motion dynamics. Thus we can use $\mathbf{M} = (A, C)$ to represent the motion sequence descriptor. Such a descriptor captures both the dynamics and appearance of human action sequence, which is much different from local spatio-temporal gradient descriptors. However, there exist two problems when using $\mathbf{M}$ as the descriptor. The first one is that the traditional LDS solvers ignore the stability of dynamical systems. This may result in a degenerate LDS model. The second one is that the descriptor $\mathbf{M} = (A, C)$ lives in a non-Euclidean space and is in non-vector form. There is no a straightforward way to compute the non-vector descriptor distance in a non-Euclidean space.

2.2. Learning Robust LDS

Stability is a very important property for LDSs. An unstable LDS may become degenerate quickly and fail to generate long sequences which share the same characteristics as the training data. We show in Section 2.3 that generating consistent sequences is critical for our shift invariant distance metric.

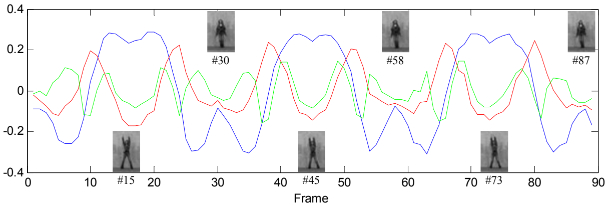


Fig. 1. Periodic state components and sample frame images.

In the field of dynamical systems, stability means the poles of the model are all inside the complex unit circle. If the poles are on the unit circle, the system is called to be marginally stable. Marginally stable systems are very useful as they generate sustained oscillations in the output, which in our case describes the periodic patterns in motion sequences. In Fig. 1, we show the three state components trajectories of a jacking sequence learned with robust LDSs, which indicates that robust LDSs capture the intrinsic periodic mode.

Let $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ denote the eigenvalues of dynamic matrix A in decreasing order of magnitude. A LDS is called stable or marginally stable if and only if $\lambda_1 \leq 1$. However, traditional LDS learning method of (3) does not enforce this stability criterion. This may cause the solution to be unstable. However, most of previous work employ this approach to learn LDSs, thus the similarity results do not stand firmly.

In the system identification literature, stability has been intensively studied. However, most of the methods are computationally expensive to reach the optimization result. In our task, we prefer an efficient approximation solution if only it satisfies the stability criterion. Here we introduce a constraint generation method [8], which achieves the stable result efficiently by iteratively checking stability criterion and generating new constraints.

For the least squares problem in (3), it can be reformulated by expanding polynomial as

$$\hat{A} = \arg\min_a \{a^T Pa - 2q^T a + r\} \tag{4}$$

where $a = vec(A)$, $q = vec(X_{1:\tau-1}X_{2:\tau}^T)$, $P = I_n \otimes (X_{1:\tau-1}X_{1:\tau-1}^T)$, $r = tr(X_{2:\tau}^T X_{2:\tau})$. Here $vec(\cdot)$ is a linear operator which flattens the matrix to vector in column order.

From the stability criterion, the constraint is obtained by decomposing $\hat{A}$ with SVD $\hat{\Sigma} = \hat{U}^T \hat{A} \hat{V}$ and inferring as

$$\tilde{\lambda}_1 = tr(\tilde{u}_1^T \hat{A} \tilde{v}_1) = tr(\tilde{v}_1 \tilde{u}_1^T \hat{A}) = g^T \hat{a} \leq 1 \tag{5}$$

where $g = vec(\tilde{u}_1 \tilde{v}_1^T)$, and $\hat{a} = vec(\hat{A})$.

Therefore, the quadratic program can be written as

$$\begin{array}{ll} \text{minimize} & a^T Pa - 2q^T a + r \\ \text{subject to} & g^T a \leq 1 \end{array} \tag{6}$$

which can be solved efficiently with *quadprog*.

