Pattern Recognition 46 (2013) 1710–1718

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Action recognition using linear dynamic systems

Haoran Wang [a,b], Chunfeng Yuan [b], Guan Luo [b], Weiming Hu [b,*], Changyin Sun [a]

[a] School of Automation, Southeast University, Nanjing, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China

ARTICLE INFO

Article history:
Received 20 April 2012
Received in revised form
26 November 2012
Accepted 1 December 2012
Available online 12 December 2012

Keywords:
Linear dynamic system
Kernel principal angle
Multiclass spectral clustering
Supervised codebook pruning
Action recognition

ABSTRACT

In this paper, we propose a novel approach based on Linear Dynamic Systems (LDSs) for action recognition. Our main contributions are two-fold. First, we introduce LDSs to action recognition. LDSs describe the dynamic texture which exhibits certain stationarity properties in time. They are adopted to model the spatiotemporal patches which are extracted from the video sequence, because the spatiotemporal patch is more analogous to a linear time invariant system than the video sequence. Notably, LDSs do not live in the Euclidean space. So we adopt the kernel principal angle to measure the similarity between LDSs, and then the multiclass spectral clustering is used to generate the codebook for the bag of features representation. Second, we propose a supervised codebook pruning method to preserve the discriminative visual words and suppress the noise in each action class. The visual words which maximize the inter-class distance and minimize the intra-class distance are selected for classification. Our approach yields the state-of-the-art performance on three benchmark datasets. Especially, the experiments on the challenging UCF Sports and Feature Films datasets demonstrate the effectiveness of the proposed approach in realistic complex scenarios.

1. Introduction

Automatic recognition of human actions in videos is useful for surveillance, content-based summarization, and human–computer interaction applications. Yet, it is still a challenging problem. In recent years, a large number of researchers have addressed this problem as evidenced by several survey papers [1–4].

Action representation is important for action recognition. There are appearance-based representation [5,40], shape-based representation [6,41], optical-flow-based representation [7,42], volume-based representation [8,43] and interest-point-based representation [9,44]. Among them, methods using local interest point features together with the bag of visual words model are greatly popular, due to their simple implementation and good performance. The bag of visual words approaches are robust to noise, occlusion and geometric variation, without requirement for reliable tracking on a particular subject. Despite recent developments, the representation of local regions in videos is still an open field of research.

Dynamic textures are sequences of images of moving scenes that exhibit certain stationarity properties in time, such as sea-waves, smoke, foliage, whirlwind etc. They capture the dynamic information in the motion of objects. Doretto et al. [10] show that dynamic textures can be modeled using a LDS. Tools from system identification are borrowed to capture the essence of dynamic textures. Once learned, the LDS model has predictive power and can be used for extrapolating dynamic textures with negligible computational cost. In tradition, LDS is used to describe dynamic textures of video sequence [11,12]. But a video sequence is usually not a linear time invariant system due in part to its long time span and complex changes. Compared with video sequence, the spatiotemporal patch is analogous to a linear time invariant system. Moreover, LDS exhibits more dynamic information, which is important for the representation of moving scenes, than traditional local features.

Several categorization algorithms have been proposed based on the LDS parameters, which live in a non-Euclidean space. Among these methods, Vishwanathan et al. [13] use Binet–Cauchy kernels to compare the parameters of two LDSs. Chan and Vasconcelos [14] use both the KL divergence and the Martin distance [12,15] as a metric between dynamic systems. Woolfe and Fitzgibbon [16] use the family of Chernoff distances, and the distances between cepstrum coefficients are adopted as the metrics between LDSs. These methods usually define a distance measurement between the model parameters of two dynamic systems. Once such a metric has been defined, classifiers such as nearest neighbors or support vector machines can be used to categorize a query video sequence based on the training data. However, all the above approaches are supervised classification. They are not suitable for the codebook generation in the bag of words representation.

* Corresponding author. Tel.: +86 13910900826.
E-mail address: wmhu@nlpr.ia.ac.cn (W. Hu).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.12.001

Dictionary learning is still an open problem. Successful extraction of good features from videos is crucial to action recognition. Several studies [17–19] have been made to extract discriminative visual words for classification. In a codebook, some visual words are discriminative, but some visual words are noise which has negative influence on classification. An effective dictionary learning method, which selects those discriminative visual words for classification, can improve the accuracy of recognition.

In this paper, we introduce the Linear Dynamic Systems to action recognition. We replace traditional gradient and optical flow features of interest points by LDS. LDS describes the temporal evolution in a spatiotemporal patch which is analogous to a linear time invariant system. It captures more dynamic information than traditional local features. So, we utilize LDS as the local descriptor which lives in a non-Euclidean space. To obtain the codebook for the bag of features representation, existing methods typically cluster local features by *K*-means. But the *K*-means method does not fit the non-Euclidean space. In [15], the high-dimensional non-Euclidean space is mapped to a low-dimensional Euclidean space, and then the clustering algorithm suitable for the Euclidean space is used to generate the codebook. But the transformation is an approximation, and it is not the optimal solution. In our method, we adopt the kernel principal angle [20] to measure the similarity between LDSs, and then use the multiclass spectral clustering [21–23] to compute the codebook of LDSs. As a discriminative approach, the spectral clustering does not make assumptions about the global structure of data. What makes it appealing is the global-optimal solution in the relaxed continuous domain and the nearly global-optimal solution in the discrete domain. In the codebook, not all the visual words are discriminative for classification. To extract the discriminative visual words and remove the noise in each action class, we propose a supervised codebook pruning method. The algorithm is effective and linear-complexity. Furthermore, it is fairly general and can be used to deal with many areas relative to the codebook. Fig. 1 shows the flowchart of our framework.

