ELSEVIER
Volume 45, Issue 11
November 2012
ISSN 0031-3203
PATTERN RECOGNITION
Available online at www.sciencedirect.com
SciVerse ScienceDirect

Pattern Recognition 45 (2012) 3902–3911

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Supervised class-specific dictionary learning for sparse modeling in action recognition

Haoran Wang [a,b,*], Chunfeng Yuan [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 9 June 2011
Received in revised form
13 February 2012
Accepted 25 April 2012
Available online 4 May 2012

Keywords:
Compound feature
Motion segment
Sparse model
Supervised dictionary learning
Action recognition

ABSTRACT

In this paper, we propose a new supervised classification method based on a modified sparse model for action recognition. The main contributions are three-fold. First, a novel hierarchical descriptor is presented for action representation. To capture spatial information about neighboring interest points, a compound motion and appearance feature is proposed for the interest point at low level. Furthermore, at high level, a continuous motion segment descriptor is presented to combine temporal ordering information of motion. Second, we propose a modified sparse model which incorporates the similarity constrained term and the dictionary incoherence term for classification. Our sparse model not only captures the correlations between similar samples by sharing dictionary, but also encourages dictionaries associated with different classes to be independent by the dictionary incoherence term. The proposed sparse model targets classification, rather than pure reconstruction. Third, in the sparse model, we adopt a specific dictionary for each action class. Moreover, a classification loss function is proposed to optimize the class-specific dictionaries. Experiments validate that the proposed framework obtains the performance comparable to the state-of-the-art.

1. Introduction

In recent years, a large number of approaches have been proposed to fulfill action recognition. Among them, bag of visual words approaches are greatly popular, due to their simple implementation and good reliability. A video clip is summarized by the histogram of its local features. By fully exploiting local space–time features, the bag of visual words approaches are robust to noise, occlusion and geometric variation, without requiring reliable tracks on a particular subject. Recent work has shown promising results using local space–time features together with bag of visual words models. The methods in Refs. [1–3] are classical interest-point-based methods for action recognition. These approachs extract the local feature from a single interest point, and achieve good results. However, the conventional interest-point-based methods describe the feature of a single interest point. They are mainly based on the individual power of the interest point, and therefore do not consider the spatio-temporal relationship between them. As an improvement, Gilbert et al. [4] perform dense interest point detection, and compute the distribution of interest points in a small area. Although this approach utilizes some spatial information, it does not exhibit the temporally ordering information in actions. A key limitation of interest-point-based representation is failing to capture adequate spatial or temporal information.

Sparse representation has received a lot of attention from the signal processing community due in part to the fact that various signals such as audio and natural images can be well approximated by a linear combination of a few elements of some redundant bases, usually called dictionary. Recent publications about sparse representation have shown that this approach is very effective, leading to state-of-the-art results, e.g., in image restoration, image denoising, texture classification and texture synthesis. In the supervised or weakly supervised methods, algorithms adopt features of the sparse coding of signals for classification [5–9]. But the sparse models mainly consider minimizing the reconstruction error. Little attention is paid to better classification.

Recent research on dictionary learning for sparse coding has been targeted on learning discriminative sparse models instead of the purely reconstructive ones. Mairal et al. [10] generalize the reconstructive sparse dictionary learning process by optimizing the sparse reconstruction jointly with a linear prediction model. Bradley and Bagnell [11] propose a novel differentiable sparse prior rather than the conventional L_1 norm, and employ a back propagation procedure to train the dictionary for sparse coding in order to minimize the training error. These approaches need to

[*] Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China. Tel.: +86 136 93277219.
E-mail address: whr1fighting@gmail.com (H. Wang).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.04.024

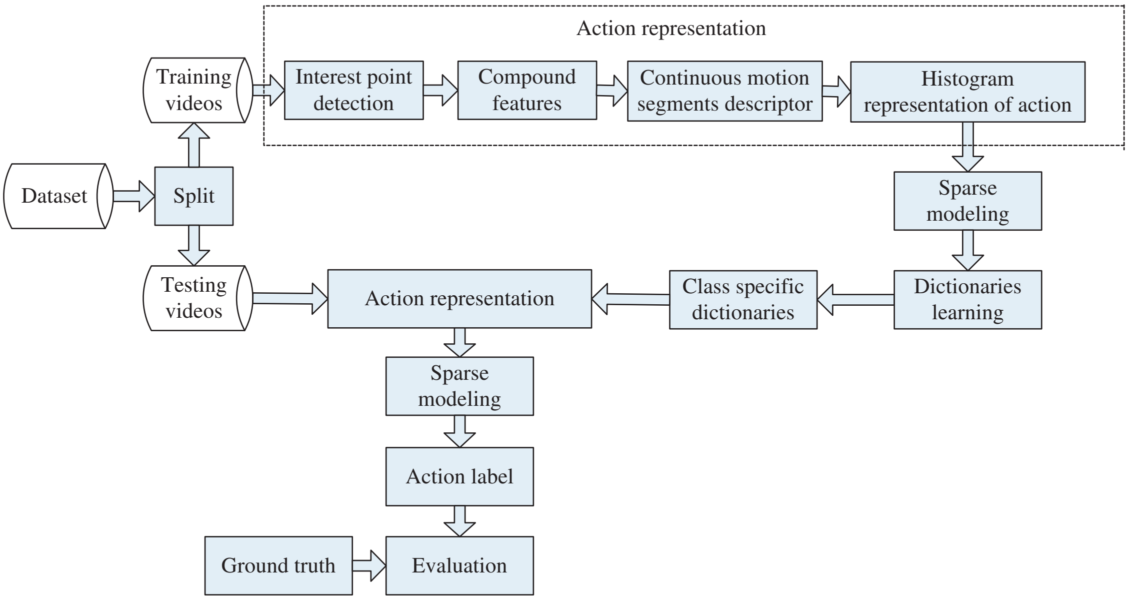

Fig. 1. Flowchart of the proposed framework.

explicitly associate each sample with a label in order to perform the supervised training. How to learn a discriminative dictionary for both sparse data representation and classification is still an open problem.

In this paper, we solve the three problems aforementioned and present a novel framework for action recognition. Fig. 1 shows the flowchart of our framework. We make the following three contributions.