2.3. Shift Invariant Similarity Metric for LDS

Given two motion sequences, we use the robust LDS parameters $\mathbf{M}_1 = (A_1, C_1)$ and $\mathbf{M}_2 = (A_2, C_2)$ to represent the motion sequence descriptors. Since the model space has a non-Euclidean structure and the descriptors are in non-vector form, this naturally raises the issue of how to measure the similarity between these two descriptors. De Cock and De Moor [9] propose to compare dynamical models by using the subspace angles between two systems. The subspace angles are obtained by solving the Lyapunov equation

$$\mathcal{Q} = \mathcal{A}^T \mathcal{Q} \mathcal{A} + \mathcal{C}^T \mathcal{C} \tag{7}$$

where $\mathcal{Q} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$, $\mathcal{A} = \begin{pmatrix} A_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix}$, $\mathcal{C} = (C_1 \quad C_2)$.

The solution of (7) is guaranteed when the systems are *stable*. The cosines of the subspace angles $\cos^2\theta_i$ are calculated as eigenvalues of matrix $Q_{11}^{-1}Q_{12}Q_{22}^{-1}Q_{21}$.

Therefore the subspace angles distance is defined as

$$d_{LDS}(\mathbf{M}_1, \mathbf{M}_2)^2 = -\log \prod_{i=1}^{n} \cos^2\theta_i \tag{8}$$

However, the experiments for human action recognition show that the subspace angles distance varies greatly when two sequences have only temporal shift (see Fig. 2). We are prone to a similarity measure that is insensitive to the initial state. In other words, two walking sequences should be classified into the same category no matter what frame they begin with. Hence we develop an offset alignment strategy by evolving each sequence for τ steps so that the similarity between them is maximized. That is

$$d(\mathbf{M}_1, \mathbf{M}_2) = \min_{\tau_1, \tau_2 \in \mathbb{N}} d_{LDS}(\mathbf{M}_1(\tau_1), \mathbf{M}_2(\tau_2)) \tag{9}$$

where $\mathbf{M}(\tau)$ denotes the model parameter of evolved sequence which is generated by shifting the original one τ steps ahead. We notice that the evolved sequences should keep the same characteristics as the original one. In order to achieve this purpose, the original model must be *stable*.

It is unfortunate that there is no an explicit way to obtain the optimization solution of (9). However in many applications the periods of most motion patterns are short. Thus we can handle this problem by searching through all the combinations of τ_1 and τ_2 exhaustively. Assuming the maximal shift is $\mathcal{T}$, the complexity of this problem is $O(\mathcal{T}^2)$.

We show in Fig. 2 that our aligned shift invariant distance outperforms the traditional one in two aspects. First, the aligned distance shows higher similarity than the original subspace angles distance. This results in better recognition results as illustrated in Fig. 4. Second, the aligned distance shows a more stable similarity measure that is insensitive to the starting frame, whereas the subspace angles distance shows sudden changes in some conditions.

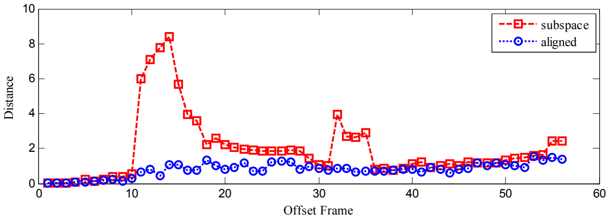


Fig. 2. Subspace versus aligned distances.

3. EXPERIMENTS

To evaluate the performance of our proposed method for human action recognition, we carry out detailed experiments on the Weizmann data set.

The Weizmann data set consists of 90 video sequences from 9 different people, each performing 10 natural actions. These actions include bending, jumping jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, running, galloping-sideways, skipping, walking, waving-one-hand, and waving-two-hands.

