Learning Human Actions by Combining Global Dynamics and Local Appearance

Guan Luo, Shuang Yang, Guodong Tian, Chunfeng Yuan, Weiming Hu, *Senior Member, IEEE*, and Stephen J. Maybank, *Fellow, IEEE*

Abstract

In this paper, we address the problem of human action recognition through combining global temporal dynamics and local visual spatio-temporal appearance features. For this purpose, in the global temporal dimension, we propose to model the motion dynamics with robust linear dynamical systems (LDSs) and use the model parameters as motion descriptors. Since LDSs live in a non-Euclidean space and the descriptors are in non-vector form, we propose a shift invariant subspace angles based distance to measure the similarity between LDSs. In the local visual dimension, we construct curved spatio-temporal cuboids along the trajectories of densely sampled feature points and describe them using histograms of oriented gradients (HOG). The distance between motion sequences is computed with the Chi-Squared histogram distance in the bag-of-words framework. Finally we perform classification using the maximum margin distance learning method by combining the global dynamic distances and the local visual distances. We evaluate our approach for action recognition on five short clips datasets, namely Weizmann, KTH, UCF sports, Hollywood2 and UCF50, as well as three long continuous datasets, namely VIRAT, ADL and CRIM13. We show competitive results as compared with current state-of-the-art methods.

Index Terms

Action recognition, linear dynamical system, local spatio-temporal feature, non-vector descriptor, distance learning.

1 Introduction

Analysis of human activities concerns detecting, tracking, recognizing and understanding human movements from image sequences. Among these problems, human action recognition is one of the most active research areas and has attracted much research interest over the past couple of decades. The surveys by Turaga *et al.* [1] and Poppe [2] provide a broad overview of numerous approaches for analyzing human motion sequences in videos. In terms of the motion sequence representation, previous work can be roughly classified into appearance-based methods and motion-based methods. Appearance-based methods usually characterize the motion sequence with various local [3], [4], [5],

- *G. Luo, S. Yang, G. Tian, C. Yuan and W. Hu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No. 95, Zhongguancun East Road, PO Box 2728, Beijing, 100190, P.R. China. E-mail: {gluo, syang, gdtian, cfyuan, wmhu}@nlpr.ia.ac.cn.*
- *S. J. Maybank is with the Department of Computer Science and Information Systems, Birkbeck College, Malet Street, London WC1E 7HX, United Kingdom. E-mail: sjmaybank@dcs.bbk.ac.uk.*

[6] or global [7], [8], [9], [10] visual features extracted from raw video data. For example, Niebles *et al.* [6] use a bag-of-words model to represent human actions. The bag-of-words model is learned by extracting and clustering local spatio-temporal interest points. The major problem in these methods is that they discard the temporal information inherent to actions and thus fail to capture the temporal dynamics of human activities. Motion-based methods generally model the motion sequence with temporal state-space models [11], [12], [13], [14] and view the human action recognition as a temporal classification problem. For example, Yamato *et al.* [11] propose to recognize tennis shots by modeling a sequence of grid-based silhouette features as outputs of class-specific HMMs. However, these methods are of comparatively high complexity and require detailed statistical modeling and parameter learning. They also do not model the motion dynamics in an explicit way.

It is known that appearance and dynamics are two important cues for human action recognition. However, the capture of both global temporal dynamics and local visual appearance features of action sequences has always been a challenging task, and few attempts have been made to explore it. Mikolajczyk and Uemura [15] extract both local appearance features (i.e. MSER, Harris-Laplace and Hessian-Laplace) and local motion features (i.e. dominant motion compensation) to jointly represent human actions. These appearance-motion features are clustered to form many vocabulary trees. Votes for action categories are given by matching features extracted from query frames to the trees. Artizzu *et al.* [16] combine weak trajectory features with popular spatio-temporal features to recognize long-time continuous social behaviors. Weak trajectory features, such as agent distance, movement direction, velocities and accelerations are computed based on the tracked agent positions. Finally, temporal context features are adopted to improve the classification performance. They conclude that the combination of appearance and motion features outperforms their use separately, and suggest that behavior understanding should be based on such heterogeneous features. Since these methods can be regarded as feature-level combinations of local appearance and motion features, Wang *et al.* [10] use a multi-channel approach to fuse different distance matrices computed with HOG (histograms of oriented gradients), HOF (histograms of optical flow) and MBH (motion boundary histograms). They calculate the χ^2 distance matrix separately for each channel, and then sum them up, weighted by their respective mean values, to form the final RBF-χ^2 kernel. This method can be regarded as a decision-level combination. Our approach is in part similar to that of Wang *et al.* [10], but different in two aspects. First, we compute the robust LDS over the whole action sequence to describe the global dynamics. In this way, we expect to capture the global temporal evolution of action sequences, even when this can not be achieved with local motion features. Fig. 1 gives an example of how a robust LDS captures the temporal dynamics of action sequences. With robust LDS, an action sequence is decomposed into subspace components and state sequence, in which the latter represents the temporal variations. Second, we combine the global dynamics and local appearance in a maximum margin distance learning framework. The combination weights are learned, to balance the importance of motion and appearance features, such that the action classes are maximally separated. In addition, we point out that our combination framework can be easily extended to three or more heterogeneous features without any modification.

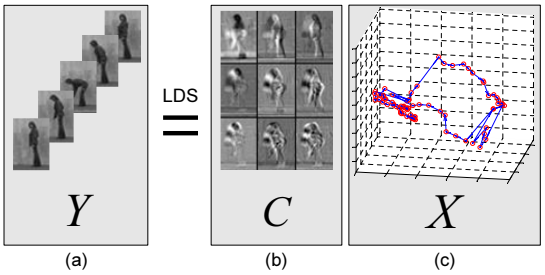

Fig. 1. This example illustrates the robust LDS representation of action sequences. (a) A video sequence Y is decomposed into two parts with robust LDS. (b) The subspace components C show the principal action appearance. (c) The state sequence X illustrates the motion dynamics over time.

1.1 Motivation and Overview

LDS and its extensions [17], [18], [19], [20] have long been studied for human motion analysis. They demonstrate superiority over common HMMs on classification task, but generally require complex Bayesian modeling and inference. Recent advances in system identification theory for learning LDS parameters [21], [22] and similarity measures between LDSs [23], [24], [25], [26] have made LDS successful for classification of high-dimensional time-series data in the field of dynamic texture [27], [28]. This poses us a new way to model and compare action sequence dynamics. By modeling the temporal variations with LDS, system theoretic methods specifically consider the global dynamics of action sequences. The similarity between two LDSs is directly measured with a distance or kernel metric defined on the LDS space. Considering that sequence similarity based on local appearance features is commonly computed with a distance metric defined on the histogram space, this inspires us to combine the global dynamics and local appearance features in a unified framework.

Following this idea, our study in this paper aims to combine global dynamics, which are described with robust LDS, and local appearance, which is represented by dense curved spatio-temporal cuboids [10], for human action recognition in a maximum margin distance learning framework. The main motivation behind this method is to model the distance between two sequences as a weighted sum of the respective distances obtained from global dynamics and local appearance features. Recognition is carried out by learning the weights in a way that maximizes the class discrimination. By this means, LDSs and cuboids are complementary in describing action sequences by capturing both motion dynamics and visual gradients. Fig. 2 shows the system diagram of our proposed method.

Furthermore, we find that stability is a crucial property for dynamical systems, whereas it is generally omitted by most of previous work [9], [26], [27], [29], though it has been intensively studied in the system identification literature [21], [30], [31]. This motivates us to introduce an efficient optimization method to learn robust LDS. We also develop a shift invariant distance metric based on the subspace angles distance [24]. We demonstrate that our distance metric is insensitive to the starting frame of action sequences, and thus outperforms the traditional metrics in recognition rate.

1.2 Contributions

The main contributions of this paper are summarized as follows:

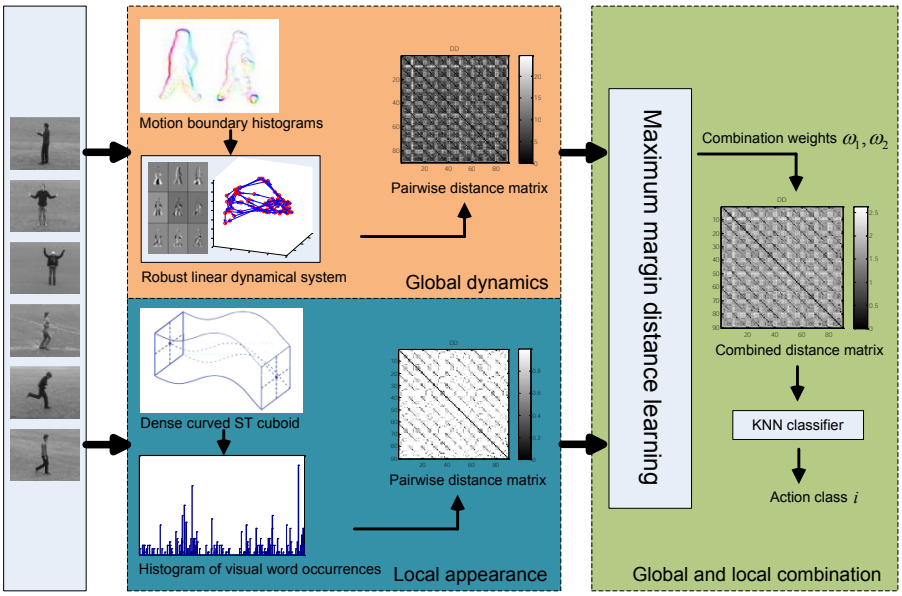

Fig. 2. System diagram of the proposed method. For global dynamics, MBH is extracted in each frame and robust LDS is learned based on MBH sequences. Distance matrix is computed using a shift invariant distance metric. For local appearance, HOG is calculated for each dense curved spatio-temporal cuboid. Occurrence histogram is built in the bag-of-words framework and Chi-Squared distance is computed to measure the pairwise distance. Finally, these two distance matrices are combined in a maximum margin distance learning framework. Action classification is achieved by using a KNN classifier.