The remainder of this paper is organized as follows. Section 2 gives a review of related approaches about action recognition. Section 3 introduces the LDS-based codebook formation. Section 4 proposes the supervised codebook pruning method. Section 5 demonstrates the experimental results. Section 6 concludes this paper.

2. Related work

We review the related work on interest-point-based representations and dictionary learning methods for action recognition.

2.1. Interest-point-based representations

Much work has recently demonstrated the effectiveness of the interest-point-based representation on action recognition tasks. Local descriptors based on normalized pixel values, brightness gradients and windowed optical flows are evaluated for action recognition by Dollar et al. [34]. Experiments on three datasets (KTH human actions, facial expressions and mouse behavior) show that gradient descriptors achieve excellent results. Those descriptors are computed by concatenating all gradient vectors in a region or by building histograms on gradient components. Primarily based on gradient magnitudes, they suffer from sensitivity to illumination changes. Laptev and Lindeberg [35] investigate single-scale and multi-scale *N*-jets, histograms of optical flows, and histograms of gradients as local descriptors for video sequences. Best performance is obtained with optical flow and spatio-temporal gradients. Instead of the direct quantization of the gradient orientations, each component of the gradient vector is quantized separately. In later work, Laptev et al. [9,36] apply a coarse quantization to gradient orientations. As only spatial gradients have been used, histogram features based on optical flow are employed in order to capture the temporal information. The computation of optical flow is rather expensive and the result depends on the regularization method [37]. Klaser et al. [38] build a descriptor with pure spatio-temporal 3D gradients which are robust and cheap to compute. They perform orientation quantization with up to 20 bins by using regular polyhedrons. But this descriptor is still computed by building histograms on gradient components, and ignores the spatiotemporal information.

The strategy of generating compound neighborhood-based features, which are explored initially for static images and object recognition [39–41], has been extended to videos. One approach is to subdivide the space–time volume globally using a coarse grid of histogram bins [9,42–44]. Another approach places grids around the raw interest points, and designs a new representation using the positions of the interest points that fall within the grid cells surrounding that central point [29]. In contrast to previous methods which compute the feature of each interest point, the neighborhood-based features use the distribution of interest points as the descriptor.

2.2. Dictionary learning

To the best of our knowledge, not much work on dictionary learning has been reported for action recognition. Liu and Shah [17] propose an approach to automatically discover the optimal number of visual word clusters by utilizing maximization of mutual information. In later work, Liu et al. [19] use Page Rank to mine the most informative static features. In order to further construct compact yet discriminative visual vocabularies, a divisive information-theoretic algorithm is employed to group semantically related features. However, their formulation is intractable and requires approximation, and it may not learn the optimal dictionary. Brendel and Todorovic [18] store multiple diverse exemplars per activity class, and learn a sparse dictionary of most discriminative exemplars. But the exemplars are extracted based on human-body postures which are difficult to

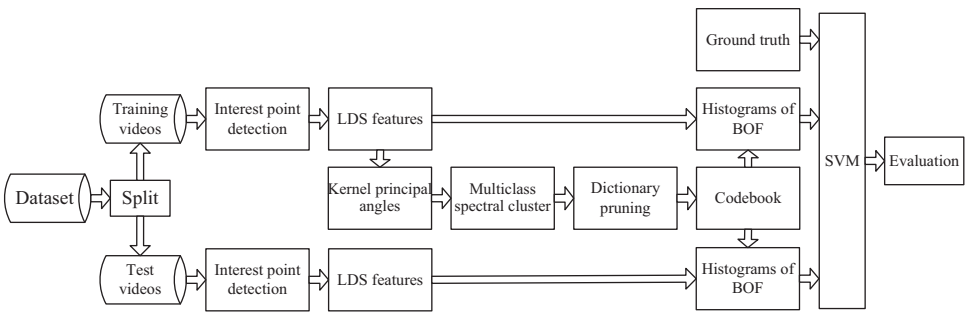

Fig. 1. Flowchart of the proposed framework.

detect. Qiu et al. [45] propose a gaussian process model to optimize the dictionary objective function. The dictionary learning algorithm is based on sparse representation which has recently received a lot of attention.

3. LDS-based codebook formation

In the traditional bag-of-words framework, once interest points and the corresponding descriptors are extracted, the descriptors are clustered using a clustering algorithm such as *K*-means to form the codebook. In our method, the Linear Dynamic System is introduced to model the spatiotemporal patch. LDSs live in a non-Euclidean space. We cannot directly apply the clustering algorithms that are used in the Euclidean space. So we propose a new method to form the codebook. The process mainly includes three steps. First, we adopt LDS to model the local spatiotemporal patch. The LDS descriptor exhibits certain stationarity properties in time, and it extracts more dynamic information than other traditional descriptors. Second, we utilize the kernel principal angle to measure the similarity between LDSs. Third, we use the multiclass spectral clustering to generate the codebook.

3.1. Linear dynamic system based descriptor

The dynamic system is a powerful tool to deal with temporally ordered data. It has been used in several applications in computer vision, such as tracking, human recognition from gait, and dynamic texture. The main idea is to use a dynamic system to model the temporal evolution of a measurement vector $I(t)\in R^n$ as a function of a relatively low dimensional state vector $z(t)\in R^d$ that changes over time.