For learning the global dynamics, we prefer a sequence representation that is insensitive to the subject appearance. So we directly use the foreground silhouettes. Since some of the silhouettes are rather noisy, we introduce a shape feature [10] which is robust for our motivation. Given a silhouette image, we first compute two orthogonal projection lines passing through its center of mass. Then we uniformly divide the bounding box of silhouette into m bins on horizontal and vertical directions, respectively. Finally we encode the shape feature as the average distance of the points on the silhouette from projection lines for each bin. In our experiments, we use $m = 16$ on both sides of projection lines and obtain a shape feature with 64 components. In Fig. 3, we show a sample silhouette sequence and the corresponding shape features. We can see that though the silhouettes are noisy, the shape features suppress the defects effectively.

In Fig. 4, we examine the relationship between the correct classification rates and the model dimension n up to 20. We also evaluate the effectiveness of our proposed shift invariant distance metric compared with the standard subspace angles distance. Average results are reported based on the leave-one-out cross-validation method, from which we can see that: 1) the recognition rates do not change much with respect to the model dimension. This means that we can choose a comparatively small model dimension, say $n = 3$, to gain nearly the same results with much less computation cost; 2) the proposed shift invariant distance always performs better than the traditional subspace angles distance. It achieves a best recognition rate of 96.67% with $n = 3$, which is almost as well as the state-of-the-art recognition result of 97.83% [11].

In Fig. 5, we show the confusion matrix of action classification by LDSs when the model dimension $n = 3$. We can see that LDSs make misclassification primarily among

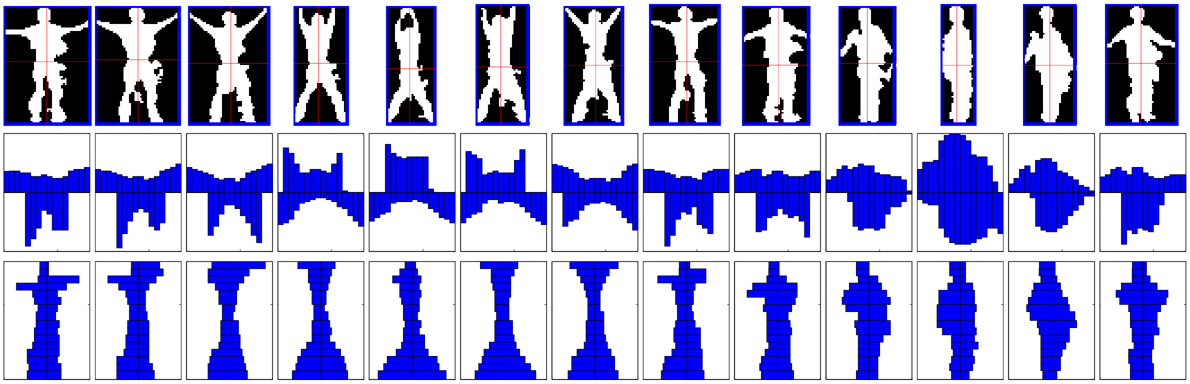


Fig. 3. Silhouette sequence and associated shape features.

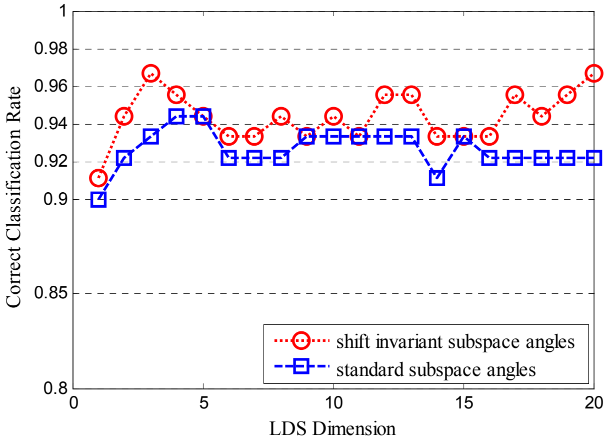
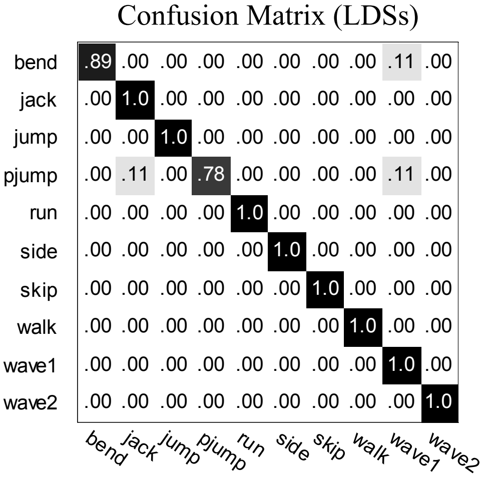