- We propose to combine LDSs and cuboids for human action recognition in a maximum margin distance learning framework. The LDSs and cuboids capture not only local visual information of action appearance, but also global dynamic information of human movements. To the best of our knowledge, there is no other work which considers this approach to recognize human actions.
- We introduce a simple yet efficient robust LDS learning algorithm to describe action sequence dynamics. A suboptimal solution is achieved by iteratively checking stability criteria and generating new constraints. The process is repeated until it finds a stable solution. We show that sequences evolved with robust LDS share the same dynamic characteristics as the training data. This is essential for our proposed shift invariant distance metric.
- We develop a shift invariant distance metric which is robust to the initial conditions of the LDS, thus is insensitive to the starting frame of the action sequences. This is based on the considerations that the classification of an action sequence should be independent of the frame that it begins with. The state-of-the-art methods do not achieve this purpose.
- We carry out extensive experiments on eight public datasets. We evaluate the robust LDS on the selection of model parameters, and compare to the traditional LDS as well as three other temporal methods, namely MEMM [32], CRF [13] and switching LDS [19] to quantify the improvement in recognition rates. Furthermore, we evaluate the performance achieved with LDSs and cuboids alone, as well as the combination of these two components. We compare to current state-of-the-art results and show the significant potential of our method.

1.3 Organization

Section 2 reviews related work on appearance-based and motion-based methods for human action recognition. Section 3 firstly introduces LDS and describes how LDS parameters are learned in the literature. Then a robust LDS learning algorithm is introduced and analyzed in detail. Finally a shift invariant distance metric is proposed to measure the distance between LDSs. Section 4 briefly outlines the recognition pipeline of bag-of-words and shows how dense curved spatio-temporal cuboids are constructed and described in our work. Section 5 presents the maximum margin distance learning algorithm and illustrates how LDSs-based distance and cuboids-based distance are combined for human action recognition. Section 6 shows the experimental results including comparisons with competing methods for learning human actions. Section 7 gives the conclusions plus some ideas for further work.

2 Related Work

In Section 1, we briefly reviewed the work on motion representation in order to make clear the motivation for this paper. In this section, we review in detail the human action recognition methods in order to put our work into context. For clarity, we divide the existing work into two categories: appearance-based methods and motion-based methods.

2.1 Appearance-based Methods

Appearance-based methods usually extract local or global visual features from video data for human motion analysis. Local approaches involve detecting and describing spatio-temporal interest points to represent human activity. Laptev [3] extends the Harris detector to video in both spatial and temporal domains. However, this detector only finds a small number of stable interest points, which are usually insufficient for motion analysis and classification. Dollár *et al.* [4] improve the 3D Harris detector by applying Gabor filtering in the spatial and temporal dimensions separately. By choosing proper spatial and temporal scales, the detector can yield a large number of cuboids. Oikonomopoulos *et al.* [33] present a saliency detector based on a computed entropy of each cuboid. The cuboids with the local maximum entropy value are selected as salient points. Willems *et al.* [34] identify saliency with the determinant of a 3D Hessian matrix, which can be calculated efficiently using the integral videos. Seo and Milanfar [35] use space-time local steering kernels to capture the underlying geometric structures of action sequences. They show good performance with only one training example. Mathe and Sminchisescu [36] propose saliency map prediction models, which are learned from human eye fixated regions for action recognition. They illustrate comparable results to those obtained using dense sampling with sparse uniform sampling. These methods treat space and time in the same manner. However, action videos normally show different characteristics in space and time. It is therefore more suitable to handle space and time differently, rather than to treat them in a joint 3D space. Wong and Cipolla [37] detect interest points separately on subspace images and coefficient vectors by decomposing the sequence using the non-negative matrix factorization algorithm. Shabani *et al.* [38]

address this problem by designing a novel anisotropic filter in the temporal dimension and applying an isotropic Gaussian filter in the spatial domain. Wang *et al.* [10] handle space and time differently by tracking densely sampled interest points through video sequences. The resulting trajectories and the aligned space-time volumes are used to represent the videos. To describe spatio-temporal points, feature descriptors are calculated in the neighborhoods of selected points using image measurements such as spatial or spatio-temporal gradients. Schüldt *et al.* [39] calculate normalized derivatives in both space and time. Dollár *et al.* [4] experiment with image brightness, gradient and optical flow information. Scovanner *et al.* [5] extend the popular SIFT descriptor [40] to the space-time domain and develop a 3D SIFT. Willems *et al.* [34] generalize the SURF descriptor [41] by computing a weighted sum of space-time Haar-wavelets in grid cells. Wang *et al.* [10] combine local HOG, HOF and MBH descriptors to achieve state-of-the-art results. Local methods have shown many encouraging results because of their reliability in the presence of background noise, geometric variation and occlusion. However, there exists a limitation for these approaches in that they do not incorporate global information and ignore the temporal variations of human actions.

Global approaches generally use global features such as optical flow and silhouettes to capture the motion information in action sequences. Efros *et al.* [7] use blurred optical flow histograms to match actions in sport videos. Ali and Shah [42] derive kinematic features from optical flows. PCA is applied to determine the dominant kinematic modes. Action recognition is performed by using multiple instance learning approach. Wang *et al.* [10] introduce a descriptor by tracking dense points based on the optical flow field. Instead of extracting optical flow, the silhouette of a person in the image is used for recognition. Bobick and Davis [43] extract silhouettes and use temporal templates for the representation and recognition of aerobics actions. Wang and Suter [44] project silhouette sequences into a low-dimensional subspace to characterize the space-time properties of human actions. However, optical flow and silhouettes are very sensitive to background noise, viewpoint variation as well as occlusions. To address these issues, Tran *et al.* [8] divide the silhouettes and optical flow fields into rectangular grids. A motion descriptor is developed by combining both horizontal and vertical components of optical flow as well as the silhouette of person. Gorelick *et al.* [45] stack silhouettes over time to form a space-time volume. Local space-time saliency and orientation features are derived by solving a Poisson equation. Global approaches perform well because they encode much of the motion information. However, they depend strongly on the recording conditions, in order to realize accurate localization, background subtraction or tracking. In addition, these approaches also do not model the temporal dynamics of global features.

2.2 Motion-based Methods

Motion-based methods usually model the action sequences with temporal state-space methods such as HMMs, conditional random fields (CRFs) or dynamical systems. Brand *et al.* [12] propose a coupled HMM to represent the interaction between subjects. Caillette *et al.* [46] use a variable length Markov model (VLMM) to describe the observations and 3D poses for each action. Hongeng and Nevatia [47] incorporate domain knowledge as an a priori probability for state duration into the HMM framework,

resulting in a hidden semi-Markov model (semi-HMM) to achieve event detection. HMMs are efficient for modeling time series data. However, their application is restricted due to the assumptions of conditionally independent observations and Markov property for the sequence of hidden states. CRFs, on the other hand, avoid these two assumptions and allow non-local dependencies between states and observations. Sminchisescu *et al.* [13] use CRFs for human motion recognition. They show that CRFs outperform both HMMs and maximum entropy Markov model (MEMM) when a longer observation history is taken into account. Vail *et al.* [14] compare CRFs and HMMs in detail and conclude that CRFs perform as well as or better than HMMs. Natarajan and Nevatia [48] propose a two-layer model and use CRFs to encode actions and viewpoint-specific poses. However, despite that HMMs and CRFs model action sequence as a time-varying series, they do not model the motion dynamics in an explicit way.

Dynamical system methods capture the temporal variations by decoupling the action sequence into subspace poses and latent dynamics. Bregler [17] proposes a multi-level framework for learning and recognizing human dynamics. LDSs are used to describe the mid-level simple movements, while HMMs are learned to represent the high level complex behaviors. Blake *et al.* [18] represent the multi-class motion sequence with a mixed auto-regressive process. Model parameters are learned by combining the expectation maximization (EM) with the Condensation algorithm. Pavlovic and Rehg [19] model the nonlinear dynamics in human motion with switching LDSs, whereas model learning and inference are based on a variational technique. Turaga *et al.* [20] model an action sequence as a cascade of LDSs. They simultaneously segment the sequence in the temporal dimension and learn an LDS for each segment. Wang *et al.* [49] explore the nonlinearity of motion sequences with Gaussian process dynamical models. Model parameters are marginalized out in closed form rather than being estimated. This results in a nonparametric model for dynamical systems. Though dynamical system approaches are powerful for describing the motion sequence dynamics, they usually require detailed statistical modeling and parameter learning. In addition, exact inference is generally intractable and approximation methods have to be developed.

Recent work reported in the system identification literature makes the comparison between dynamical systems straightforward by directly defining distance or kernel metrics in the model space. Martin [23] defines a metric for stable ARMA models based on a comparison of their cepstrum coefficients. De Cock and De Moor [24] extend the concept and propose to compare stable ARMA models by using the subspace angles between two systems. Chan and Vasconcelos [25] derive a probabilistic kernel based on Kullback-Leibler divergence and use it for dynamic texture classification. Vishwanathan *et al.* [26] present a generic similarity metric for dynamic scene analysis based on the Binet-Cauchy theorem. Since most of the work is designed for dynamic textures, few attempts have been made for human motion recognition. Bissacco *et al.* [50] extend the work in [26] and define a novel kernel-based distance on LDSs for human gait recognition. Chaudhry *et al.* [9] encode each frame of an action sequence using a histogram of oriented optical flow (HOOF) and employ Binet-Cauchy kernels [26] to describe the HOOF sequence. In this paper, we argue that previous work for learning dynamical systems mostly ignores the model stability, and the similarity metrics are sensitive

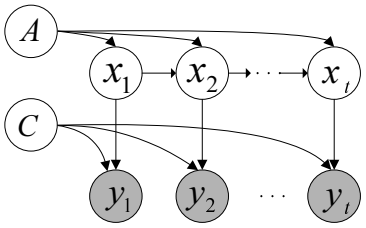

Fig. 3. Graphical model representation of LDS. The grey element y_i designates the observable feature or image. The white element x_i represents the unseen or latent state variable. The parameter A indicates the dynamic matrix, while parameter C indicates the subspace mapping matrix.

to the initial conditions. To address these issues, we introduce a robust LDS learning algorithm and develop a shift invariant distance metric. We demonstrate encouraging recognition results on both short clips datasets and long continuous datasets.