We perform space–time interest point detection and the associated local feature extraction. To detect interest points, the method in [24] is adopted, which is a space–time extension of the Harris operator. Instead of performing scale selection, we detect interest points at multiple levels of spatiotemporal scales. The Linear Dynamic System [10] is adopted as the local descriptor because the spatiotemporal patch is more analogous to the linear time invariant system compared with the whole video.

The spatiotemporal patch is denoted as $I(t)_{t=1}^{F}$, where $I(t)$ is the feature vector of the tth frame and F is the number of frames in the patch. We compute the normalized histograms of optical flow (HOF) [9] and oriented gradient (HOG) [39], and concatenate them into a vector descriptor to characterize the measurement vector $I(t)$. The spatiotemporal patch is modeled as the output of a LDS:

$$z(t+1)=Az(t)+Bv(t) \tag{1}$$

$$I(t)=Cz(t)+w(t) \tag{2}$$

where $z(t)\in R^n$ is the hidden state at time t, $A\in R^{n\times n}$ represents the dynamics of the system, $C\in R^{p\times n}$ maps the hidden state to the output of the system, $w(t)\sim N(0,R)$ and $v(t)\sim N(0,Q)$ are the normal distributions which denote the measurement and process noise respectively, R and Q are the variances of the normal distributions. The order of the system is given by n, and p is the dimension of the feature extracted from one frame of the patch.

We describe the spatiotemporal patch using the tuple $M=(A,C)$. The advantage of the LDS descriptor is that, it encodes both the appearance of the spatiotemporal patch, which is modeled by C, and the dynamics, which is represented by A. Moreover, LDSs describe the dynamic textures which are sequences of images in moving scenes, and exhibit certain stationarity properties in time, which are never considered in previous descriptors such as optical flow, image gradient, shape based descriptor and so on. They characterize the evolution in continuous images, and pose the problem of modeling dynamic textures on an analytical footing. LDSs describe rich dynamic information in the spatiotemporal patch.

3.2. Kernel principal angle based similarity measurement between LDSs

The principal angle measures the similarity between two subspaces. It effectively measures whether the subspaces intersect, which is somewhat similar to the "nearest neighbor" approach. The kernel principal angle [20,12,15,46] introduces an appropriate function over the principal angle to form a positive definite kernel. It is suitable for measuring the similarity between LDSs which live in a non-Euclidean space.

Given two LDS descriptors, $M_1=(A_1,C_1)$ and $M_2=(A_2,C_2)$, we adopt the kernel principal angle to measure the similarity between them. The subspace angle is defined as the principal angle between the observability subspaces associated with two model parameters A and C. The calculation of the subspace angle between two models is performed by the solution for P from the Lyapunov equation $A^TPA-P=-C^TC$, where

$$P=\begin{bmatrix}P_{11} & P_{12}\\ P_{21} & P_{22}\end{bmatrix}\in R^{2n\times 2n},\quad A=\begin{bmatrix}A_1 & 0\\ 0 & A_2\end{bmatrix}\in R^{2n\times 2n}$$

$$C=[C_1C_2]\in R^{p\times 2n}$$

The cosines of the principal subspace angles $\theta_{i\,i=1}^{n}$ are calculated by the first n largest eigenvalues of matrix $P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}$:

$$\cos^2\theta_i=i\text{th eigenvalue}\left(P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}\right) \tag{3}$$

The similarity between M_1 and M_2 is defined as

$$S(M_1,M_2)=\prod_{i=1}^{n}\cos^2\theta_i \tag{4}$$

3.3. Codebook generation based on multiclass spectral clustering

Given the similarity between each pair of local LDS descriptors through the kernel principal angle mentioned above, we adopt the multiclass spectral clustering method to generate the codebook for the bag of words representation. The reasons are threefold. First, the multiclass spectral clustering method is not limited by the Euclidean case. Second, compared with other clustering methods, the multiclass spectral clustering method converges fast. Furthermore, it is robust to random initialization, and the resulting discrete solution is nearly global-optimal.

A weighted graph is specified by $G=(V,E,W)$, where V is the set of all nodes; E is the set of edges connecting the nodes; W is the affinity matrix, which is assumed to be non-negative and symmetric. In our method, V is the collection of all the LDS descriptors extracted from videos, E is the set of edges connecting each pair of the LDS descriptors, and $W(i,j)=S(M_i,M_j)$ is the similarity between LDSs. Let $V',V''\subset V$ be two subsets of V. As in [21], the $links(V',V'')$ is defined to be the total weighted connections from V' to V'':

$$links(V',V'')=\sum_{i\in V',j\in V''}W(i,j) \tag{5}$$

The *degree* of a set V' is the total links to all the nodes:

$$degree(V')=links(V',V) \tag{6}$$

Using the degree as a normalization term, the $linkratio(V',V'')$ is defined as

$$linkratio(V',V'')=\frac{links(V',V'')}{degree(V')} \tag{7}$$

The K-way partition is denoted by $\Gamma_V^K = V_1,\ldots,V_K$. The normalized cut $knassoc\left(\Gamma_V^K\right)$ is defined as

$$knassoc\left(\Gamma_V^K\right) = \frac{1}{K}\sum_{l=1}^{K} linkratio(V_l,V_l) \tag{8}$$

The $N\times K$ partition matrix X is used to represent Γ_V^K, where N denotes the number of the LDS descriptors, and K denotes the size of the codebook. Let $X=[X_1,\ldots,X_K]$, where X_l is a binary indicator for V_l.