First, traditional interest-point-based representation only utilizes features of a single interest point, and fails to capture adequate spatial or temporal information. It is sensitive to the noise. We consider that spatial and temporal information is very important for action representation. So a novel hierarchical descriptor is presented. The proposed compound appearance and motion feature captures spatial information of neighboring interest points. Furthermore, we propose a continuous motion segment descriptor to represent human action by capturing the temporal ordering information in actions. Spatial and temporal context information is utilized for action representation.

Second, we propose a modified sparse model for classification. Different from traditional sparse representation whose only task is to minimize the reconstruction error, the proposed sparse model targets at classification. Given K action classes, we learn K class-specific dictionaries for representing the data, and then classify the test sample into the class whose dictionary generates the minimum reconstruction error. The similarity-constrained term is utilized to project each descriptor into its local coordinate system which captures the correlations between similar descriptors by sharing bases. The dictionary incoherence term ensures that samples from different classes are reconstructed by independent dictionaries. Our proposed sparse model ensures samples are best reconstructed by their own class specific dictionary.

Third, we introduce a classification loss function for the class-specific dictionary learning. The dictionaries are trained by minimizing the classification loss function. The test sample is classified into the class whose dictionary generates the minimum reconstruction error. The learned dictionaries are remarkably more discriminative.

The remainder of this paper is organized as follows. Section 2 gives a review of related approaches for action representation and sparse representation. Section 3 introduces the compound appearance and motion feature and the continuous motion segment descriptor. Section 4 presents a modified sparse model and a supervised class-specific dictionary learning method for classification. Section 5 demonstrates experimental results. Section 6 concludes this paper.

2. Related work

Over the last few years, many methods for action recognition have been presented and made impressive progress. Approaches can be categorized on the basis of action representation. There are appearance-based representation [13–15], shape-based representation [16–18], optical-flow-based representation [19–21], volume-based representation [22–24] and interest-point-based representation [1–3]. A number of approaches adopt the bag of space–time interest points [2] representation for human action recognition. This representation can be combined with either discriminative [1,25] classifiers, semi-latent topic models [26] or unsupervised generative [3,27] models. Such holistic representation of video sequences does not capture adequate time ordering and arrangement of features in the sequence. To represent actions accurately, some researchers have studied the use of temporal structures for recognizing human activities. Methods based on dynamical Bayesian networks and Markov models improve the performance but require either manual design by experts [28] or detailed training data that are expensive to collect [29]. Other work has aimed at constructing plausible temporal structures in the actions of different agents but does not consider the temporal composition within the movements of a single subject, due in part to their holistic representation. On the other hand, discriminative models of temporal context have also been applied to classification of simple motions in rather simplified environments [30,31].

In recent years, sparse representation has received a lot of attention. It approximates the input signal in terms of a sparse linear combination of the given overcomplete bases in dictionary. Such sparse representations are usually derived by linear programming as an L_1 norm minimization problem. Many efficient algorithms have been proposed to capture sparse coding features in the past several years [12,32]. A number of algorithms have also been proposed to learn dictionaries for sparse representations of signals [32,33]. The sparse representation has been successfully applied to many problems, e.g., image restoration

[34], image denoising and also well applied to classification tasks [6]. Wright et al. [35] consider the recognition problem as one of finding a sparse representation of the test image in terms of the training set as a whole, up to some sparse errors due to occlusion. The algorithm achieves impressive results on public datasets, but fails to handle practical face variations such as alignment and pose. Others try to train a compact dictionary for sparse coding, and the sparse representations of the signals are used as image features trained latter with some classifiers [36].

3. Action representation

Traditional interest-point-based representation only describes features of a single interest point, and fails to capture adequate spatial or temporal information. So it is easily influenced by noise. A novel hierarchical descriptor for action representation is introduced in this section. We present a compound appearance and motion feature, and then design a continuous motion segment descriptor based on our compound features. The compound feature considers the relationship between neighboring interest points, and describes an area around the central interest point, not a single interest point. It captures the spatial information, and it is not sensitive to noise. Furthermore, the continuous motion segment descriptor describes the time ordering information in motion. Different from previous descriptors, our proposed hierarchical descriptor incorporates more spatial and temporal information. So it is more discriminative for action representation and improves the performance of our framework.

3.1. Compound appearance and motion feature

We perform space–time interest point detection and their associated local feature extraction. To detect interest points, the method in [2] is adopted, which is a space–time extension of the Harris operator. A multi-scale approach is adopted. For the initial features, we use the histograms-of-optical-flow (HOF) and histograms-of-oriented-gradients (HOG), which characterize the motion and appearance within a volume surrounding the interest point.

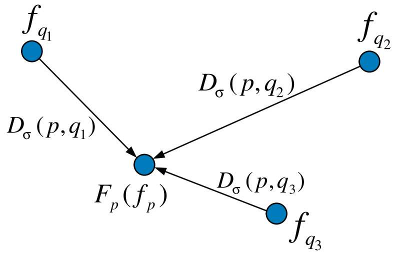

Fig. 2. The neighborhood formation of compound feature.

We specifically introduce the compound feature, which incorporates neighborhood information around the central interest point. For a given space–time point, its N closest interest points are collected, where the distance is measured by the normalized Euclidean distance on its 3D position coordinates:

$$D_\sigma(p,q) = \left(\sum_{i=1}^{3} \frac{1}{\sigma_i} (p(i)-q(i))^2 \right)^{1/2} \tag{1}$$

where $p=(x_1,y_1,t_1)$ and $q=(x_2,y_2,t_2)$ record the spatial position and the frame number of two interest points respectively. $p(i)$ is the ith dimension of vector p, and σ_i is a weight that scales the x, y or t dimension.

Let $\varphi(p)=\{p,q_1,\ldots,q_{N-1}\}$ denote the N nearest neighboring interest points for the central interest point p. The compound feature of the central point p is formed from the features of the nearest neighboring points. The contribution that a neighboring point makes to the central interest point is adaptive to the distance between the central point and the neighboring point. Therefore, the compound feature of the interest point p is defined as follows:

$$F_p = f_p + \sum_{j=1}^{N-1} \omega_j f_{q_j} \tag{2}$$

$$\omega_j = \alpha \frac{1}{D_\sigma(p,q_j)} \tag{3}$$

where F_p is the compound feature of the interest point p, f_p is the HOG and HOF features of interest point p, and f_{q_j} is the HOG and HOF features of the jth nearest interest point to the central interest point, α is a parameter, and $D_\sigma(p,q_j)$ is the distance between the central point and the jth nearest neighboring interest point. The neighborhood formation is shown in Fig. 2. Spatio-temporal words are built by applying k-means clustering to compound features of interest points. Each interest point is assigned to a closest spatio-temporal word. Compared with features of single interest point, the compound features adopt information of neighboring points. So they describe features of a larger area and they are more robust for action representation.

