Computers, Materials & Continua CMC, vol.63, no.1, pp.243-262, 2020

Human Action Recognition Based on Supervised Class-Specific Dictionary Learning with Deep Convolutional Neural Network Features

Binjie Gu[1, *], Weili Xiong[1] and Zhonghu Bai[2]

Abstract: Human action recognition under complex environment is a challenging work. Recently, sparse representation has achieved excellent results of dealing with human action recognition problem under different conditions. The main idea of sparse representation classification is to construct a general classification scheme where the training samples of each class can be considered as the dictionary to express the query class, and the minimal reconstruction error indicates its corresponding class. However, how to learn a discriminative dictionary is still a difficult work. In this work, we make two contributions. First, we build a new and robust human action recognition framework by combining one modified sparse classification model and deep convolutional neural network (CNN) features. Secondly, we construct a novel classification model which consists of the representation-constrained term and the coefficients incoherence term. Experimental results on benchmark datasets show that our modified model can obtain competitive results in comparison to other state-of-the-art models.

Keywords: Action recognition, deep CNN features, sparse model, supervised dictionary learning.

1 Introduction

In recent years, human action recognition has been successfully applied in scenarios such as smart home, intelligent video surveillance, public security and so on. Due to large differences in human action types, such as gesture and height, human action recognition is still difficult.

Action representation as a key issue will greatly influence the classification performance of human action recognition. Motion and appearance information as low-level features are the main cues embedded in action video sequence. In early human action recognition, the spatiotemporal representation methods have demonstrated their superiorities. Recently, researchers reveal that the salient low-level features are one of the key issues in the field of image processing [Hou, Li, Wang et al. (2018)] and pattern recognition [Bian,

[1] Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi, China.

[2] National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi, China.

[*] Corresponding Author: Binjie Gu. Email: gubinjie1980@jiangnan.edu.cn.

Received: 09 April 2019; Accepted: 13 June 2019.

Tao and Rui (2012); Lu, Fang, Shao et al. (2012); Fahad, Joost, Rao et al. (2018)]. For example, the weber's law states that the change in a stimulus that will be just discriminable is a constant ratio of the original stimulus. Hence, the weber local descriptor (WLD) can be utilized to represent the characteristics of the local area [Chen, Shan, He et al. (2010); Wang, Li, Yang et al. (2011); Wang, Yuan, Hu et al. (2012); Li, Gong and Yuan (2013)]. To date, WLD has been successfully used in object recognition domain [Kong and Wang (2012); Zhou, Shen, Peng et al. (2012)].

Suitable feature representation is always useful in intelligent video surveillance domains. Ali et al. [Ali and Shah (2010)] explored the utility of kinematic features derived from motion information for human action recognition in videos. Junejo et al. [Junejo, Dexter, Laptev et al. (2010)] explored the temporal self-similarities of action sequences over time to address human action recognition under different view changes. Fanello et al. [Fanello, Gori, Metta et al. (2013)] designed an effective real-time system for one-shot action modeling and recognition, they obtained very good results on benchmark datasets and human-robot interaction setting. Zhang et al. [Zhang, Xu, Shi et al. (2015)] proposed a robust spatiotemporal saliency algorithm for action recognition. Caetano et al. [Caetano, Santos and Schwartz (2016)] proposed a new spatiotemporal feature descriptor based on co-occurrence matrices, and this feature extraction method proved to be discriminative in some action recognition datasets. Cherian et al. [Cherian, Fernando, Harandi et al. (2017)] developed generalized rank pooling to summarize the action dynamics in video sequences. The extensive experiments on action recognition datasets demonstrated the advantages of the proposed schemes.

Local representation methods, holistic representation methods, and machine learning methods are the most commonly used human action recognition approaches. Local representation methods are mainly based on the spatiotemporal interest points (STIP) [Laptev (2005); Sapuppo, Umana, Frasca et al. (2006); Oikonomopoulos, Patras and Pantic (2006); Shao, Zhen, Liu et al. (2011); Hara, Kataoka and Satoh (2017)]. Among these approaches, the model of bag of words (BOW) and its variants proved to be very successful. This is because that the BOW model is insensitive to partial occlusion and abandons some preprocessing steps. However, BOW representation suffers two defects in describing a behavior. First, in this model, one word denotes one interest point, which will lead to a big reconstruction error. In addition, the closest word type completely determines the location of corresponding interest point, which will lead to the situation that different points of interest may belong to the same type. Holistic representation methods consider the raw video sequences as a volume in space/time and extract spatiotemporal features from this volume directly rather than detecting spatiotemporal interest points using STIP detectors. However, holistic representation methods depend on accurate location, background extraction and tracking. Furthermore, preprocessing steps such as accurate localization and tracking are often necessary [Minhas, Mohanmmed and Wu (2012); Zhen, Shao, Tao et al. (2013)]. Machine learning methods are widely used in computer vision due to their powerful abilities to handle large-scale training data. For many computer vision recognition tasks, deep learning methods have shown their superiorities. However, traditional convolutional neural network (CNN) is not always suitable due to its poor generalization ability. In order to obtain better features, Wen et al. [Wen, Zhang, Li et al. (2016)] proposed a new learning regulation, named center loss.