$$X(i,l) = \langle i \in V_l \rangle,\ i \in V,\ l \in [1,\ldots,K] \tag{9}$$

where $\langle\ .\rangle$ is 1 if the argument is true and 0 otherwise. A node is assigned to one and only one partition. The degree matrix D is defined for the symmetric weight matrix W to be:

$$D(i,j) = \begin{cases} \sum_{k=1}^{N} W(i,k) & \text{if } i=j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

According to Eqs. (5) and (6), the *links* and *degree* can be rewritten as

$$links(V_l,V_l) = X_l^T W X_l \tag{11}$$

$$degree(V_l) = X_l^T D X_l \tag{12}$$

According to Eq. (8), the K-way normalized cuts criterion is expressed in an optimization program of the variable X:

$$\begin{aligned} \max\varepsilon(X) = knassoc\left(\Gamma_V^K\right) &= \frac{1}{K}\sum_{l=1}^{K} linkratio(V_l,V_l) \\ &= \frac{1}{K}\sum_{l=1}^{K} \frac{links(V_l,V_l)}{degree(V_l)} \\ &= \frac{1}{K}\sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l} \\ \text{subject to } X \in 0,1^{N\times K},\quad & \sum_{l=1}^{K} X_l = 1_N \end{aligned} \tag{13}$$

In this way, we obtain the codebook according to the partition matrix X. Each descriptor is assigned to a corresponding visual word.

4. Supervised codebook pruning

To extract the discriminative visual words and suppress the noise in each action class, we propose a supervised codebook pruning method to refine the original codebook. Our method mainly contains two steps. First, given a codebook with n visual words, each visual word is associated with a weight. We propose a supervised weight updating method to optimize the weight vector w of the codebook, and aim at assigning larger weights to more discriminative visual words. Second, we preserve the visual words with large weights, and remove the visual words with small weights. In this way, the codebook is pruned in order to improve the classification. We specify the whole process as follows.

Let x_p and x_q denote the histogram representations of two videos. The distance between them is defined as $d(x_p,x_q)=|(x_p-x_q)^T|w$, where $|\cdot|$ is to compute the absolute value for each element of the vector. Let x_i^j, where $i=1,\ldots,K'$ and $j=1,\ldots,m_i$, denotes the histogram representation of a training sample in the action class i. There are m_i videos in class i. Our goal is to extract the most discriminative visual words in each action class. For action class k, we define a class specific weight vector w_k. Afterwards, we propose a supervised method to update the weight vector by maximizing the distances between videos belonging to class k and all out of class ones, and minimizing the intra-class distance in class k. The supervised weight updating process for class k is specified as follows:

(1) To maximize the inter-class distance, we solve the following linear program:

$$\underset{w_k}{\operatorname{argmax}}\left\{\sum_{\substack{i=1\\ i\neq k}}^{K'}\sum_{j=1}^{m_i}\left|\left(x_i^j-\bar{x}_k\right)^T\right|w_k\right\} \tag{14}$$

$$\bar{x}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} x_k^j \tag{15}$$

$$\text{subject to } \|w_k\|_1 = \gamma,\quad w_k \geq 0 \tag{16}$$

leading to

$$\underset{w_k}{\operatorname{argmax}}\left\{\sum_{\substack{i=1\\ i\neq k}}^{K'}\sum_{j=1}^{m_i}\left|\left(x_i^j-\bar{x}_k\right)^T\right|\frac{w_k}{\|w_k\|_1}\right\} \tag{17}$$

$$\text{subject to } w_k \geq 0 \tag{18}$$

The non-negative constraint in Eq. (17) can be reformulated by using the following substitutions $\left|\left(x_i^j-\bar{x}_k\right)^T\right| = [z_{i1}^j,\ldots,z_{it}^j,\ldots,z_{in}^j]$ and $w_k=[v_1^2,\ldots,v_t^2,\ldots,v_n^2]^T$, where $v=[v_1,\ldots,v_t,\ldots,v_n]^T$ is an auxiliary variables, t is the index over all the visual words. This gives

$$\underset{v_t}{\operatorname{argmax}}\left\{\sum_{\substack{i=1\\ i\neq k}}^{K'}\sum_{j=1}^{m_i}\frac{1}{U(v)}\sum_t z_{it}^j v_t^2\right\} \tag{19}$$

where $U(v)=\|v\|_2^2$. Through all the training samples, we update the weight vector w_k for action class k.

$$v_t \leftarrow v_t + \eta\frac{z_{it}^j U(v) - \sum_t z_{it}^j v_t^2}{U^2(v)} v_t \tag{20}$$

where η is the learning rate.

(2) The intra-class distance minimization process is as the following linear program:

$$\underset{w_k}{\operatorname{argmin}}\left\{\sum_{r=1}^{m_k}\sum_{\substack{j=1\\ j\neq r}}^{m_k}\left|\left(x_k^j-x_k^r\right)^T\right|w_k\right\} \tag{21}$$

$$\text{subject to } \|w_k\|_1 = \gamma,\quad w_k \geq 0 \tag{22}$$

leading to

$$\underset{w_k}{\operatorname{argmin}}\left\{\sum_{r=1}^{m_k}\sum_{\substack{j=1\\ j\neq r}}^{m_k}\left|\left(x_k^j-x_k^r\right)^T\right|\frac{w_k}{\|w_k\|_1}\right\} \tag{23}$$

$$\text{subject to } w_k \geq 0 \tag{24}$$

The non-negative constraint in Eq. (23) is reformulated by the following substitutions: $\left|\left(x_k^j-x_k^r\right)^T\right| = [z_{k1}^j,\ldots,z_{kt}^j,\ldots,z_{kn}^j]$ and