3.2. Continuous motion segment descriptor

Based on spatio-temporal words extracted from compound appearance and motion features, a continuous motion segment descriptor is proposed as shown in Fig. 3(a). The process includes three steps. First, we compute the histogram of spatio-temporal

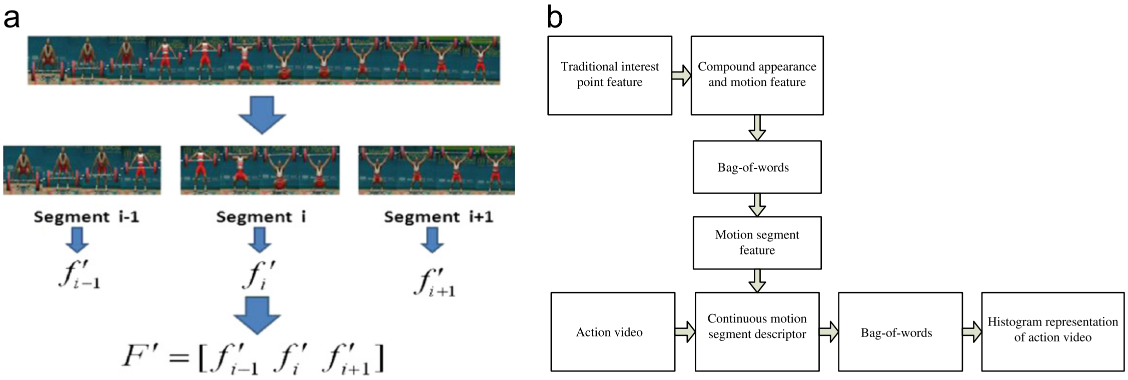
a
Segment i-1
Segment i
Segment i+1
f'_{i-1}
f'_i
f'_{i+1}
$F' = [f'_{i-1}\ f'_i\ f'_{i+1}]$
b
Traditional interest point feature
Compound appearance and motion feature
Bag-of-words
Motion segment feature
Action video
Continuous motion segment descriptor
Bag-of-words
Histogram representation of action video

Fig. 3. Action representation: (a) shows the construction of our continuous motion segment descriptor: f' is a histogram of spatio-temporal words over a segment of motion and F' is the concatenation of continuous motion segments. (b) Shows the action representation process of each action video.

words over a temporal span to represent a motion segment feature f', similar to the traditional bag-of-features approach. Second, we concatenate several continuous segment features as our continuous motion segment descriptor F' to capture temporal context information, particularly the time ordering information in the motion. The concatenation of continuous segment features can represent a continuous motion process which is discriminative. Third, we assign a label to each continuous motion segment descriptor F' by applying k-means clustering to all the continuous motion segment descriptors extracted from the video. We accumulate the occurrences of each label in the video, so each video is represented by a histogram vector. For example, we represent continuous m frames as a motion segment and concatenate n continuous motion segments as our descriptor. So the continuous motion segment descriptor is extracted from continuous $m \cdot n$ frames. From every continuous $m \cdot n$ frames in the video, we extract a continuous motion segment descriptor. Also there is not any manual annotation. Fig. 3(b) illustrates the action representation process of each action video.

The compound appearance and motion feature captures the spatial information of neighboring interest points and our continuous motion segment descriptor makes use of the time ordering information in motion. Incorporating spatial and temporal context information makes our descriptor more discriminative.

4. Sparse representation and dictionary learning

Sparse representation has been successfully applied to solve some problems in computer vision, such as image restoration, image denoising, texture synthesis and texture classification. In this section, we propose a modified sparse representation for action recognition.

Sparse representation means to represent a signal as a linear combination of a few bases of a given dictionary. Mathematically, given a signal $x \in R^n$ and a dictionary $D \in R^{n \times k}$, the sparse representation problem is stated as $\min_a ||a||_0$, s.t. $x = Da$, where $||a||_0$ is the L_0 pseudo-norm of the coefficient vector $a \in R^k$, the number of non-zero elements. As minimizing L_0 is NP-hard, a common approximation is to replace it with the L_1-norm. In the noisy case, the equality constraint must be relaxed as well. An alternative then is to solve the unconstrained problem

$$\min_a \|x - Da\|_2^2 + \lambda \|a\|_1 \tag{4}$$

where λ is a parameter that balances the tradeoff between the reconstruction error and the sparsity.

The only target of traditional sparse representation is to minimize the reconstruction error, rather than consider classification. As an improvement, each sample is locally approximated by a linear combination of its nearby samples, and the linear weights become its local coordinate coding in LCC [51] and LLC [12]. This method turns a difficult high dimensional nonlinear learning problem into a simple linear learning problem. We consider that the samples in different action classes have different features, so we adopt the class-specific dictionary. In our method, the dictionary incoherence term encourages dictionaries associated to different classes to be independent. Similar samples use similar bases and samples belonging to different classes use absolutely different bases. The reconstruction error is minimized when samples are sparsely represented by the bases in their own dictionaries. We incorporate the similarity constrained term and the dictionary incoherence term. Our proposed sparse representation algorithm is more effective for classification.

For dictionary learning, we propose a classification loss function. The target is to improve the class specific dictionaries to better reconstruct samples of their own classes than that of other classes. To minimize the classification loss function, we optimize the class specific dictionaries for more effective classification. The optimization is carried out using an iterative approach that is composed of two steps: the sparse representation step on a fixed D and the dictionary update step on fixed a.