Fig. 4. Recognition rate versus LDS dimension.



Confusion Matrix (LDSs)

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	.89	.00	.00	.00	.00	.00	.00	.00	.11	.00
jack	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
jump	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00
pjump	.00	.11	.00	.78	.00	.00	.00	.00	.11	.00
run	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
side	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
skip	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00
walk	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
wave1	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00
wave2	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

Fig. 5. Confusion matrix of action classification.

Table 1. Comparison of recognition rates.

Methods	**Best accuracy**
Ours - LDSs	96.67%
Gorelick *et al.* [11]	**97.83%**
Ali and Shah [12]	95.75%
Chaudhry *et al.* [3]	95.66%
Niebles *et al.* [13]	90.00%
Bregonzio *et al.* [14]	96.66%

the actions ’bend’, ’pjump’, ’jack’, and ’wave1’. This means that these actions share high degrees of similarity in terms of dynamics in some aspects.

Table 1 compares the performance of the state-of-the-art methods on the Weizmann data set. We can see that our approach achieves comparable recognition results to the best method [11]. In addition, [12] extracts a set of kinematic features such as divergence, vorticity, etc., from the optical flow for human action recognition. [3] uses Binet-Cauchy kernels to capture the non-linear dynamics of histograms of optical flow to recognize human action. Compared with these dynamics methods, our approach outperforms them in both accuracy and efficiency.

4. CONCLUSIONS

In this paper, we have proposed to model and recognize silhouette dynamics with robust LDSs. This is inspired by the fact that most of the previous work focuses on appearance information, or models the human action with unstable dynamical systems. We have proposed to learn robust LDSs with a simple yet efficient suboptimal algorithm, and then developed a shift invariant distance metric to measure the similarity between LDSs. We have validated our method on the human action data set and achieved encouraging results.

5. REFERENCES

[1] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatiotemporal features," in *VSPETS*, 2005, pp. 65–72.

[3] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, 2009, pp. 1932–1939.

[4] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *CVPR*, 1992, pp. 379–385.

[5] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *ICCV*, 2005, pp. 1808–1815.

[6] A. Bissacco, A. Chiuso, and S. Soatto, "Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport," *IEEE Transactions on PAMI*, vol. 29, no. 11, pp. 1958–1972, 2007.

[7] G. Doretto, A. Chiuso, Y.N. Wu, and S. Soatto, "Dynamic textures," *IJCV*, vol. 51, no. 2, pp. 91–109, 2003.

[8] S.M. Siddiqi, B. Boots, and G.J. Gordon, "A constraint generation approach to learning stable linear dynamical systems," in *NIPS*, 2007.

[9] K. De Cock and B. De Moor, "Subspace angles between ARMA models," *Systems and Control Letter*, vol. 46, pp. 265–270, 2002.

[10] F. Cuzzolin, "Using bilinear models for view-invariant action and identity recognition," in *CVPR*, 2006.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on PAMI*, vol. 29, no. 12, pp. 2247–2253, 2007.

[12] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on PAMI*, vol. 32, no. 2, pp. 288–303, 2010.

[13] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[14] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *CVPR*, 2009, pp. 1948–1955.