$w_k=[v_1^2,\ldots,v_t^2,\ldots,v_n^2]^T$. This gives

$$\underset{v_t}{\operatorname{argmin}}\left\{\sum_{r=1}^{m_k}\sum_{\substack{j=1\\ j\neq r}}^{m_k}\frac{1}{U(v)}\sum_t z_{kt}^j v_t^2\right\} \tag{25}$$

We update the weight vector w_k for action class k.

$$v_t \leftarrow v_t-\eta\frac{z_{kt}^j U(v)-\sum_t z_{kt}^j v_t^2}{U^2(v)}v_t \tag{26}$$

where η is the learning rate. Once v_t is estimated, the weights of visual words are computed by $w_k=[v_1^2,\ldots,v_t^2,\ldots,v_n^2]^T$.

After weight updating, we preserve the visual words with large weights. The preserved visual words are considered to be representative for class k. We repeat the above process for each action class, and then collect all the preserved visual words together to form the pruned codebook for action classification. The whole process of codebook pruning is illustrated in Algorithm 1.

Algorithm 1. Supervised codebook pruning method

Input: Codebook
for $k=1$ to K' **do**
Initialize: w_k, $k\in(1,\ldots,K')$
1. Maximize the inter-class distance.
2. Minimize the intra-class distance.
3. Preserve the representative visual words for class k.
end
Output: Pruned codebook

Under the pruned codebook, each video is represented by a histogram of the preserved visual words. The SVM is adopted as the classifier, and we employ the gaussian kernel with the χ^2 distance. The χ^2 distance between two bag-of-words histograms H_i and H_j is defined as

$$\chi^2(H_i,H_j)=\frac{1}{2}\sum_{b=1}^{N}\left(\frac{(H_i(b)-H_j(b))^2}{H_i(b)+H_j(b)}\right) \tag{27}$$

where b indexes over each of the N histogram bins. The gaussian kernel is defined as

$$K(H_i,H_j)=\exp\left(-\frac{1}{A}\chi^2(H_i,H_j)\right) \tag{28}$$

where A is the kernel's scale parameter, and is set to the mean distance between training samples.

5. Experiments

Our experiments evaluate the proposed approach for action recognition with a variety of categories. In addition to reporting the overall accuracy, we analyze the reasons of our performance. We compare the LDS descriptor with some other popular descriptors under the bag of words model. Moreover, we test the influence of our proposed codebook pruning method on the recognition rate.

5.1. Datasets and implementation details

We evaluate our approach on three benchmark datasets for human action recognition: the KTH action dataset [25], the UCF Sports dataset [26], and the Feature Film dataset [26]. All the video clips contain primarily a single action of interest. Examples of the datasets are shown in Fig. 2.

The KTH action dataset contains six types of human actions (boxing, hand waving, hand clapping, walking, jogging, and running), performed repeatedly by 25 persons in four different scenarios: outdoors, outdoors with camera zoom, outdoors with different clothes, and indoors. Twenty-four persons' videos are used as the training set and the remaining one person's videos as the test set. The results are the average of 25 times runs.

In order to evaluate the effectiveness of our method to recognize complex actions, we conduct experiments on the UCF Sports dataset. Different from the KTH dataset, the UCF Sports is a challenging dataset for action recognition. It collects a large set of action clips from various broadcast sport videos. The actions are captured in a wide range of scenes and viewpoints. The dataset is tested in a 9-way recognition task in a leave-one-out manner, cycling each example as a test video one at a time.

Rodriguez et al. [26] collect a dataset named Feature Films. It contains actions performed in a range of film genres consisting of classic old movies, comedies, scientific movies, fantasy movies, and romantic movies. This dataset provides a representative pool of natural samples including 92 samples of "Kissing" and 112 samples of "Hitting/slapping". The extracted samples appear in a wide range of scenes and view points, and are performed by different actors. Testing for this dataset proceeds in a leave-one-out framework. Given the significant intra-class variability presented in the movie scenes, the recognition task is challenging.

We extract sparse Harris 3D points for the KTH dataset, and perform dense and multi-scale interest point extraction for the UCF Sports and Feature Films dataset. The spatiotemporal patch centered around the interest point is extracted, and its size is empirically set to $13\times13\times8$. To generate the codebook, we empirically set the codebook size K for the KTH dataset to 400, and that for the UCF Sports and Feature Films dataset to 3000. When we prune the codebook, the discriminative visual words whose weights increase after codebook learning are preserved, and the visual words whose weights decrease are removed.

5.2. Action recognition performance

We report the overall accuracy on three datasets. The KTH dataset is a standard benchmark for human action recognition.

Fig. 2. Representative frames from videos in three datasets: examples in row 1 are from the KTH dataset, examples in row 2 are from the UCF Sports dataset, and examples in row 3 are from the Feature Films dataset.

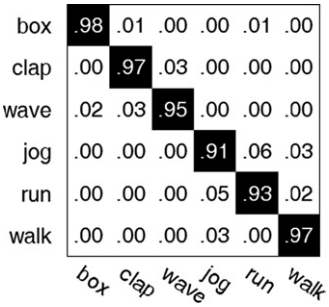

Fig. 3. Confusion matrix on the KTH dataset.

Table 1
Comparison with previous work on the KTH dataset.