3 Recognition with Robust LDSs

Dynamical system methods have been studied extensively in fields ranging from control engineering to visual process. Among all the methods, LDS is used broadly because of its simplicity and efficiency. For instance, many dynamic texture recognition methods represent the texture's temporal variation as an LDS [27], [28]. From the graphical model's perspective, as illustrated in Fig. 3, LDS is indeed a generative state-space model with Gaussian observations and Markov states. For human action sequences, many inherent nonlinearities such as phase transition, turbulence and delay can be eliminated by choosing proper coordinates or mapping into high-dimensional spaces [50]. Therefore in this paper, we focus on how to model the human motion dynamics with LDS.

3.1 Linear Dynamical Systems

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 parameter tuple $\mathbf{M} = (A, C)$ 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 distributions.

In (1) the hidden state is modeled as a first-order Gauss-Markov process, where x_{t+1} is determined by the previous state x_t. The output y_t depends linearly on the current state x_t. 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 it is generally solved by using the least squares estimation [27].

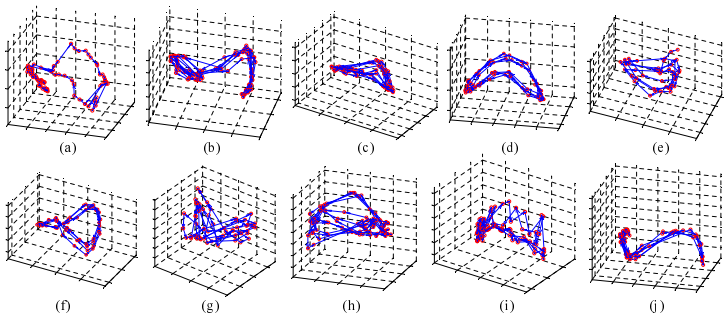

Fig. 4. State sequences ($n = 3$) of 10 different actions performed by one subject from the Weizmann dataset. (a) bend, (b) jack, (c) jump, (d) pjump, (e) run, (f) side, (g) skip, (h) walk, (i) wave1, (j) wave2.

Let column matrix $Y_{1:\tau} = [y_1, y_2, ..., y_\tau]$ and $X_{1:\tau} = [x_1, x_2, ..., x_\tau]$ represent the observation sequence and state sequence, respectively. In order to get a closed form estimate of the model parameter $\mathbf{M}$, the observation matrix is firstly decomposed with the singular value decomposition (SVD), i.e. $Y_{1:\tau} = U\Sigma V^T$, where U, V are orthogonal and Σ is a rectangular diagonal matrix with non-negative real numbers on the leading diagonal. An estimate of the subspace mapping matrix and the underlying state sequence is obtained by setting

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

The model dimension n is determined by retaining the singular values which exceed a given threshold.

Then the least squares estimation of A is

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

where $\|\cdot\|_F$ denotes the Frobenius norm and $^+$ denotes the Moore-Penrose inverse. Given the above estimates of $\hat{A}$ and $\hat{C}$, the covariance matrices $\hat{Q}$ and $\hat{R}$ can be estimated directly from residuals.

According to Eq. (2), LDS implicitly models the observation $Y_{1:\tau}$ with a subspace mapping matrix C and its corresponding coefficients $X_{1:\tau}$. In human action recognition task, as shown in Fig. 1, the subspace matrix C describes the action components, while matrix A, which is derived from $X_{1:\tau}$, 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 the embedding components of an action sequence, and is thus very different from the local spatio-temporal gradient descriptors. In Fig. 4, we show the state trajectories of ten action classes from the Weizmann dataset, from which we can see that the trajectories behave very differently among different actions. This makes it possible to classify and recognize them by exploring the dynamics.

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 is represented using the tuple $\mathbf{M} = (A, C)$, which lives in a non-Euclidean space and is in non-vector form. There is no straightforward way to compute the non-vector descriptor distance in a non-Euclidean space, let alone find distance metrics which are invariant to initial conditions.

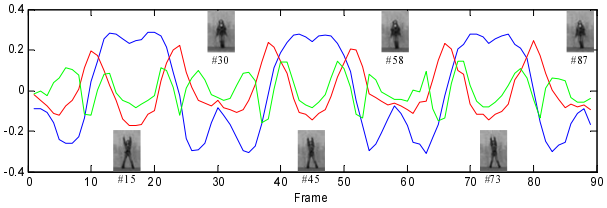

Fig. 5. State sequence $X_{1:\tau}$ learned with a jumping jack video from the Weizmann dataset. The sequence has total 89 frames with three cycles. A robust LDS is learned with the state dimension $n = 3$. The three state components are figured as blue, red and green curves, respectively. The sample images corresponding to the key periodic frames are also demonstrated.

3.2 Learning Robust LDS

Stability is a very important property for LDSs and it has been intensively studied in the system identification literature. An unstable LDS may fail to generate long sequences which share the same characteristics as the original data. However, we find that most of the previous work [9], [26], [27], [29] has omitted this problem. This may cause the failure of distance computations between descriptors because the distance or kernel metrics for dynamical systems are mostly defined on stable models [23], [24], [26]. We show in Section 3.3 that generating consistent data is critical for our proposed shift invariant distance metric.

Let $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ denote the eigenvalues of dynamic matrix A in decreasing order of magnitude. The system spectral radius is defined as $\rho(A) \equiv |\lambda_1|$. An LDS is called stable if and only if $\rho(A) \leq 1$. If $\rho(A) = 1$, the system is said 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. 5, we show an example of a periodic action sequence from the Weizmann dataset. By learning the robust LDS, we obtain a state sequence which varies periodically with the observation sequence. This means that robust LDS captures the intrinsic dynamics of cycle movements. However, traditional LDS learning methods such as (3) do not enforce this stability criterion. This may cause the solution to be unstable even if the true system is stable.

Some robust techniques in the literature have been developed based on searching through a full feasible set of dynamic matrices [31] or on improving the least squares objective function with regularization terms [30]. The N4SID method [21] achieves optimal solutions in the sense of maximum likelihood. However, it is computationally expensive to obtain the optimal solutions for these methods. In our case of human action recognition, we do not require precisely optimal results. We prefer an efficient approximation solution, provided it satisfies the stability criterion. Here we introduce a constraint generation method [51], which achieves a stable result efficiently by iteratively checking the stability criterion and generating new constraints. The main idea of this approach lies in formulating the least squares problem in (3) as a quadratic program with linear constraints. These linear constraints, which are initially void, are inferred iteratively using the stability criterion. The optimal process stops once it finds a stable solution.

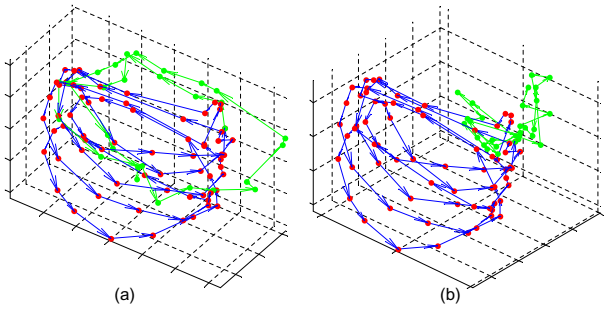

Fig. 6. Generation comparison between robust LDS and traditional LDS. The red trajectories in (a) and (b) illustrate the state sequences learned with robust LDS (a), and traditional LDS (b) respectively, using the same walking video in each case. The green trajectories are generated by evolving the red sequences 28 steps ahead with their respective LDS, from which we see that the unstable LDS (b) cannot generate new data consistent with the given data.

The least squares problem in (3) can be reformulated by expanding the right hand side as

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

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

The linear constraints are obtained by decomposing $\hat{A}$ using SVD $\hat{\Sigma} = \hat{U}^T\hat{A}\hat{V}$ and inferring as

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

where $g = vec(\tilde{u}_1\tilde{v}_1^T)$, $\hat{a} = vec(\hat{A})$, $\tilde{u}_1$ and $\tilde{v}_1$ are the singular vectors corresponding to $\tilde{\lambda}_1$.

Therefore, the quadratic program can be written as

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

This quadratic program is repeatedly invoked until the stability criterion is satisfied. At each iteration, a new linear constraint is calculated with (5). Considering that g depends on a, it is therefore an approximation solution to hold g fixed. In Fig. 6, we compare the generation capability between robust LDS and traditional LDS, from which we can see that the robust LDS generates new data illustrating the same dynamic characteristics as the training sequence, while the traditional one generates cluttered data which are very different from the original data.

3.3 Shift Invariant Distance Metric for LDS

Given an action sequence, we use the robust LDS model parameter $\mathbf{M} = (A, C)$ as a motion sequence descriptor, with the dynamic matrix $A \in \mathbb{GL}(n)$, where $\mathbb{GL}(n)$ is the group of all $n \times n$ invertible matrices, and with the mapping matrix $C \in \mathbb{ST}(p, n)$, where $\mathbb{ST}(p, n)$ is the Stiefel manifold. Since the model space has a non-Euclidean structure and the descriptor is in non-vector form, this naturally raises the issue of how to measure the similarity between two descriptors. In one of the first papers to address this issue, Martin [23] defines a metric for stable ARMA models based on a comparison of

their cepstrum coefficients. De Cock and De Moor [24] improve Martin's work by using the subspace angles between two LDSs. The subspace angles are defined as the principal angles between the column spaces of infinite observability matrices $\mathcal{O}_\infty(\mathbf{M}_i) = [C_i^T \quad (C_iA_i)^T \quad (C_iA_i^2)^T \quad \ldots]^T \in \mathbb{R}^{\infty\times n}$ for $i = 1, 2$.

Let $\mathbf{M}_1 = (A_1, C_1)$ and $\mathbf{M}_2 = (A_2, C_2)$ denote the two motion sequence descriptors. The computation of subspace angles is obtained by solving the Lyapunov equation

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

for $\mathcal{Q}$, where $\mathcal{Q} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$, $\mathcal{A} = \begin{pmatrix} A_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$, $\mathcal{C} = (C_1 \quad C_2) \in \mathbb{R}^{p\times 2n}$.

The solution of (7) is guaranteed to exist when $\mathbf{M}_1$ and $\mathbf{M}_2$ are stable. The cosines of the subspace angles $\cos^2\theta_i$ are calculated as eigenvalues of matrix $Q_{11}^{-1}Q_{12}Q_{22}^{-1}Q_{21}$, where $Q_{kl} = \mathcal{O}_\infty(\mathbf{M}_k)^T\mathcal{O}_\infty(\mathbf{M}_l)$ for $k, l = 1, 2$.