4.1. Sparse representation and classification

Let $x_i^j, i = 1,\ldots,K, \quad j = 1,\ldots,m_i$ denotes a video representation in class i as described in Section 3 and B_i is the corresponding dictionary trained for class i. The proposed similarity-constrained and dictionary-incoherence sparse model (SDSM) is computed as

$$\min_{\substack{\{B_i, a_i^j\} \\ i = 1,\ldots,K \\ j = 1,\ldots,m_i}} \left\{ \sum_{i=1}^{K} \sum_{j=1}^{m_i} (\|x_i^j - B_i a_i^j\|_2^2 + \lambda \|d_i^j a_i^j\|_1) + \eta \sum_{p \neq q} \|B_p^T B_q\|_F^2 \right\} \tag{5}$$

where "$\cdot$" denotes the element-wise multiplication, and the notation a_i^j is the sparse code corresponding to the video descriptor $j \in [1,\ldots,m_i]$ in class i. $\|B_p^T B_q\|_F^2$ denotes the dictionary incoherence term. $\|d_i^j a_i^j\|_1$ is the similarity constrained term, and d_i^j is the similarity adapter that gives different freedom for each basis vector proportional to its similarity to the input signal x_i^j. Specifically

$$d_i^j = \exp\left(\frac{dist(x_i^j, B_i)}{\sigma} \right) \tag{6}$$

As in [12], $dist(x_i^j, B_i) = [dist(x_i^j, b_1), \ldots, dist(x_i^j, b_N)]^T$, and $dist(x_i^j, b_n)$ is the Euclidean distance between x_i^j and b_n. σ is used for adjusting the weight decay speed for the similarity adapter. Usually, we further normalize d_i^j to be between (0, 1] by subtracting $\max(dist(x_i^j, B_i))$ from $dist(x_i^j, B_i)$.

The SDSM reconstruction error is

$$\hat{R}(x_i^j, B_i) = \min_{a_i^j, B_1, \ldots, B_K} \|x_i^j - B_i a_i^j\|_2^2 + \lambda \|d_i^j a_i^j\|_1 + \eta \sum_{p \neq q} \|B_p^T B_q\|_F^2 \tag{7}$$

For classification, once the dictionaries have been learned, the class i_0 for a given new sample x is found by solving $i_0 = \text{argmin}_{i=1,\ldots,K} \hat{R}(x, B_i)$.

Fig. 4 shows the comparison between standard sparse coding (SC), locality-constrained linear coding (LLC) and our SDSM. For SC, the bases are selected from all the samples. LLC only selects bases similar to input. LLC code captures the correlations between similar inputs. SDSM not only has the properties of LLC, but also considers dictionary incoherence. The proposed sparse model selects samples that can best represent their own action class as

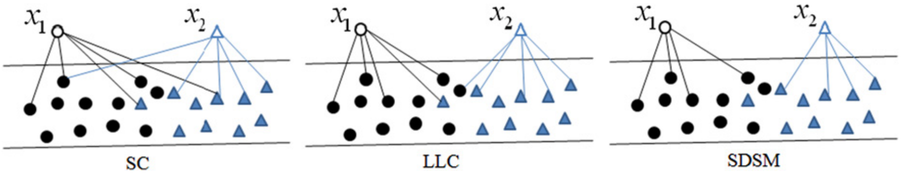

Fig. 4. Comparison between SC, LLC and SDSM. x_1 and x_2 are two inputs from different classes.

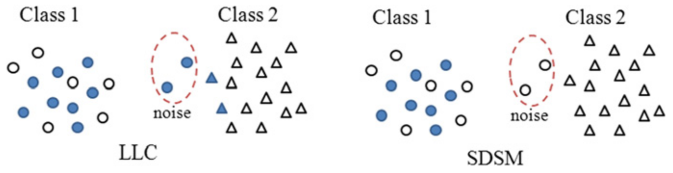

Fig. 5. The selected bases for class 1 are highlighted in blue. The two noise samples and their neighbors from class 2 are selected as the bases of class 1 by LLC. However, SDSM discards the noise successfully. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the bases of their class-specific dictionary. It does not select samples similar to other classes as bases. In the training samples, we compute the distance between the histogram representation of a video and the center of each action class. In practice, there may be complex background in some action videos. So the histogram representations of these videos may be far from the center of their own class, even they are close to the samples of other classes. We call these action videos "noise". For this reason, LLC may select samples of other classes as the bases because these selected samples belonging to other classes are similar to the noise. To overcome this problem, the dictionary incoherence term in SDSM discards the noise samples and selects the samples, which are more close to the center of their own action class than that of other classes, as the bases of their class-specific dictionary, further illustrated in Fig. 5. Different results are shown in bases selection. Obviously, the bases selected by SDSM can better represent their own action class and be more discriminative for classification than that selected by LLC.

4.2. Supervised class-specific dictionary learning

To learn a discriminative sparse model instead of pure reconstruction, we propose a supervised method to learn the dictionaries $\{B_i\}_{i=1}^K$ of sparse representation. To evaluate the effect of dictionaries for classification, we propose a classification loss function:

$$E(\{B_i\}_{i=1}^K) = \sum_{i=1}^{K}\sum_{j=1}^{m_i}\left(\hat{R}(x_i^j,B_i) - \sum_{k\neq i}\hat{R}(x_i^j,B_k)\right) \tag{8}$$

Minimizing $E(\{B_i\}_{i=1}^K)$ over B_i, the learned class specific dictionaries will better reconstruct samples of their own classes than dictionaries of other classes, and therefore, be more discriminative for classification. The dictionaries optimization is carried out using an iterative approach that is composed of two steps: the sparse coding step on a fixed $\{B_i\}_{i=1}^K$ according to our SDSM model and the dictionary update step on fixed a_i^j. We elaborate the class specific dictionary update step on fixed a_i^j as follows:

$$\begin{aligned}\frac{\partial E}{\partial B} &= \sum_{i=1}^{K}\sum_{j=1}^{m_i}\frac{\partial(\hat{R}(x_i^j,B_i) - \sum_{k\neq i}\hat{R}(x_i^j,B_k))}{\partial B}\\ &= \sum_{i=1}^{k}\sum_{j=1}^{m_i}\left(\frac{\partial\hat{R}(x_i^j,B_i)}{\partial B} - \frac{\sum_{k\neq i}\hat{R}(x_i^j,B_k)}{\partial B}\right)\end{aligned} \tag{9}$$

where $B = B_1,B_2,\ldots,B_K$.

$$\frac{\partial\hat{R}(x_i^j,B)}{\partial B_{mn}} = \lim_{\Delta B_{mn}\to 0}\frac{\hat{R}(x_i^j,B+\Delta B_{mn}) - \hat{R}(x_i^j,B)}{\Delta B_{mn}} \tag{10}$$

$$= \lim_{\Delta B_{mn}\to 0}\frac{\left\|x_i^j - (B+\Delta B_{mn})\left(a_i^j + \left(\partial a_i^j/\partial B_{mn}\right)\Delta B_{mn}\right)\right\|_2^2 - \|x_i^j - Ba_i^j\|_2^2}{\Delta B_{mn}} \tag{11}$$

where B_{mn} is the element of matrix B. Therefore, the problem is reduced to compute the gradients of the sparse representation vector a_i^j with respect to the dictionaries $\{B_i\}_{i=1}^K$.