Approach	Year	Accuracy (%)
Laptev et al. [9]	2008	91.80
Bregonzio et al. [28]	2009	93.17
Liu et al. [19]	2009	93.80
Gilbert et al. [29]	2009	94.50
Niebles et al. [30]	2010	91.30
Brendel and Todorovic [18]	2010	94.22
Li et al. [11]	2011	93.60
Le et al. [31]	2011	93.90
Wang et al. [47]	2012	94.17
Our method		95.17

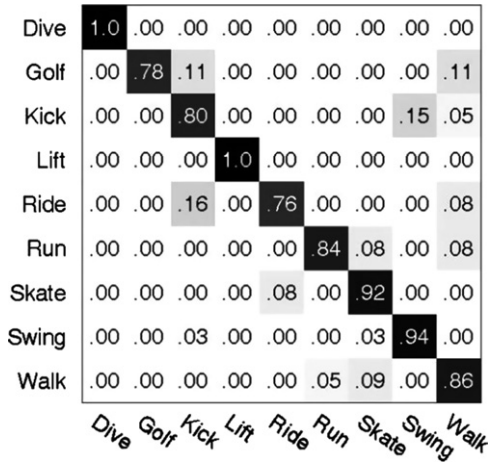

Fig. 4. Confusion matrix on the UCF Sports dataset.

Fig. 3 shows the confusion matrix of our approach on the KTH dataset. Most classes are almost perfectly predicted, except for running and jogging, which are a little confused with each other. Table 1 compares our result with those from previous work. The recognition rate is raised to 95.17%. Our method outperforms previously published results, and achieves the state-of-the-art performance. It validates that the LDS descriptors offer discriminative power, and capture temporal stationarity properties which are indeed helpful for action representation.

Fig. 4 presents the confusion matrix across all scenarios on the UCF Sports dataset. Our method works well on most actions. For example, the accuracies for some actions are high up to 100%, such as "diving" and "lifting". Table 2 compares the overall mean accuracy of our method with the results reported by previous researchers. The average accuracy of our method is 87.3%, which is the state-of-the-art performance. This result further indicates that our LDS based approach is robust. Even in the challenging and realistic action dataset, the proposed method also achieves reliable recognition rate.

Table 2
Comparison with previous work on the UCF Sports dataset.

Approach	Year	Accuracy (%)
Rodriguez et al. [26]	2008	69.2
Yeffet and Wolf [32]	2009	79.2
Wang et al. [33]	2009	85.6
Yao et al. [27]	2010	86.6
Le et al. [31]	2011	86.5
Wang et al. [47]	2012	86.6
Our method		87.3

Table 3
Results on the Feature Films dataset.

Class	Rodriguez et al. [26] (%)	Yeffet and Wolf [32] (%)	Our method (%)
Kissing	66.4	77.3	86.3
Slapping	67.2	84.2	89.6
Average	66.80	80.75	87.95

Table 4
Comparison of different descriptors on three datasets.

	KTH (%)	UCF (%)	Feature Films (%)
HOG	83.14	79.3	76.77
HOF	90.15	83.4	79.13
HOG+HOF	92.65	84.7	80.31
HOG3D	91.48	82.7	81.10
LDS	**95.17**	**87.3**	**87.95**

Table 3 shows the performance of our method on the Feature Films dataset, and compares our approach with the LTP proposed in [32] and the original method in [26]. In both categories, our method outperforms previously reported results, and achieves the state-of-the-art performance. There is an increase in average accuracy between our method (87.95%) and the closest result (80.75%). The excellent performances on the challenging Feature Films and UCF Sports datasets further validate the effectiveness of our method.

5.3. Impact of the LDS descriptor

In our approach, we adopt the Linear Dynamic System as the descriptor along with the bag-of-words model. The bag-of-words model is usually associated with some descriptors, such as HOG, HOF and HOG3D. HOG and HOG3D only describe the gradient feature which contains no temporal information. The optical flow only describes the motion of objects between two continuous frames in a spatiotemporal patch, and it cannot capture the temporal evolution process of features in a whole patch. LDS describes the feature of dynamic textures which are sequences of moving scenes that exhibit stationarity properties in time. Compared with traditional descriptors, the LDS model captures more dynamic information. Different from static images, the dynamic information is very important for describing actions in videos. So, LDS is more suitable for action representation than traditional descriptors.

In order to confirm the advantage of LDS in action representation, we compare different descriptors in the same condition. Under the bag-of-words model, we compare LDS with some other popular and powerful descriptors on three datasets, as shown in Table 4. On the KTH dataset, the combination of HOG and HOF has a better result than HOG3D which obtains a comparable performance. LDS achieves the best performance. There is an increase in recognition rate between our method (95.17%) and the closest competitive descriptor

(92.65%). The experiments on the challenging UCF Sports and Feature Films datasets further validate the robustness of LDS. Similar to the results on the KTH dataset, LDS still performs better than other descriptors. Experiments on the three datasets validate that LDS is effective for action representation.

5.4. Analysis of the codebook pruning

In the codebook, not all the visual words are representative for their own action class. So our codebook pruning method aims at preserving discriminative visual words and removing noise. Next we run experiments to support our claim. Two experiments are performed. First, for the codebook, we compare the recognition rates corresponding to different numbers of preserved visual words. When the number of visual words removed increases, we observe how the recognition rate changes. Second, for different sizes of the codebook, we compare the influence of our proposed codebook pruning method on the improvement of the recognition rate.