The subspace angles distance is defined as

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

Vishwanathan *et al.* [26] present a generic kernel metric for dynamical systems based on the Binet-Cauchy Theorem. For two LDSs $\mathbf{M}_1$ and $\mathbf{M}_2$ which evolve with independent noise realization, the Binet-Cauchy kernel is computed by firstly solving a Sylvester equation

$$S = e^{-\lambda}A_1^TSA_2 + C_1^TC_2 \tag{9}$$

for S, where $e^{-\lambda}$ is an exponential discounting term with $\lambda > 0$.

The above Sylvester equation can be solved efficiently by rewriting it as

$$\begin{aligned} vec(S) &= vec(e^{-\lambda}A_1^TSA_2) + vec(C_1^TC_2) \\ &= (A_2^T \otimes e^{-\lambda}A_1^T)vec(S) + vec(C_1^TC_2) \end{aligned} \tag{10}$$

Thus we have

$$vec(S) = (\mathbf{I} - A_2^T \otimes e^{-\lambda}A_1^T)^{-1}vec(C_1^TC_2) \tag{11}$$

After reshaping $vec(S)$ into matrix S, the Binet-Cauchy kernel is defined as

$$k_{LDS}(\mathbf{M}_1, \mathbf{M}_2) = x_1(0)^TSx_2(0) \tag{12}$$

on condition that $e^{-\lambda}\|A_1\|\,\|A_2\| < 1$. Here $x_1(0)$, $x_2(0)$ are the initial states of $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively.

From (12), we see that the Binet-Cauchy kernel depends directly on the initial conditions. Experiments also show that subspace angles distances and Binet-Cauchy kernels vary greatly when two sequences differ by only a temporal shift (see Fig. 7, blue square). In our case of human action recognition, we require a distance 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 a number of steps such that the distance between them is minimized. That is

$$d(\mathbf{M}_1, \mathbf{M}_2) = \min_{\tau_1,\tau_2\in\mathbb{N}} d_{LDS}(\mathbf{M}_1(\tau_1), \mathbf{M}_2(\tau_2)) \tag{13}$$

or

$$k(\mathbf{M}_1, \mathbf{M}_2) = \max_{\tau_1, \tau_2 \in \mathbb{N}} k_{LDS}(\mathbf{M}_1(\tau_1), \mathbf{M}_2(\tau_2)) \tag{14}$$

where $\mathbf{M}(\tau)$ denotes the model parameter of evolved sequence which is generated by shifting the original sequence τ steps ahead. We notice that a robust LDS model must be learned to ensure that the evolved sequences have the same characteristics as the original sequence. This is why model stability is especially important for our algorithm.

It is unfortunate that there is no an explicit way to obtain the solutions of (13) and (14). However in many applications, the periods of most motion patterns are short. Thus we can solve this problem by searching through all the combinations of τ_1 and τ_2 exhaustively. If $\mathcal{T}$ is an upper bound for the shift, then the complexity of this problem is $O(\mathcal{T}^2)$. In practice, we can speed up the computation by caching both the original $\mathbf{M}$ and its $\mathcal{T}$ shifted versions $\mathbf{M}(\tau)$ for $\tau = 1, ..., \mathcal{T}$. The detail of the shift invariant distance metric algorithm is presented in Algorithm 1.

Algorithm 1 Shift Invariant Distance Metric Algorithm

Input: Given two sequences Y_1 and Y_2;

Learn robust LDS $\mathbf{M}_1 = (A_1, C_1)$ and $\mathbf{M}_2 = (A_2, C_2)$ according to (6);
for τ=1:$\mathcal{T}$ **do**
  % *Subroutine to evolve Y_1 and Y_2 for τ steps respectively*
  Initialize x_0 as the last state of the learned state sequence;
  for $i = 0 : \tau - 1$ **do**
    $x_{i+1} = A * x_i + mvnrnd(\mathbf{0}, Q)$
    $y_{i+1} = C * x_{i+1} + mvnrnd(\mathbf{0}, R)$
  end for
  Learn robust LDS $\mathbf{M}_1(\tau)$ and $\mathbf{M}_2(\tau)$ based on evolved sequences $Y_1(\tau)$ and $Y_2(\tau)$;
end for
Compute all pairwise distance metrics between $\mathbf{M}_1(\tau_1)$ and $\mathbf{M}_2(\tau_2)$ according to (8) or (12);
Compute the shift invariant distance metric based on (13) or (14);

Output: Shift invariant distance metric $d(\mathbf{M}_1, \mathbf{M}_2)$ or $k(\mathbf{M}_1, \mathbf{M}_2)$;

We show in Fig. 7 that our shift invariant distance metric outperforms the standard metrics in two aspects. First, the aligned distance/kernel exhibits a better similarity measure than the original subspace angles distance and Binet-Cauchy kernel. This results in better recognition results as illustrated in Fig. 9a. Second, the aligned distance/kernel is more stable with respect to frame offset, whereas the subspace angles distance shows sudden changes in some conditions and the Binet-Cauchy kernel exhibits periodic variations with the periodic change of the starting frame. In addition, in Fig. 7a, we compare the aligned distance and subspace angles distance between different action classes, i.e. walk vs. skip. We see that although the aligned distances between different classes are reduced (red triangle vs. blue triangle), the separation between the intra-class and inter-class distances (red square vs. red triangle) is very clear. On the other hand, this separation is not so clear for the original subspace angles distances (blue square vs. blue triangle). Specifically, we choose the maximum walk vs. walk subspace angles distance as the baseline, and show that the walk vs. skip subspace angles distances

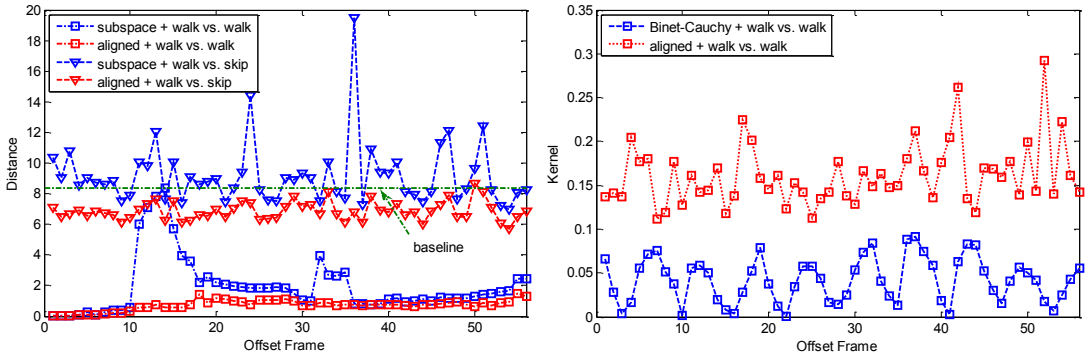

Fig. 7. Comparison of subspace versus aligned distances (a), and Binet-Cauchy versus aligned kernels (b). The 'square' distances/kernels are computed between a model learned from a walking sequence and the set of models learned from the same sequences with different starting frames. The 'triangle' distances are computed in the same way between the walking sequence and shifted skipping sequences. The baseline is chosen as the maximum walk vs. walk subspace angles distance. The 'blue' distances between different classes, say walk vs. skip, which are below the baseline may cause misclassification.

which are below the baseline may cause misclassification.

4 Recognition with Curved Spatio-Temporal Cuboids

Local spatio-temporal appearance features are successful for many visual tasks. They capture characteristic shape, texture and local motion in video sequences. These characteristics are good complements to the global dynamics. In Section 2, we reviewed the state-of-the-art local approaches. In this paper, we use the dense curved spatio-temporal cuboids [10] as local appearance features and describe them with HOG descriptors. This is based on the consideration that the dense sampling method has proved its effectiveness and achieves state-of-the-art recognition results on a wide range of human action datasets.

Given an action sequence, we first densely sample interest points on a grid in each frame. Then these points are tracked based on the dense optical flow field to form dense trajectories. The curved spatio-temporal cuboids are constructed as the local space-time volumes along the trajectories. The size of the cuboid is $N * N$ pixels and L frames long. Since the trajectories are curves in space-time domain, we call these space-time volumes 'curved cuboids', to differentiate from the traditional 'straight cuboids' [4].

To describe the curved spatio-temporal cuboids, we subdivide each cuboid into a spatio-temporal grid of size $N_\sigma * N_\sigma * N_\tau$. We compute HOG in each cell of the grid to create a sub-histogram, and concatenate all the sub-histograms to obtain the final descriptor.

Once we obtain all the HOG descriptors on the dataset, we quantize them into visual words by k-means clustering. Then we represent each sequence as the frequency histogram over the visual words. Let H_1 and H_2 denote the histograms of two sequences. They are compared using the Chi-Squared

($\mathcal{X}^2$) distance

$$d_{STC}(H_1, H_2) = \frac{1}{2}\sum_i \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)} \tag{15}$$

5 Combining LDSs and Cuboids by Maximum Margin Distance Learning

In this section, we combine LDSs-based distance d_{LDS} and cuboids-based distance d_{STC} in a maximum margin learning framework. The final distance between two sequences is computed as a linear combination of d_{LDS} and d_{STC}. The weights for the linear combination are learned with the maximum margin distance learning algorithm [29] in such a way that the action classes are maximally separated.

Let a set of training sequences be given with class labels in $c = \{1, ..., N\}$ and let l_i be the label of sequence i. The global dynamic distance from sequence j to i is denoted as $d_{LDS}(ij)$, and the local appearance distance is $d_{STC}(ij)$. Then the combined distance from sequence j to i is defined as

$$D(ij) = \omega_1(l_i)d_{LDS}(ij) + \omega_2(l_i)d_{STC}(ij) = \boldsymbol{\omega}(l_i)^T \boldsymbol{d}(ij) \tag{16}$$

where $\boldsymbol{\omega}(l_i)$ is the vector of class dependent weights that characterize the corresponding distance for class l_i. Similarly, we define the class independent distance by dropping the class label l_i from (16).