In order to establish the relationship between a sparse code a_i^j and B_i, we first find the fixed point equations by computing the gradient with respect to a_i^j on Eq. (7) at its minimum $\hat{a}$:

$$\left.\frac{\partial(\|x_i^j - B_i a_i^j\|_2^2)}{\partial a_i^j}\right|_{a_i^j=\hat{a}} = -\lambda\left.\frac{\partial(\|d_i^j a_i^j\|_1)}{\partial a_i^j}\right|_{a_i^j=\hat{a}} \tag{12}$$

leading to

$$2(B_i^T B_i a_i^j - B_i^T x_i^j)|_{a_i^j=\hat{a}} = -\lambda d_i^j\, sign(a_i^j)|_{a_i^j=\hat{a}} \tag{13}$$

where $sign(a_i^j)$ is a vector function on each element of vector a_i^j.

In Eq. (13), a_i^j is not linked with B_i explicitly. To calculate the gradient of a_i^j with respect to B_i, we take derivative of B_i on both sides of Eq. (13):

$$\frac{\partial\{2(B_i^T B_i a_i^j - B_i^T x_i^j)\}}{\partial B_{imn}} = \frac{\partial\{-\lambda d_i^j\, sign(a_i^j)\}}{\partial B_{imn}} \tag{14}$$

The "sign" function on the right side does not continue at zero. However, since the left side of Eq. (14) cannot be infinite, $\partial\{-\lambda d_i^j\, sign(a_i^j)\}/\partial B_{imn} = 0$.

$$\frac{\partial\{2(B_i^T B_i a_i^j - B_i^T x_i^j)\}}{\partial B_{imn}} = 0 \tag{15}$$

$$\frac{\partial a_i^j}{\partial B_{imn}} = (B_i^T B_i)^{-1}\left(\frac{\partial B_i^T a_i^j}{\partial B_{imn}} - \frac{\partial B_i^T B_i}{\partial B_{imn}}a_i^j\right) \tag{16}$$

Substituting Eq. (16) into (11), $\partial\hat{R}(x_i^j,B)/\partial B_{mn}$ is solved. Then based on Eq. (9), we get the gradient of classification loss E with respect to dictionaries $B_1,B_2,\ldots,B_K$ for the class specific dictionary update.

4.3. The process of SDSM classification

For each class, we first obtain bases of the dictionary B_i by k-means clustering to all the training samples of class i for initialization. Second, we select bases from the initialization of dictionary B_i through the proposed sparse model. To regard the similarity constrained term, we only keep the set of bases with large weights. To regard the incoherence of different dictionaries, we discard the bases which are far from the center of the bases in their own action class. All the bases kept form the class-specific dictionary B_i. By the same way, we obtain all the class specific dictionaries. Third, dictionaries are updated by looping through all the training samples. The classification loss function is minimized by gradient descent method. Finally, for classification, the test sample is classified into the class whose dictionary generates the minimum reconstruction error. The process of SDSM classification is illustrated in Algorithm 1.

Algorithm 1. SDSM classification

```
Input: $B_{init} \in \Re^{Q\times H}$,   $B_{init} \in \{B_1,B_2,\ldots,B_K\}$
Output: test video $x_0 \in class\ i_0$.
Dictionary learning:
    $B \leftarrow B_{init}$
    for $i=1$ to $K$ do
        for $j=1$ to $m_i$ do
            $d_i^j \leftarrow H \times 1$ zero vector,
            for $m=1$ to $H$ do
                $d_{im}^j \leftarrow \exp^{-1}(-\|x_i^j - b_m\|^2/\sigma)$
            end
            $d_i^j \leftarrow normalize_{(0,1]}(d_i^j)$
            $a_i^j \leftarrow \arg\min_a \|x_i^j - Ba\|_2^2 + \lambda\|d_i^j \cdot a\|_1$
            $id \leftarrow \{p \,|\, |a_i^j(p)| > 0.01\}, B_i \leftarrow B(:, id)$
        end
    end
```

```
  for $i=1$ to $K$
      $C_i = \sum_{j=1}^{N_i} B_i(:,j)/N_i$, $b$ is the number of bases in b.
  end
  if $\|B_i(:,q)-C_{i' \neq i}\|_2 < \|B_i(:,q)-C_i\|_2$
      delete $Bi(:,q)$,
  end
  for $l=1$ to $L$ do
      for $n=1$ to $N$ do, $N$ is the number of training samples.
          for $i=1$ to $K$ do
          $\Delta B_i \leftarrow \frac{\partial E}{\partial B_i}$   $(from\ Eq.\ (8)\ to\ Eq.\ (16))$
          $Bi \leftarrow Bi - \mu\ \Delta Bi,\ \mu \leftarrow \sqrt{1/l}.$
          end
      end
  end
Classification:
  Dictionaries $(B_1,B_2,\ldots,B_K)$ have been learned.
  $i_0 = \arg\min_{i=1,\ldots,K} \hat{R}(x,B_i)$
```

5. Experiments

We evaluate our approach on three benchmark datasets for human action recognition: the KTH actions dataset [49], the Weizmann action recognition dataset [50], and the UCF Sports dataset [43]. All the video clips contain primarily a single action of interest. Examples of the datasets are shown in Fig. 6.