In the first experiment, we empirically set $K=400$ for the codebook size. We record the number of preserved most discriminative visual words and their corresponding recognition rates on the KTH dataset, as illustrated in Fig. 5. When all the visual words are used for classification, the recognition rate is 94.22%, which is a comparable result. Then we prune the codebook, and the number of preserved visual words decreases gradually. When the number of visual words is pruned to 340, the accuracy is raised to 95.17%. This accuracy is maintained until the number is 260. The preserved visual words are the most discriminative features for classification according to our approach. Then, the accuracy begins to fluctuate. When the number of preserved visual words is less than 130, the accuracy descends linearly to 80%. When the visual words are redundant or insufficient, it is difficult to obtain the best performance. The whole process illustrates that the proposed codebook pruning method is effective for increasing the recognition rate, but the high recognition rate requires not only removing noise but also enough discriminative visual words.

In the second experiment, we further test the performance of the codebook pruning method. We compare the recognition rate before codebook pruning with that after codebook pruning when the size of the codebook ranges from 100 to 1000 on the KTH dataset, as illustrated in Fig. 6. The recognition rate fluctuates from 85% to 95% when the size of the codebook is larger than 250.

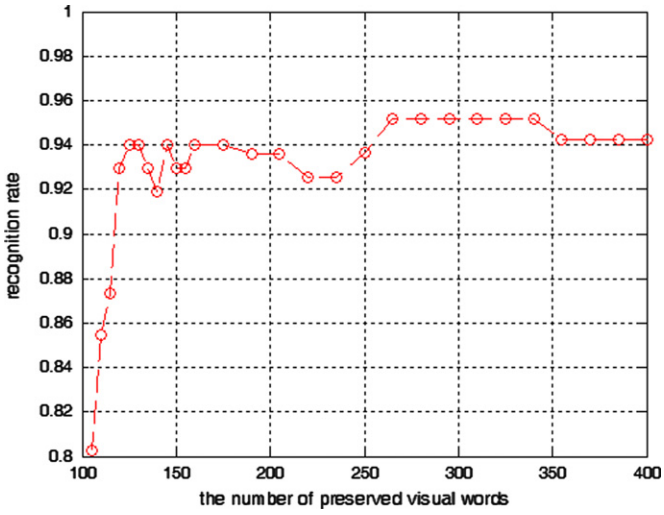

Fig. 5. Performance of supervised codebook pruning.

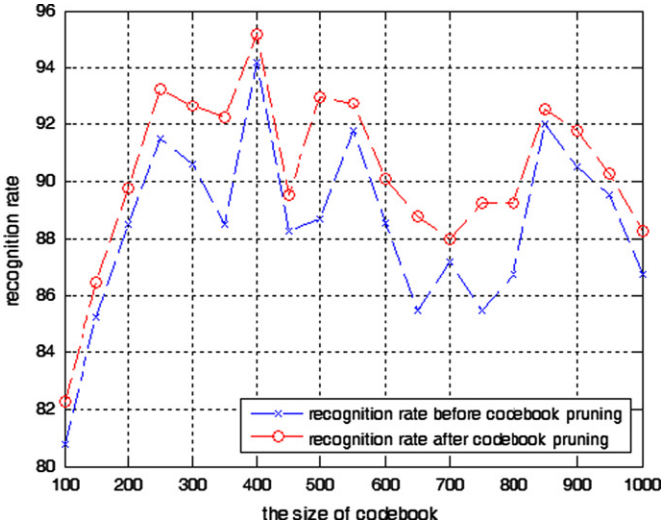

Fig. 6. The performances of pruning in different codebook sizes.

The best performance is obtained when the size is 400. The dependency of the recognition rate on the size of the codebook is not serious. But when the size is too small, the recognition rate is very low. So enough visual words are necessary for a high accuracy. If the number of visual words is too small, the codebook is not discriminative for classification. No matter what the size of the codebook is, the recognition rates after codebook pruning are all higher than that before codebook pruning. It validates the effectiveness and robustness of the proposed codebook pruning method.

6. Conclusions

In this paper, we have presented a method based on Linear Dynamic Systems for action recognition. LDS is a powerful tool to represent temporally ordered data. Compared with traditional descriptors, LDS exhibits more dynamic information in actions. It is suitable for action representation. Because the LDS feature lives in a non-Euclidean space, the kernel principal angle and the multiclass spectral clustering are used to generate the codebook for the bag-of-features model. In the codebook, not all the visual words are discriminative. There are also some visual words which have negative influence on classification. We have proposed a supervised codebook pruning method to extract those discriminative visual words and remove noise for better performance. Our framework has achieved the state-of-the-art performances on the above datasets. Experimental results have validated the effectiveness and robustness of our LDS based approach for action recognition.

Acknowledgment

This work is carried out at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. It is supported in part by the scientific research foundation of Graduate School of Southeast University, the NSFC (Grant nos. 60825204, 60935002, 61100099 and 61005008), Beijing Natural Science Foundation (4121003), and the National 863 High-Tech R&D Program of China (Grant no. 2012AA012504).