In order to maximally separate the classes of sequences, we define a representative set $\mathcal{R}(i)$ of sequence i as its k-nearest neighbors within its class, and a comparative set $\mathcal{C}(i)$ as all other sequences outside its class. We assume that a sequence is closer to the sequences in its representative set than to those in its comparative set. That is, for all $i \neq j$, $j \in \mathcal{R}(i)$ and $k \in \mathcal{C}(i)$, we have the following distance constraints

$$D(ij) \leq D(ik) \Rightarrow \boldsymbol{\omega}(l_i)^T \triangle \boldsymbol{d}(i) \geq 0 \tag{17}$$

where $\triangle \boldsymbol{d}(i) = \boldsymbol{d}(ik) - \boldsymbol{d}(ij)$ is the distance difference vector from the sequence in the comparative set to the sequence in the representative set.

In the class dependent case, the total number of such constraints in class c is $L = \sum_{l_i=c} |\mathcal{R}(i)||\mathcal{C}(i)|$. By embedding these constraints in a maximum margin framework, $\boldsymbol{\omega}(c)$ can be found by minimizing

$$\min_{\boldsymbol{\omega}(c)} \frac{\lambda}{2}\|\boldsymbol{\omega}(c)\|^2 + \frac{1}{L}\sum_{i=1}^{L} \max(0, 1 - \boldsymbol{\omega}(c)^T \triangle \boldsymbol{d}(i)) \tag{18}$$

where λ is the regularization parameter. This problem is similar to the problem of learning a support vector machine (SVM). It can be efficiently solved with the Pegasos algorithm [52].

After obtaining $\boldsymbol{\omega}(c)$ for each class, a test sequence is classified as the class which satisfies the most number of distance constraints generated by the test sequence.

6 Experiments

To evaluate the performance of our proposed method for human action recognition, we carry out detailed experiments on the state-of-the-art datasets. We first briefly introduce the experimental setup, and then describe the experiments and results.

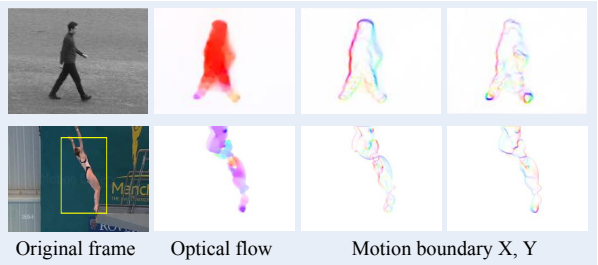

Fig. 8. Illustration of raw, optical flow and MBH (x, y) images of two action sequences from the KTH dataset (top) and the UCF sports dataset (bottom), respectively. For the optical flow and MBH images, gradient/flow orientation is indicated by color (hue) and magnitude by saturation.

6.1 Experimental Setup

6.1.1 Feature Extraction

To compute robust LDS, we extract sequential features from all the videos. Our proposed method can be used with different types of features, including even raw pixels, provided that the features form a time series. Silhouettes or shape features [44] are useful, but they are difficult to obtain in unconstrained environments. In this paper, we use the motion boundary histograms [53] to characterize the action profile. The MBH encodes the relative motion between pixels by computing gradients of the x and y optical flow components separately. It suppresses most of the camera motion and background texture, and thus highlights the foreground subject. Some examples are illustrated in Fig. 8.

As suggested in [53], we resize the sequences into 64×128 pixels. The MBH is computed by quantizing the orientations into 9 bins with 2×2 blocks of 8×8 pixel cells. To improve the performance, block overlap (0.5) is also incorporated. Thus we obtain a total of 7×15 blocks, where each block is described by a 4×9 histogram. The final histogram size is 3780 for both x and y components of MBH (i.e., $7 \times 15 \times 36$).

For local appearance, we take the same parameters as used in [10]. The size of the curved cuboid is 32×32 pixels and 15 frames long. Each cuboid is divided into $2 \times 2 \times 3$ cells, and each cell is described with a 8-bin histogram. Thus the final HOG descriptor size for each curved cuboid is 96. We randomly choose $100,000$ training cuboids to form a codebook of $4,000$ visual words with k-means. The resulting frequency histograms over the visual words are used as sequences representation.

6.1.2 Robust LDS

There are two parameters related with robust LDS: model dimension n and upper bound of temporal shift $\mathcal{T}$. We evaluate the performance for different values of these parameters in Section 6.3.1 and 6.3.2. We choose $n = 3$ and $\mathcal{T} = 16$ as default parameters in our experiments by considering the balance of recognition performance and computational cost.

6.1.3 Baseline Temporal Models

To quantify the improvement obtained with the robust LDS, we compare to the traditional LDS as well as three other temporal methods, namely MEMM [32], CRF [13] and switching LDS [19]. Details are

presented in the following. For all these models, we use the same MBH sequences as input features.

We implement the traditional LDS via least squares estimation according to Eq. (2) and (3). We use the same setup as for the robust LDS, except that the stability criterion is not incorporated. We compute LDS with model dimension $n = 3$ and use subspace angles distance to define the pairwise distance.

We learn linear models of MEMM and CRF, where every action class has a corresponding label. For CRF, we fix the context window size to 3. At testing, marginal probabilities are computed for each label and each frame using belief propagation. The frame label is selected as the label with the highest marginal probability. The sequence label is selected as the majority among all its frame labels.

We train three-state first order switching LDS models for each action class, where the underlying LDS dimension is set to $n = 3$. Testing sequence is classified into an action category by means of MAP estimation.

6.1.4 Maximum Margin Distance Learning

The regularization parameter of the combined distances is set to $\lambda = 0.05$ empirically. Both class dependent and independent weights are learned. Component distance matrices are normalized by their respective $\mu + 3\sigma$ as did in [29], where μ is the mean and σ the standard deviation.

6.1.5 Baseline Combination Scheme

We fix combination weights $\boldsymbol{\omega} = [1, 1]$ as a baseline fusing strategy to evaluate the contributions of maximum margin distance learning framework. Component distance matrices are normalized by their corresponding mean values. This is in fact the same scheme used in [10].

6.2 Datasets

Our experiments are conducted on five short clips datasets, namely Weizmann, KTH, UCF sports, Hollywood2 and UCF50, and three long continuous datasets, namely VIRAT, ADL and CRIM13.

The **Weizmann** dataset [45] consists of 93 video sequences from nine different people, each performing ten natural actions. These actions are either periodic or non-periodic, include bending, jumping jack, jumping-forward, jumping-in-place, running, galloping-sideways, skipping, walking, waving-one-hand, and waving-two-hands. Following the original experimental setup, we compute the recognition accuracy using the leave-one-out cross-validation method. We report the average accuracy over all classes as recognition performance.

The **KTH** dataset [39] consists of six human action classes: boxing, hand clapping, hand waving, jogging, running and walking. Each action class is performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In total, the data consists of 2391 video samples. We follow the original experimental setup by dividing the sequences into a test set (subject 2, 3, 5, 6, 7, 8, 9, 10, and 22) and a training set (the remaining subjects). Average results over all classes are reported as performance measures.

The **UCF sports** dataset [54] contains 150 video sequences of ten human actions collected from various sports: diving, golf swinging, kicking, lifting, horse-riding, running, skating, swinging on

bench, swinging on high bar, and walking. These action sequences are challenging because they are recorded in unconstrained environments and show large intra-class variabilities. Similar to the Weizmann dataset, we use the leave-one-out cross-validation method and report average accuracy over all classes as performance.

The **Hollywood2** dataset [55] collects 1707 video clips from 69 Hollywood movies which are classified into 12 action classes: answering the phone, driving car, eating food, fighting person, getting out of car, hand shaking, hugging person, kissing, running, sitting down, sitting up, and standing up. The sequences are divided into a training set (823 sequences) and a test set (884 sequences). Training and test sequences come from different movies, to ensure that the background and subject matter both change. This dataset is very challenging because it provides many different realistic scenarios for human action recognition. We follow the the original experimental setup and evaluate the performance by computing the average precision (AP) for each action class and reporting the mean AP over all classes (mAP).

The **UCF50** dataset [56] contains 6618 video clips, 50 action categories collected from the Youtube website. The actions range from general sports to daily life exercises. Considering that previous datasets are mostly staged by actors, UCF50 directly takes realistic videos uploaded by the users on the Youtube. This poses challenges because of the large variations in camera motion, object appearance, scale, viewpoint, background, and illumination conditions. All the sequences are split into 25 groups, such that each group consists of at least four clips. The clips in the same group share similar background and subject because they are obtained from the same long video. Thus we evaluate the performance by using the leave-one-group-out cross-validation method and report average accuracy over all classes as suggested by the authors [56].

The **VIRAT** dataset [57] is a still developing large-scale surveillance video dataset collected from both stationary ground cameras and moving aerial vehicles. It consists of 23 activity types occurring naturally in realistic outdoor scenes throughout 29 hours of video. It involves single person activities, as well as interactions between multiple persons, vehicles and facilities. The data is fully annotated with bounding boxes for both moving object tracks and localized spatio-temporal activities. This dataset is more challenging than the above datasets in terms of its resolution, background clutter, diversity in scenes and viewpoints. In order for comparison, we evaluate the performance on two subsets as used in [58] and [59]. We use the leave-one-scene-out cross-validation method and report average accuracy for the pre-segmented setting and mAP for the continuous setting.

The **ADL** dataset [60] consists of 1 million frames of 10 hours of video collected from 20 people performing 18 types of unscripted, natural everyday activities in diverse indoor environments. These naturally occurring activities are often related to hygiene or food preparation of daily living. The dataset provides detailed annotations which include activity labels, object tracks, hand positions and interaction events. The ADL is challenging because of its long-scale temporal structure and complex object interactions. We follow the original experimental setup and evaluate the performance by using the leave-one-out cross-validation method. We report average accuracy over all classes as performance.