5.1. Parameters

We extract sparse Harris 3D points on the KTH and Weizmann datasets, and perform dense and multi-scale interest point extraction on the UCF Sports dataset. For the compound appearance and motion feature, we collect 6 nearest neighboring interest points. To form spatio-temporal words, we empirically set $k=300$ for the vocabulary size of KTH and Weizmann datasets, $k=3000$ for the vocabulary of UCF Sports dataset. Motions in different datasets have different temporal span. Even the motions in the same dataset also have different temporal span. Through observing all the motions in datasets, we find that some motions last for a short time range from 20 to 30 frames such as box (KTH), handclap (KTH) and pjump (Weizmann) and some motions last for a long time range from 40 to 60 frames such as bend (Weizmann), lift (UCF) and Golf Swing (UCF). In practice, we empirically represent 10 continuous frames as a motion segment and concatenate two continuous motion segments as our descriptor. In sparse representation, we use a penalty parameter $\lambda=0.1$, and set $\sigma=100$ for the similarity adapter. In dictionary learning, we set the iteration number $L=30$. On the KTH dataset, there are about 80 bases in each class-specific dictionary. The Weizmann dataset is small, and contains videos with static camera and simple background. There is almost no "noise" in this dataset. So we use all the training samples in each action class as the bases in the class-specific dictionary. On the UCF Sports dataset, the number of videos in each action class is different. So the number of bases in each class-specific dictionary is different. Because the dataset is not very large, the number of bases in each class-specific dictionary is near to the number of training samples in corresponding action class.

Fig. 6. Representative frames from videos in three datasets: Row 1 are sampled from KTH dataset, Row 2 are from Weizmann dataset, and Row 3 are from UCF dataset.

5.2. Experiments on the KTH dataset

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

Table 1 shows the average confusion matrix across all scenarios. It is seen that our approach works excellently on most actions. For example, the recognition accuracies for some actions are high up to 97%, such as "box" and "hand clap". Fig. 7 illustrates that the proposed compound appearance and motion feature improves the performance of our framework. Traditional interest point based methods only utilize features of single interest point. It can only describe a very small area. So the accuracy can easily be influenced by noise. The recognition rate is only 92.86%. However, the proposed compound feature makes full use of the information of neighboring interest points. It describes a larger area than single interest point. Obviously, the compound appearance and motion feature is more robust for action representation. The recognition rate is raised to 94.17% when we collect six nearest neighboring interest points to generate the compound feature. Table 2 shows the contribution of the proposed hierarchical descriptor. Each of

Table 1
Confusion matrix on the KTH dataset.

	Box	Handclap	Handwave	Jog	Run	Walk
Box	**0.97**	0.02	0.00	0.00	0.01	0.00
Handclap	0.00	**0.97**	0.03	0.00	0.00	0.00
Handwave	0.02	0.02	**0.96**	0.00	0.00	0.00
Jog	0.00	0.00	0.00	**0.90**	0.06	0.04
Run	0.00	0.00	0.00	0.06	**0.91**	0.03
Walk	0.01	0.00	0.00	0.04	0.01	**0.94**

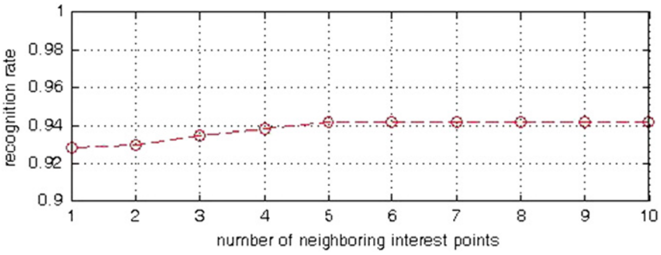

Fig. 7. Performance of compound feature.

Table 2
Contribution of proposed features.

Action feature	Accuracy (%)
Traditional single interest feature	92.18
(A) Compound feature	93.19
(B) Continuous motion segment feature	92.69
A+B	94.17

the proposed features offers more discriminative power than traditional single interest point feature, and in combination our hierarchical descriptor provides a richer representation than any single proposed feature. The hierarchical descriptor achieves the highest recognition rate. The compound feature incorporates neighborhood spatial information and our continuous motion segment descriptor captures the action temporal ordering information. Experimental results validate that the proposed hierarchical descriptor is effective for capturing the spatial and temporal information which is important for action representation.

To prove the effectiveness of the proposed sparse model and the class specific dictionary learning algorithm for classification, we make comparisons with other classifiers. The results are shown in Table 3. The performance of SVM is better than that of sparse representation without dictionary learning (WDL), and our sparse model with supervised dictionary learning (SDL) algorithm achieves the best performance. The result of WDL is comparable with that of SVM which is a powerful classifier. It proves that the proposed sparse model based on class specific dictionaries is discriminative for classification. The accuracy of SDL is higher than that of WDL and SVM. It validates that our proposed dictionary learning method is effective for boosting the recognition rate. We also compare the accuracy of our sparse model with that without the dictionary incoherence term (WDI). When we remove the dictionary incoherence term, the accuracy of WDI is lower than that of SDL. It validates that the dictionary incoherence term can improve the accuracy. Comparing the results when different lambda is chosen in the sparse model, the performance is robust as shown in Fig. 8. Experiments show that our method achieves the state-of-the-art result on the KTH dataset as shown in Table 4.

Table 3
Effectiveness of supervised dictionary learning.

	Box	Handclap	Handwave	Jog	Run	Walk	**Average (%)**
SVM	0.98	0.95	0.93	0.90	0.88	0.96	**93.33**
WDL	0.95	0.92	0.93	0.91	0.89	0.95	**92.50**
WDI	0.97	0.96	0.95	0.91	0.87	0.94	**93.33**
SDL	0.97	0.97	0.96	0.90	0.91	0.94	**94.17**

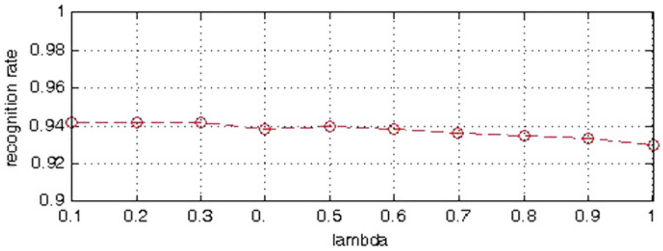

Fig. 8. The influence of lambda in the sparse model on recognition rate.

Table 4
Comparison with previous work on the KTH dataset.