References

[1] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (6) (2010) 976–990.
[2] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.
[3] A. Bobick, V. Kruger, On human action, in: Proceedings of the Visual Analysis of Humans, 2011, pp. 279–288.
[4] T.B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2–3) (2006) 90–126.
[5] I. Kotsia, S. Zafeiriou, I. Pitas, Texture and shape information fusion for facial expression and facial action unit recognition, Pattern Recognition 41 (3) (2008) 833–851.
[6] J. Zhang, S. Gong, Action categorization with modified hidden conditional random field, Pattern Recognition 43 (1) (2010) 197–203.
[7] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.
[8] T. Syeda-Mahmood, A. Vasilescu, S. Sethi, Recognizing action events from multiple viewpoints, in: Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video, 2001.
[9] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[10] G. Doretto, A. Chiuso, Y. Wu, S. Soatto, Dynamic textures, International Journal of Computer Vision 51 (2) (2003) 91–109.
[11] B. Li, M. Ayazoglu, T. Mao, O. Camps, M. Sznaier, Activity recognition using dynamic subspace angles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[12] A. Chan, N. Vasconcelos, Classifying video with kernel dynamic textures, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[13] S. Vishwanathan, A. Smola, R. Vidal, Binet–Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes, International Journal of Computer Vision 73 (1) (2007) 95–119.
[14] A. Chan, N. Vasconcelos, Probabilistic kernels for the classification of auto-regressive visual processes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[15] A. Ravichandran, R. Chaudhry, R. Vidal, View-invariant dynamic texture recognition using a bag of dynamical systems, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[16] F. Woolfe, A. Fitzgibbon, Shift-invariant dynamic texture recognition, in: Proceedings of European Conference on Computer Vision, 2006.
[17] J. Liu, M. Shah, Learning human actions via information maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[18] W. Brendel, S. Todorovic, Activities as time series of human postures, in: Proceedings of European Conference on Computer Vision, 2010.
[19] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[20] L. Wolf, A. Shashua, Kernel principal angles for classification machines with applications to image sequence interpretation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003.
[21] S.X. Stella, J. Shi, Multiclass spectral clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2003.
[22] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.
[23] J. Malik, S. Belongie, T. Leung, J. Shi, Contour and texture analysis for image segmentation, International Journal of Computer Vision 43 (1) (2001) 7–27.
[24] I. Laptev, On space–time interest points, International Journal of Computer Vision 64 (2) (2005) 107–123.
[25] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local SVM approach, in: Proceedings of International Conference on Pattern Recognition, 2004.
[26] M.D. Rodriguez, J. Ahmed, M. Shah, Action mach: a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[27] A. Yao, J. Gall, L.V. Gool, A Hough transform-based voting framework for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[28] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space–time interest points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[29] A. Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal feature, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[30] J. Niebles, C. Chen, L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, in: Proceedings of European Conference on Computer Vision, 2010.
[31] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[32] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[33] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of the British Machine Vision Conference, 2009.
[34] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.
[35] I. Laptev, T. Lindeberg, Local descriptors for spatio-temporal recognition, Spatial Coherence for Visual Motion Analysis (2006).
[36] I. Laptev, P. Perez, Retrieving actions in movies, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.
[37] S. Baker, S. Roth, D. Scharstein, M. Black, J. Lewis, R. Szeliski, A database and evaluation methodology for optical flow, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.
[38] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in: Proceedings of the British Machine Vision Conference.
[39] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[40] H. Jiang, M. Crew, Z. Li, Successive convex matching for action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[41] S. Xiang, F. Nie, Y. Song, C. Zhang, Contour graph based human tracking and action sequence recognition, Pattern Recognition 41 (12) (2008) 3653–3664.
[42] M. Ahmad, S. Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (7) (2008) 2237–2252.
[43] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[44] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial–temporal words, International Journal of Computer Vision 79 (3) (2008) 299–318.
[45] Q. Qiu, Z. Jiang, R. Chellappa, Sparse dictionary-based representation and recognition of action attributes, in: Proceedings of the IEEE International Conference on Computer Vision, 2011.
[46] K.D. Cock, B.D. Moor, Subspace angles between linear stochastic models, in: Proceedings of the IEEE Conference on Decision and Control, 2000.
[47] H. Wang, C. Yuan, W. Hu, C. Sun, Supervised class-specific dictionary learning for sparse modeling in action recognition, Pattern Recognition 45 (11) (2012) 3902–3911.

Haoran Wang received the B.S. degree from the Department of Information Science and Technology, Northeast University, Shenyang, China, in 2008. Now, he is a Ph.D. student in School of Automation at the Southeast University, Nanjing, China.

Chunfeng Yuan received the B.S. and M.S. degrees in information science and technology from the Qingdao University of Science and Technology, China, in 2004 and 2007, respectively, and the Ph.D. degree in 2010 from the National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. She is currently working as an assistant professor at Institute of Automation, Chinese Academy of Sciences. Her main research interests include activity analysis and pattern recognition.

Guan Luo is currently an Assistant Professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He received the B.Eng., M.Eng. and Ph.D. 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 June 2004 to August 2005. His current research interest is on web content analysis, video data mining, pattern recognition, and machine learning. Until now he has published tens of relevant papers in the national core journals and conferences.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University. From April 1998 to March 2000, he was a Postdoctoral Research Fellow with the Institute of Computer Science and Technology, Founder Research and Design Center, Peking University. Since April 2000, he has been

with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Now he is a Professor, a Ph.D. Student Supervisor in the laboratory. He is a senior member of the IEEE.

Changyin Sun is a professor in School of Automation at the Southeast University, China. He received the M.S. and Ph.D. degrees in Electrical Engineering from the Southeast University, Nanjing, China, respectively, in 2001 and 2003. His research interests include Intelligent Control, Neural Networks, SVM, Pattern Recognition, Optimal Theory, etc. He has received the First Prize in Nature Science from Ministry of Education, China. Professor Sun is a member of the IEEE, an Associate Editor of IEEE Transactions on Neural Networks, Neural Processing Letters and International Journal of Swarm Intelligence Research, Recent Patents on Computer Science.