The **CRIM13** dataset [16] consists of 237 annotated videos of pairs of mice engaging in social

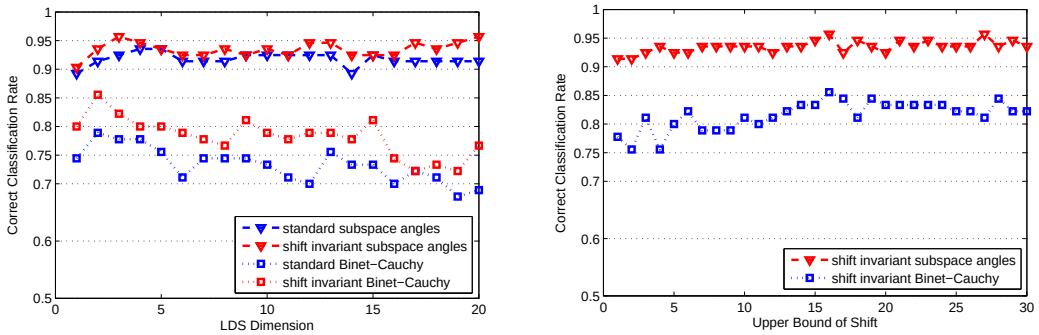

Fig. 9. Evaluation of robust LDS dimension n (a), and upper bound of temporal shift $\mathcal{T}$ (b). The performance of shift invariant and standard distance metrics is illustrated for comparison.

behaviors, which are categorized into 13 different actions. Each scene is recorded both from top-view and side-view using two fixed, synchronized cameras. Each video lasts about 10 minutes, for a total of over 88 hours of video and 8 million frames. We follow the original experimental setup by using 104 videos for training and 133 for test. We report average accuracy over all classes as performance.

6.3 Experiments on Short Clips Datasets

6.3.1 Evaluation of Robust LDS Dimension

In Fig. 9a, we examine the relationship between the correct classification rate (CCR) and the model dimension n up to 20 using only global dynamics on the Weizmann dataset. We also evaluate the performance of our proposed shift invariant distance metric compared with the standard subspace angles distance and the Binet-Cauchy kernel. The kernel distance is computed as

$$d(\mathbf{M}_1, \mathbf{M}_2) = k(\mathbf{M}_1, \mathbf{M}_1) + k(\mathbf{M}_2, \mathbf{M}_2) - 2k(\mathbf{M}_1, \mathbf{M}_2)$$

We see that: 1) the recognition rate does not improve much with the increase of model dimension n for the shift invariant and standard subspace angles distance (see Fig. 9a 'triangle' data). This means that we can choose a comparatively small model dimension, say $n = 3$, to gain a large saving of computation expense with only a small reduction in the recognition rate; 2) the recognition rate even drops with the increase of model dimension n for the shift invariant and standard Binet-Cauchy kernel (see Fig. 9a 'square' data). This indicates that the Binet-Cauchy kernel is not suitable for high-dimensional linear dynamical models; 3) the proposed shift invariant distance/kernel always performs better than the standard subspace angles distance and the Binet-Cauchy kernel (see Fig. 9a, red vs. blue). The shift invariant/standard subspace angles distance performs better than the shift invariant/standard Binet-Cauchy kernel (see Fig. 9a, triangle vs. square); 4) the proposed shift invariant subspace angles distance outperforms the other three distance metrics and achieves a best recognition result of 95.70% with $n = 3$. So in the following experiments, we use the shift invariant subspace angles distance as default metric for LDS unless otherwise specified.

6.3.2 Evaluation of Upper Bound of Shift

In Fig. 9b, we evaluate the recognition performance with respect to upper bound of temporal shift $\mathcal{T}$ from $\mathcal{T} = 1$ to $\mathcal{T} = 30$. We see that increasing $\mathcal{T}$ improves the performance up to $\mathcal{T} = 16$. Further

TABLE 1
Comparison of different temporal models (%)

	Weizm.	KTH	U. sports	Hollywood2	UCF50
Rb-LDS	**95.70**	**91.67**	**81.33**	52.72	**77.82**
T-LDS	91.40	87.47	76.81	46.28	68.15
MEMM	89.25	81.36	74.84	43.54	55.73
CRF	92.47	89.50	80.98	48.58	74.34
SLDS	93.55	89.55	81.30	**53.66**	76.56

increasing of temporal shift does not yield better results. In addition, with the increase in $\mathcal{T}$, the computational cost increases dramatically. Considering that most of the concerned sequences are non-periodic or have only short periods, we choose $\mathcal{T} = 16$ in all the experiments.

6.3.3 Comparison of Different Temporal Models

The different temporal models are compared in Table 1. We see that robust LDS (Rb-LDS) consistently outperforms the other models on all the datasets except switching LDS (SLDS) on the Hollywood2 dataset.

Rb-LDS gives an average 3%-9% improvement over the traditional LDS (T-LDS). This indicates that the stability criterion of dynamical systems plays an important role in the results. For T-LDS, the computation of the shift invariant distance metric may fail if the LDS is degenerate. This is especially clear on the realistic datasets, e.g., UCF sports, Hollywood2 and UCF50, because they contain many more disturbance factors.

Notwithstanding all these disadvantages, T-LDS outperforms MEMM by nearly 2%-12% because it captures the global dynamics. CRF gives much better results than both T-LDS and MEMM. This is probably because CRF considers the temporal context. It is interesting to notice that the temporal context is especially useful to discriminate actions with a similar motion style, e.g., walk vs. jog in the KTH dataset. However, training CRF with a long window of observations is much more time consuming than for T-LDS and MEMM.

SLDS achieves a recognition performance almost as good as Rb-LDS. These two methods both use LDS to describe the sequence dynamics. SLDS further builds a stochastic model on top of a set of LDSs. The switching or transition among the set of LDSs describes the nonlinear dynamics. This is why SLDS gives a slightly better result on the more complex dataset, e.g., Hollywood2, in which many clips include multiple primitive actions. However, Rb-LDS consistently performs better than SLDS on the other datasets, because Rb-LDS specifically considers the stability of the LDS and temporal shifts of the action sequence. In addition, learning SLDS involves many more parameters and exact inference is generally not possible. On the other hand, training Rb-LDS is simple and efficient. Classification is achieved by directly defining a distance metric on the LDS space. This makes it straightforward to compare two action sequences. More importantly, Rb-LDS yields a pairwise distance matrix, which can be embedded into the maximum margin distance learning framework. This embedding cannot be achieved with SLDS.

A closely related LDS-based activity clustering method is proposed by Turaga *et al.* [20], who consider the same tools as ours and SLDS. They segment a long video sequence into meaningful subsequences by searching for action boundaries with a suboptimal algorithm. Then they fit LDSs to each segment and cluster them by building a distance metric on LDSs. Each clustering center represents an action component, and each LDS is assigned the label to its nearest center. The sequence is finally modeled as a cascade of LDSs (CLDS). CLDS differs with SLDS in two ways. First, CLDS assumes as given the number of clustering centers to model action components, while SLDS assumes the number of hidden states. Second, CLDS maintains the temporal structure of LDSs using an n-grams model, while SLDS maintains the temporal structure with a state transition matrix. The similarity between Rb-LDS and CLDS is that they both use LDS to describe motion dynamics and define a distance metric on LDSs. However, CLDS differs with Rb-LDS in that it is designed to mine repetitive patterns from long sequences which may contain multiple activities. To define and search for a stable segmentation boundary in long sequences is indeed not so straightforward. The success of the search largely depends on the definition of action components and the expected segmentation granularity. The proposed reconstruction error based algorithm [20] indeed produces many uncertainties in the segmentation. These uncertainties can be eliminated or alleviated by our proposed aligned distance metric.

6.3.4 Recognition Performance of Proposed Framework

In Fig. 10, we compare confusion matrices with respect to different distance measures: LDSs, cuboids and combined distances with both class independent and class dependent weights on three action datasets, namely Weizmann, KTH and UCF sports. For Hollywood2, binary classification is carried out for each action class as all the clips are labeled with 1 or -1, indicating whether the clip contains a specific action class or not. For this reason, we only conduct class independent combination for each action class separately and show the results in Table 2, accompanied with the learned class independent weights for each class. The class dependent combination is not applicable for this dataset. For UCF50, since it contains 50 action classes, we only show the average results in Table 3 due to limited space. We also give the class independent weights in Table 4 and class dependent weights in Fig. 11 on all the datasets.

We see that LDSs give fairly good results by themselves, and outperform cuboids on all the datasets except the UCF sports. It is worth noting that the recognition rates achieved by LDSs are nearly 10% better than those obtained by cuboids on Hollywood2 and UCF50. This indicates that LDSs are more discriminative than cuboids on complex datasets which contain large temporal variations. The UCF sports dataset involves many specific environments and items of equipment. Local cuboids are designed to capture this information, and thus achieve better results on this dataset.

We observe that the combination of LDSs and cuboids significantly improves the final performance for both class independent and class dependent modes on all the datasets. In general, the class dependent combinations achieve better results than the class independent combinations. This is intuitively plausible because the class dependent combinations specifically consider the differences

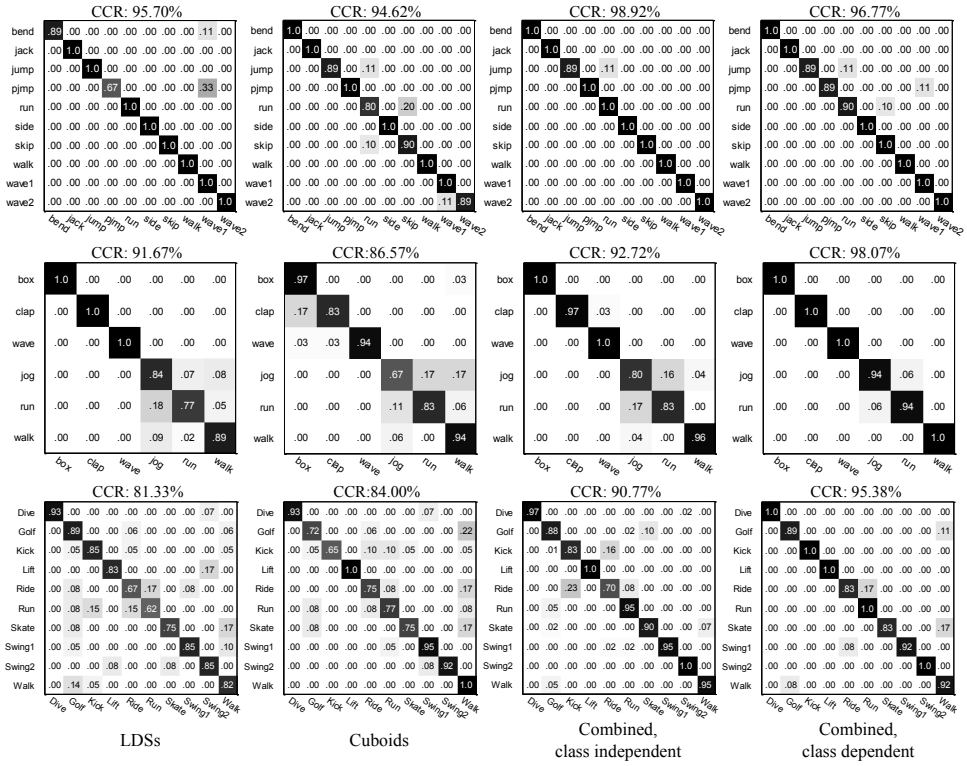

Fig. 10. Confusion matrices for depicting the results of action recognition with respect to LDSs, cuboids, combined distances with class independent and class dependent weights on three action datasets. From top to bottom: Weizmann, KTH and UCF sports, respectively.