Approach	Year	Accuracy (%)
Laptev et al. [1]	2008	91.80
Bregonzio et al. [38]	2009	93.17
Liu et al. [39]	2009	93.80
Gilbert et al. [40]	2009	94.50
Brendel and Todorovic [41]	2010	94.22
Kovashka et al. [53]	2010	94.53
Le et al. [42]	2011	93.90
Li et al. [54]	2011	93.60
Our method		94.17

5.3. *Experiments on the Weizmann dataset*

In order to further validate the performance of our algorithm, we also conducted experiments on the Weizmann dataset. The Weizmann action dataset contains 10 actions (bend, jumping, jack, jump forward, jump in place, jump sideways, skip, run, walk, wave with two hands, and wave with one hand) performed by 9 different subjects. This dataset contains videos with static cameras and simple background, but it provides a good test environment to evaluate the performance of the algorithm when the number of categories is larger compared with the KTH dataset (a total of six categories). In each run, eight actors' videos are used as the training set and the remaining one person's videos as the testing set. So the results are the average of nine times runs.

Table 5 shows the average confusion matrix. The recognition accuracies for some actions are high up to 100%. Table 6 illustrates the comparisons of two proposed features and their combination. Each feature offers discriminative power, and in combination our hierarchical descriptor provides a richer representation that can boost recognition rate. Similar to the performance on KTH dataset, our SDL method outperforms SVM and WDL as shown in Table 7. When we remove the dictionary incoherence term, the accuracy of WDI remains the same as that of SDL. Videos on the Weizmann dataset have little noise with static cameras and simple background. So the recognition rate is higher than that of KTH, but there are still some confused actions with small difference. Experiments show that our framework leads to the results comparable to the state-of-the-art performance as shown in Table 8.

5.4. *Experiments on the UCF sports dataset*

The authors of [43] have collected a large set of action clips from various broadcast sport videos. The actions in this dataset include diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a baseball bat, and pole vaulting. The pole vaulting sequences were removed from the original database due to copyright concerns. The videos on UCF sports dataset are captured from much more camera views. Different from the datasets above, the UCF Sports is a challenging dataset for action recognition. The actions are featured in a wide range of scenes

Table 5
Confusion matrix on the Weizmann dataset.

	Bend	Jack	Jump	PJump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	0.00	**0.89**	0.00	0.00	0.00	0.11	0.00	0.00	0.00
PJump	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	**0.89**	0.00	0.11	0.00	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.11	0.00	**0.89**	0.00	0.00	0.00
Walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**

Table 6
Contribution of proposed features.

Action feature	Accuracy (%)
Traditional single interest feature	94.5
(A) Compound feature	95.6
(B) Continuous motion segment feature	94.5
A+B	96.7

Table 7
Effectiveness of supervised dictionary learning.

	Bend	Jack	Jump	PJump	Run	Side	Skip	Walk	Wave1	Wave2	**Average (%)**
SVM	1.00	1.00	0.89	1.00	1.00	0.89	0.78	1.00	1.00	1.00	**95.6**
WDL	1.00	1.00	0.89	1.00	0.89	1.00	0.78	1.00	0.89	1.00	**94.5**
WDI	1.00	1.00	0.89	1.00	1.00	1.00	0.89	1.00	0.89	1.00	**96.7**
SDL	1.00	1.00	0.89	1.00	0.89	1.00	0.89	1.00	1.00	1.00	**96.7**

Table 8
Comparison with previous work on the Weizmann dataset.

Approach	Year	Accuracy (%)
Klaser et al. [46]	2008	84.3
Fathi and Mori [47]	2008	100
Bregonzio et al. [38]	2009	96.7
Ali and Shah [21]	2010	95.8
Seo and Milanfar [48]	2010	97.5
Our method		96.7

Table 9
Confusion matrix on the UCF Sports dataset.

	Diving	Golf swing	Kick	Lift	Ride horse	Run	Skateboard	Swing	Walk
Diving	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golf Swing	0.00	**0.77**	0.06	0.00	0.00	0.00	0.00	0.06	0.11
Kick	0.05	0.00	**0.75**	0.00	0.00	0.00	0.00	0.10	0.10
Lift	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.00	0.17
Ride Horse	0.00	0.00	0.08	0.00	**0.84**	0.00	0.00	0.00	0.08
Run	0.00	0.00	0.00	0.00	0.00	**0.77**	0.08	0.00	0.15
Skateboard	0.00	0.00	0.08	0.00	0.00	0.00	**0.92**	0.00	0.00
Swing	0.00	0.03	0.03	0.00	0.00	0.00	0.00	**0.94**	0.00
Walk	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.00	**0.90**

and viewpoints. The dataset is tested in a leave-one-out manner, cycling each example in as a test video one at a time.

Table 9 shows the average confusion matrix across all scenarios. Table 10 illustrates the comparison of two proposed features and their combination. As the experimental results on the datasets above, the two proposed features are both discriminative and in combination our hierarchical descriptor provides a richer representation that can boost the recognition rate. However, different from the results on KTH and Weizmann datasets, Table 11 illustrates that the performances of WDL and SDL are both better than that of SVM. The three methods use the same video descriptor. WDL and SDL both adopt the proposed sparse model for classification. So one possible reason is our sparse model is more effective on this dataset. Though there are many camera views for the same action, we only use the training samples from similar camera views to reconstruct the test sample. For example, as shown in Fig. 9, there are three action classes. Actions in class 1 are captured from two different camera views, and so are actions in class 2. Actions in class 3 are captured from a single camera view. The test sample is best reconstructed by the samples in action class 1 from camera view 1. So the test sample is classified into action class 1. But maybe it is difficult to find a separating surface to put all the videos belonging to the same action class captured from different camera views in the same class in SVM. So WDL and SDL both achieve better performance than SVM on this dataset. The performance of WDI is lower than that of SDL. It validates that the dictionary incoherence term can improve the accuracy. The overall mean accuracy we obtain on this dataset is 86.6%, which is comparable to the state-of-the-art performance as shown in Table 12. Even in the challenging and realistic action dataset, our method also performs reliable recognition rate. It further indicates that the proposed hierarchical descriptor is discriminative and our sparse model is effective for classification.

Table 10
Contribution of proposed features.

Action feature	Accuracy (%)
Traditional single interest feature	79.3
(A) Compound feature	81.3
(B)Continuous motion segment feature	82.7
A+B	86.6

Table 11
Effectiveness of supervised dictionary learning.