TABLE 2

Average precision of each action class for the Hollywood2 dataset (%)

	LDSs	Cuboids	Combined (ω_1, ω_2)
AnswerPhone	52.80	25.93	49.46 (1.25, 0.68)
DriveCar	55.49	59.17	75.33 (1.58, 0.45)
Eat	41.68	41.48	53.50 (0.28, 2.19)
FightPerson	81.86	53.99	83.91 (2.04, 0.44)
GetOutCar	48.24	44.65	51.99 (1.41, 0.92)
HandShake	29.54	24.26	33.84 (1.52, 0.12)
HugPerson	30.89	35.21	38.05 (1.06, 0.96)
Kiss	61.19	42.05	67.60 (1.64, 0.46)
Run	71.19	53.82	73.31 (1.38, 0.56)
SitDown	65.72	42.61	64.11 (1.15, 0.82)
SitUp	14.11	20.63	28.94 (0.09, 1.46)
StandUp	79.95	45.39	83.23 (1.42, 0.52)
mAP	52.72	40.77	58.61

TABLE 3
Recognition accuracy for the UCF50 dataset (%)

	LDSs	Cuboids	Combined, class independent	Combined, class dependent
CCR	77.82	67.06	80.11	84.74

TABLE 4
Class independent weights of LDSs and cuboids

	Weizm.	KTH	U. sports	Hollywood2	UCF50
LDSs	1.01	1.42	1.47	N/A	0.71
Cuboids	**1.99**	**1.45**	**2.01**	N/A	**1.15**

among all the action classes, and learn the weights required to reveal these differences. In Table 4, we see that cuboids play the main role for class independent combinations on all the datasets. This is probably because the class independent combinations consider the impact on all the action classes and give trade-off results. For the Hollywood2, the class independent weights for the whole dataset are not applicable. We give the weights for each action class in Table 2. In Fig. 11, we observe a more coherent result in that LDSs play the dominant role on most of the action classes.

In Fig. 10, we see how LDSs and cuboids contribute to the combination result. Taking the UCF sports dataset as example, we see that LDSs tend to misclassify the actions 'ride', 'run' and 'skate', while cuboids tend to misclassify the actions 'golf', 'kick', 'ride', 'run' and 'skate'. We also observe that cuboids give perfect performance for the actions 'lift' and 'walk'. After combining LDSs and cuboids, the perfect performance for the actions 'lift' and 'walk' is kept, and the poor performance for the actions 'golf', 'kick', 'ride', 'run' and 'skate' is greatly improved. We achieve average recognition rates of 81.33% for LDSs, 84.00% for cuboids, 90.77% for the class independent combination and 95.38% for the class dependent combination. From Fig. 11c, we see that for actions 'dive', 'golf', 'kick', 'ride', 'run', 'skate', 'swing1' and 'walk', LDSs play the dominant role, while for actions 'lift' and 'swing2', the cuboids are more dominant. Putting this all together, we conclude that LDSs and cuboids compensate for each other in describing different aspects of the actions and thus can be combined to improve the recognition performance. In addition, we obtain that the weightings of the class dependent combination largely depend on the recognition results achieved by LDSs and cuboids individually for the specific action classes. The method with the better recognition result is given the higher weighting.

In Table 5, we compare class independent combination (CIC) to a baseline combination scheme with the weights $\boldsymbol{\omega} = [1, 1]$. We see that the our proposed combination method outperforms the baseline combination significantly on most of the datasets. The KTH dataset is an exception, but this is because the learned weights are very close to $[1, 1]$ (see Table 4). This indicates that LDSs and cuboids are equally important for the KTH dataset. Furthermore, we observe that the baseline combination does

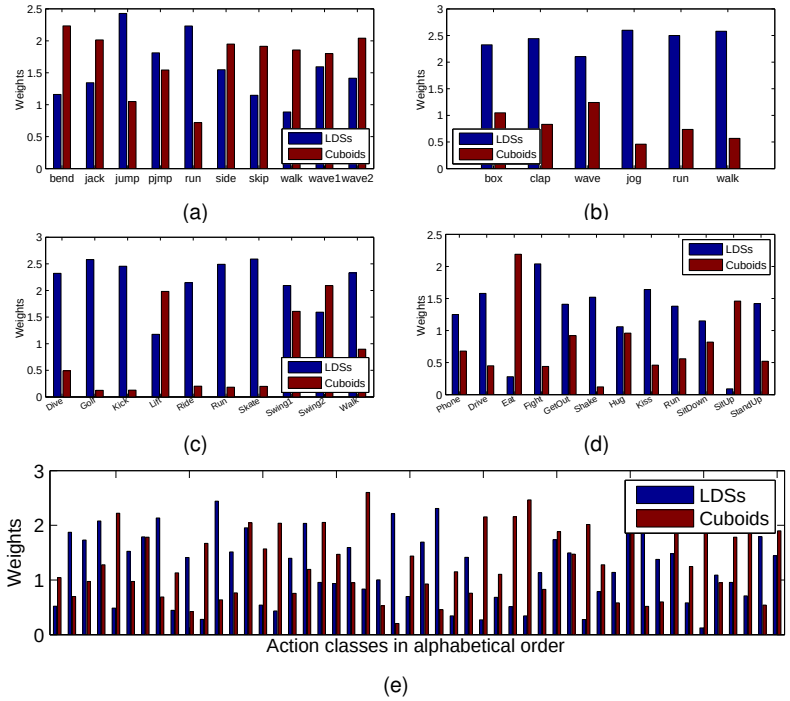

Fig. 11. Class dependent weights of LDSs and cuboids for all the action class. (a) Weizmann, (b) KTH, (c) UCF sports, (d) Hollywood2, (e) UCF50.

TABLE 5
Comparison to baseline combination scheme (%)

	Weizm.	KTH	U. sports	Hollywood2	UCF50
[1, 1]	93.55	92.47	82.12	47.81	70.86
CIC	**98.92**	**92.72**	**90.77**	**58.61**	**80.11**

not guarantee the improvement with respect to LDSs and cuboids. On the other hand, our proposed combination method significantly outperforms LDSs and cuboids on all the datasets.

6.3.5 Comparison to State-of-the-art Results

In this section, we compare our results to the state of the art on all the datasets. We show in Table 6 the most recent methods and compare our results with them. We demonstrate that our method performs consistently and stably when comparing to the state-of-the-art methods, especially those evaluated across most of the datasets.

On Weizmann, fairly high performance has been achieved by many methods, among which Wang and Suter [44] obtains the highest recognition rate of 100%. Among the methods, [42] and [9] focus on modeling and learning the motion information. Ali and Shah [42] extract a set of kinematic features such as divergence, vorticity, etc., from the optical flow for human action recognition. They report a best recognition rate of 95.75%. Chaudhry *et al.* [9] use the Binet-Cauchy kernels to capture the nonlinear dynamics of HOOF sequences, and achieve a best recognition rate of 95.66%. Our approach gives a similar recognition rate of 95.70% with single LDSs, and 98.92%, the second highest recognition rate, with class independent combination.

TABLE 6
Comparison of recognition performance to state-of-the-art results (%)

Weizmann		**KTH**		**UCF sports**	
Gorelick *et al.* [45]	97.83	Gilbert *et al.* [61]	94.50	Rodriguez *et al.* [54]	69.20
Chaudhry *et al.* [9]	95.66	Kovashka and Grauman [62]	94.53	Kovashka and Grauman [62]	87.27
Ali and Shah [42]	95.75	Wang *et al.* [10]	95.30	Wang *et al.* [10]	89.10
Bregonzio *et al.* [63]	96.66	Le *et al.* [64]	93.90	Le *et al.* [64]	86.50
Wang and Suter [44]	**100.00**	Wu *et al.* [65]	94.50	Wu *et al.* [65]	91.30
		Sadanand and Corso [66]	**98.20**	Sadanand and Corso [66]	**95.00**
LDSs alone	95.70	LDSs alone	91.67	LDSs alone	81.33
Cuboids alone	94.62	Cuboids alone	86.57	Cuboids alone	84.00
Combined, class independent	**98.92**	Combined, class independent	92.72	Combined, class independent	90.77
Combined, class dependent	96.77	Combined, class dependent	**98.07**	Combined, class dependent	**95.38**
Hollywood2		**UCF50**			
Gilbert *et al.* [61]	50.90	Reddy and Shah [56]	76.90		
Le *et al.* [64]	53.30	Sadanand and Corso [66]	57.90		
Wang *et al.* [10]	**59.90**	Wang *et al.* [10]	**85.60**		
LDSs alone	52.72	LDSs alone	77.82		
Cuboids alone	40.77	Cuboids alone	67.06		
Combined	**58.61**	Combined, class independent	80.11		
		Combined, class dependent	**84.74**		

On KTH and UCF sports, we obtain results which are better than most of the recently proposed methods and are close to the best reported in [66]. Sadanand and Corso [66] achieve their results by exploring a large set of action detectors as bases to represent action semantics. However, their method is time consuming in that it requires an average of 204 minutes to process a regular video, while our method only requires 10 minutes or so. In addition, our method outperforms their method significantly on UCF50 by nearly 20%. It is worth noting that Wu *et al.* [65] propose a similar fusion framework to ours, namely multiple kernel learning, based on two types of complementary appearance-based features. However, they report inferior results on both datasets.

On Hollywood2 and UCF50, we achieve results comparable to the best reported in [10]. Wang *et al.* [10] achieve 58.20% on Hollywood2 and 84.50% on UCF50 by combining local HOG, HOF and MBH. These results are slightly less than ours. However, they further improve their results by using spatio-temporal pyramids. They report 59.90% on Hollywood2 and 85.60% on UCF50, which are around 1% better than ours. On the other hand, our method outperforms theirs by nearly 3% on KTH and 6% on UCF sports.

6.4 Experiments on Long Continuous Datasets

6.4.1 Recognition Performance of Pre-segmented Setting

In the pre-segmented setting, we segment the videos into short clips according to the ground-truth annotations, so that each clip contains only one instance of the activities. Evaluation is carried out

TABLE 7
Recognition performance of pre-segmented setting on long continuous datasets (%)

VIRAT		ADL		CRIM13	
Amer and Todorovic [59]	76.20	Pirsiavash and Ramanan [60]	77.00	N/A	
LDSs+Cuboids	**84.49**	LDSs+Cuboids	**86.57**	LDSs+Cuboids	**88.35**

TABLE 8
Recognition performance of continuous setting on long continuous datasets (%)

VIRAT		ADL		CRIM13	
Chakraborty *et al.* [58]	11.53	Pirsiavash and Ramanan [60]	60.70	Artizzu *et al.* [16]	61.20
LDSs+Cuboids	**26.55**	LDSs+Cuboids	**66.25**	LDSs+Cuboids	**68.83**

on the resulting short clips datasets. We show in Table 7 the recognition rates on three datasets and compare to the state-of-the-art results.

On VIRAT, Amer and Todorovic [59] propose to represent each activity by a sum-product network and achieve a recognition rate of 76.20%. On ADL, Pirsiavash and Ramanan [60] achieve 77.00% by combining temporal pyramids and object detectors. Our method outperforms their results by over 8% on both datasets. On CRIM13, there is no existing result for comparison and we achieve 88.35%. We see that in the pre-segmented setting, our combination method yields good results even on large-scale datasets.

6.4.2 Recognition Performance of Continuous Setting

In the continuous setting, we test our method within temporal sliding windows, and assign the best activity labels to the center frames. The performance is measured by the percentage of frames with class agreement between the assigned labels and ground-truth annotations. We choose the sliding window sizes empirically by considering the duration of activities in each dataset. In detail, we set the window sizes to 4 seconds for the VIRAT dataset, and 10 seconds for the ADL and CRIM13 datasets, respectively. We show in Table 8 the recognition results and compare to the state-of-the-art methods.

On VIRAT, we obtain the mean AP 26.55%, which is nearly 15% better than the state-of-the-art result. This result is rather low mainly because of the small number of activity samples as discussed in [58]. On ADL and CRIM13, our method outperforms the best results by nearly 6%. We see that with no ground-truth segmentations, our method still obtains reasonable results, though the performance decreases by nearly 20% compared with those obtained in the pre-segmented setting.

7 Conclusions

In the human action recognition community, the capture of both global dynamics and local appearance features is a challenging problem. In this paper, we solve this problem by a simple yet effective ap-

proach based on combining LDSs and cuboids in a maximum margin distance learning framework. We firstly modeled the human action sequences dynamics with robust LDSs and proposed a shift invariant distance metric to measure the distance between LDSs. We then described the local appearance with HOG descriptors of the dense curved spatio-temporal cuboids. Finally we performed classification using the maximum margin distance learning method by combining the LDSs-based distances and the cuboids-based distances. We validated our method on the public human action datasets and achieved encouraging results compared to the state-of-the-art methods.

The main limitation of our method is that it is currently not designed to handle long-time sequences. Since the concerned sequences usually involve long-scale temporal variations, LDSs can not describe the embedding nonlinear dynamics alone. A possible way is to develop an automatic segmentation method to divide the long sequence into many short motion primitives, while these primitives are modeled as LDSs. The whole sequence is described as a hierarchical model by connecting these LDSs through Markov models or topic models. This is undoubtedly an interesting problem and we will consider it in the further work.

ACKNOWLEDGMENTS

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

REFERENCES

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, 2008.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] I. Laptev, "On space-time interest points," *Int'l J. Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatiotemporal features," in *Proc. IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. Int'l Conf. Multimedia*, 2007, pp. 357–360.

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

[7] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int'l Conf. Computer Vision*, 2003, pp. 726–733.

[8] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. European Conf. on Computer Vision*, 2008.

[9] 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 *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1932–1939.

[10] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int'l J. Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[11] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1992, pp. 379–385.

[12] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 994–999.

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *Proc. IEEE Int'l Conf. Computer Vision*, 2005, pp. 1808–1815.

[14] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in *Proc. Int'l Conf. Autonomous Agents and Multi-agent Systems*, 2007.

[15] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[16] X. P. Burgos-Artizzu, P. Dollár, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1322–1329.

[17] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 568–574.

[18] A. Blake, B. North, and M. Isard, "Learning multi-class dynamics," in *Proc. Ann. Conf. Neural Information Processing Systems*, 1999, pp. 389–395.

[19] V. Pavlović and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000, pp. 788–795.

[20] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "From videos to verbs: Mining videos for activities using a cascade of dynamical systems," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[21] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, 1994.

[22] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Dept. Computer Science, Univ. of Toronto, Technical Report CRG-TR-96-2, 1996.

[23] R. J. Martin, "A metric for ARMA processes," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1164–1170, 2000.

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

[25] A. B. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 846–851.

[26] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal, "Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *Int'l J. Computer Vision*, vol. 73, no. 1, pp. 95–119, 2007.

[27] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int'l J. Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[28] F. Woolfe and A. W. Fitzgibbon, "Shift-invariant dynamic texture recognition," in *Proc. European Conf. on Computer Vision*, 2006, pp. 549–562.

[29] B. Ghanem and N. Ahuja, "Maximum margin distance learning for dynamic texture recognition," in *Proc. European Conf. on Computer Vision*, 2010.

[30] T. Van Gestel, J. A. K. Suykens, P. Van Dooren, and B. De Moor, "Identification of stable models in subspace identification by using regularization," *IEEE Trans. Autom. Control*, vol. 46, no. 9, pp. 1416–1420, 2001.

[31] S. L. Lacy and D. S. Bernstein, "Subspace identification with guaranteed stability using constrained optimization," *IEEE Trans. Autom. Control*, vol. 48, no. 7, pp. 1259–1263, 2003.

[32] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. Int'l Conf. Machine Learning*, 2000, pp. 591–598.

[33] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. Syst., Man, Cybern. B*, vol. 36, no. 3, pp. 710–719, 2006.

[34] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. European Conf. on Computer Vision*, 2008.

[35] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 867–882, 2011.

[36] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *Proc. European Conf. on Computer Vision*, 2012, pp. 842–856.

[37] S. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. IEEE Int'l Conf. Computer Vision*, 2007, pp. 1–8.

[38] A. H. Shabani, D. A. Clausi, and J. S. Zelek, "Improved spatio-temporal salient feature detection for action recognition," in *Proc. British Machine Vision Conference*, 2011, pp. 100.1–100.0.

[39] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int'l Conf. Pattern Recognition*, 2004, pp. 32–36.

[40] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int'l Conf. Computer Vision*, 1999, pp. 1150–1157.

[41] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[42] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, 2010.

[43] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[44] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1646–1661, 2007.

[45] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[46] F. Caillette, A. Galata, and T. Howard, "Real-time 3-D human body tracking using learnt models of behaviour," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 112–125, 2008.

[47] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models," in *Proc. IEEE Int'l Conf. Computer Vision*, 2003, pp. 1455–1462.

[48] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[49] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, 2008.

[50] A. Bissacco, A. Chiuso, and S. Soatto, "Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1958–1972, 2007.

[51] S. M. Siddiqi, B. Boots, and G. J. Gordon, "A constraint generation approach to learning stable linear dynamical systems," in *Proc. Ann. Conf. Neural Information Processing Systems*, 2007.

[52] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM," in *Proc. Int'l Conf. Machine Learning*, 2007, pp. 807–814.

[53] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. European Conf. on Computer Vision*, 2006.

[54] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[55] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2929–2936.

[56] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications Journal*, pp. 1–11, 2012.

[57] S. Oh, A. Hoogs, A. Perera, and et al., "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3153–3160.

[58] B. Chakraborty, J. Gonzàlez, and F. X. Roca, "Large scale continuous visual event recognition using max-margin hough transformation framework," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1356–1368, 2013.

[59] M. Amer and S. Todorovic, "Sum-product networks for modeling activities with stochastic structure," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1314–1321.

[60] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2847–2854.

[61] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 883–897, 2011.

[62] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 2046–2053.

[63] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1948–1955.

[64] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3361–3368.

[65] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 489–496.

[66] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1234–1241.

Guan Luo received the BEng, MEng and PhD degrees in Electronic Engineering from Northwestern Polytechnical University in 1998, 2001 and 2004, respectively. He worked as a Senior Research Associate in RCMT, School of Creative Media, City University of Hong Kong from Jun 2004 to Aug 2005. He is currently an Assistant Professor in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include activity recognition, video analysis, and time-series data mining.

Shuang Yang received the BSc and MSc degrees in Control Science and Engineering from College of Electrical and Information Engineering, Hunan University in 2008 and 2011, respectively. She is currently a PhD candidate in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests are in motion analysis and human action understanding.

Guodong Tian received the BSc and MSc degrees in Information and Communication Engineering from Beijing University of Posts and Telecommunications in 2007 and 2010, respectively. He is currently a PhD candidate in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include activity recognition and motion pattern learning.

Chunfeng Yuan received the PhD degree from Institute of Automation, Chinese Academy of Sciences in 2010. She is currently an Assistant Professor in National Laboratory of Pattern Recognition(NLPR), Institute of Automation, Chinese Academy of Sciences. Her research interests include motion analysis, action recognition, and event detection.

Weiming Hu received the PhD degree from Department of Computer Science and Engineering, Zhejiang University in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Now he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include visual surveillance, and filtering of Internet objectionable information.

Stephen J. Maybank received the BA degree in mathematics from Kings College Cambridge in 1976 and the PhD degree in computer science from Birkbeck College, University of London in 1988. Now he is a professor in the Department of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, etc.