	Diving	Golf swing	Kick	Lift	Ride horse	Run	Skate board	Swing	Walk	**Average (%)**
SVM	1.00	0.67	0.70	0.67	0.68	0.70	0.76	0.88	0.85	**78.0**
WDL	1.00	0.72	0.65	0.83	0.68	0.77	0.84	0.85	0.85	**80.0**
WDI	1.00	0.72	0.70	0.83	0.76	0.77	0.84	0.91	0.90	**83.3**
SDL	1.00	0.77	0.75	0.83	0.84	0.77	0.92	0.94	0.90	**86.6**

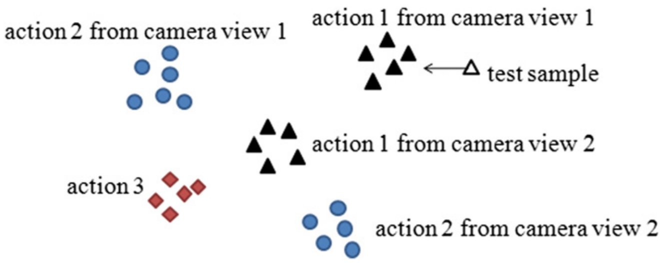
action 2 from camera view 1
action 1 from camera view 1
test sample
action 1 from camera view 2
action 3
action 2 from camera view 2

Fig. 9. The test sample is best reconstructed by samples in the same action class from similar camera view.

Table 12
Comparison with previous work on the UCF Sports dataset.

Approach	Year	Accuracy (%)
Rodriguez et al. [43]	2008	69.2
Yeffet and Wolf [44]	2009	79.2
Wang et al. [45]	2009	85.6
Yao et al. [52]	2010	86.6
Kovashka and Grauman [53]	2010	87.3
Le et al. [42]	2011	86.5
Our method		86.6

6. Conclusions

In this paper, we have presented a novel method for action representation based on compound features and continuous motion segments. Our descriptor incorporates spatial and temporal information which represents actions more accurately. We have also introduced a supervised classification based on class specific sparse representation and dictionary learning. We have proposed a classification loss function for the class specific dictionary learning as well. Our framework achieves comparable performance on the datasets above. The experiments have validated that our proposed hierarchical descriptor is discriminative, and the proposed sparse model incorporating supervised dictionary learning is effective for classification.

Acknowledgment

This work was carried out at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. It was supported in part by the NSFC (Grant nos. 60825204, 60935002, 61100099 and 61005008) and the National 863 High-Tech $R\& D$ Program of China (Grant no. 2009AA01Z318).

References

[1] 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.

[2] I. Laptev, On space–time interest points, International Journal of Computer Vision 64 (2) (2005) 107–123.

[3] 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.

[4] A. Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in: Proceedings of the IEEE International Conference on Computer Vision, 2010.

[5] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, Advances in Neural Information Processing Systems (2007).

[6] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[7] Ignacio Ramirez, Pablo Sprechmann, Guillermo Sapiro, Classification and clustering via dictionary learning with structured incoherence and shared features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[8] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (1) (2010) 331–341.

[9] M. Fan, N. Gu, H. Qiao, B. Zhang, Sparse regularization for semi-supervised classification, Pattern Recognition 44 (8) (2011) 1777–1784.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Supervised dictionary learning, Advances in Neural Information Processing Systems (2008).

[11] D.M. Bradley, J.A. Bagnell, Differential sparse coding, Advances in Neural Information Processing Systems (2008).

[12] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, Yihong Gong, Locality-constrained linear coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[13] 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.

[14] 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.

[15] T. Starner, A. Pentland, Visual recognition of American sign language using hidden Markov model, in: International Workshop on Automatic Face and Gesture Recognition, 1995.

[16] J. Zhang, S. Gong, Action categorization with modified hidden conditional random field, Pattern Recognition 43 (1) (2010) 197–203.

[17] S. Xiang, F. Nie, Y. Song, C. Zhang, Contour graph based human tracking and action sequence recognition, Pattern Recognition 41 (12) (2008) 3653–3664.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (12) (2007) 2247–2253.

[19] 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) 1257–1265.

[20] 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.

[21] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 288–303.

[22] T. Mahmood, A. Vasilescu, S. Sethi, Recognition of action events from multiple video points, in: Proceedings of the IEEE International Conference on Computer Vision, 2001.

[23] 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.

[24] P. Scovanner, S. Ali, M. Shah, A 3-dimensional SIFT descriptor and its application to action recognition, in: Proceedings of the International Conference on Multimedia, 2007.

[25] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[26] Y. Wang, G. Mori, Human action recognition by semilatent topic models, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1762–1774.

[27] S.F. Wong, T.K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[28] B. Laxton, J. Lim, D. Kriegman, Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[29] N. Ikizler, D.A. Forsyth, Searching for complex human activities with no visual examples, International Journal of Computer Vision 80 (3) (2008) 337–357.

[30] C. Sminchisescu, A. Kanaujia, D. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding 104 (2) (2006) 1808–1815.

[31] A. Quattoni, S.B. Wang, L.P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1848–1852.

[32] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, Advances in Neural Information Processing Systems (2006).

[33] M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.

[34] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Transactions on Image Processing 17 (1) (2008) 53–69.

[35] J. Wright, A. Yang, A. Ganesh, S. Satry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[36] J. Yang, K. Yu, T. Huang, Supervised translation-invariant sparse coding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[38] 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.

[39] 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.

[40] 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, 2010.

[41] W. Brendel, S. Todorovic, Activities as time series of human postures, in: Proceedings of European Conference on Computer Vision, 2010.

[42] Q. Le, W. Zou, S. Yeung, A. 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.

[43] 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.

[44] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.

[45] 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.

[46] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in: Proceedings of the British Machine Vision Conference, 2008.

[47] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[48] H.J. Seo, P. Milanfar, Action recognition from one example, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (5) (2010) 867–882.

[49] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of International Conference on Pattern Recognition, 2004.

[50] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[51] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, Advances in Neural Information Processing Systems (2009).

[52] A. Yao, J. Gall, L. Van Gool, A Hough transform-based voting framework for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[53] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space–time neighborhood features for human action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[54] 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.

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.

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.

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 of Nature Science of Ministry of Education, China. He has published more than 40 papers. Professor Sun is a member of an 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.