The main idea of center loss is that the feature of each sample in a batch is the square sum of the distance from the center of feature, the smaller the better. It is encouraging to see that their CNNs have achieved rather competitive results. Therefore, in our work, we adapt this kind of network model to obtain robust action feature.

In this paper, a novel sparse classification model for conducting action analysis is proposed. The framework of our model is plotted in Fig. 1. First, we extract the deep CNN features from the input samples. Then, the reduced low-level descriptors are transformed into mid-level features by sparse coding. To obtain a highly discriminative representation of the extracted features, we explored an improved class-specific dictionary learning approach over the whole sparse code set. Finally, a sparse representation classifier is well designed for classification.

The main contributions of this paper are summarized as follows: (1) We propose a new and robust human action recognition algorithm by integrating sparse representation-based classification model with deep CNN features. The features obtained by the deep convolutional network proved to be very robust; (2) A novel sparse model is proposed to achieve more effective recognition. In our designed model, to strengthen the learning ability of the dictionary, two terms are combined to keep discriminative ability, named as the representation-constrained term and the coefficients incoherence term.

The rest of our work is organized as follows. Section 2 briefly presents related work. Section 3 introduces our proposed sparse representation model. Section 4 demonstrates the robustness of our model. Section 5 gives the conclusion.

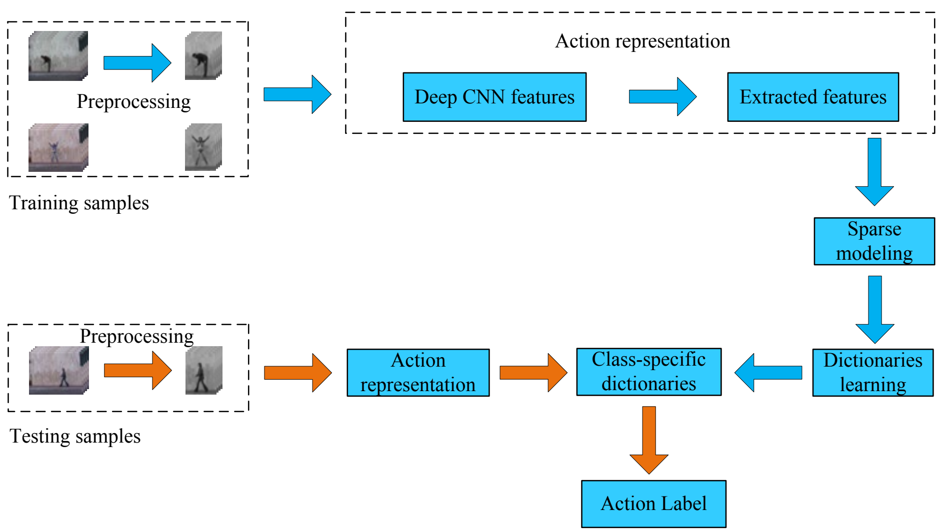


Figure 1: The framework of the proposed model

2 Related work

Learning dictionaries for sparse coding is a key factor to achieve a high recognition rate. The existing dictionary learning works can be classified into shared dictionary, class-specific dictionary, and hybrid dictionary learning.

2.1 Shared dictionary learning

The shared dictionary learning algorithm aims to learn a shared dictionary. Marial et al. [Mairal, Bach, Ponce et al. (2008)] presented a robust and discriminative dictionary learning algorithm. Inspired by K-SVD, a more discriminative K-SVD was proposed to learn a joint dictionary [Zhang and Li (2010)]. Then, Mairal et al. [Mairal, Bach and Ponce (2012)] designed a task-driven learning scheme to minimize corresponding loss functions. Based on the discriminative K-SVD, Jiang et al. [Jiang, Lin and Davis (2013)] introduced a label consistent term to enhance the discriminative power of dictionary. In general, in these schemes, a shared and discriminative dictionary can be learned simultaneously. However, shared dictionary learning algorithm does not consider the effect of a class label on a dictionary atom.

2.2 Class-specific dictionary learning

The class-specific dictionary learning scheme mainly considers the class labels information and reflects the relationship between the dictionary atoms and the class labels [Ramirez, Sprechmann and Sapiro (2010); Yang, Zhang, Feng et al. (2011); Castrodad and Sapiro (2012)]. Wang et al. [Wang, Yuan, Hu et al. (2012)] proposed a modified sparse model by minimizing two constrained terms. To ensure that the dictionaries from different kinds are independent, Ramirez et al. [Ramirez, Sprechmann and Sapiro (2010)] constructed a constraint term to improve learning power. In general, in these schemes, a class-specific dictionary can be learned by adding appropriate penalty and constraint term [Mairal, Bach, Ponce et al. (2008); Yang, Zhang, Feng et al. (2011); Castrodad and Sapiro (2012)]. Therefore, this kind of method owns a broad application prospect.

2.3 Hybrid dictionary learning

The hybrid dictionary model mainly considers the relationship among different dictionary atoms. Kong et al. [Kong and Wang (2012)] introduced a coherence penalty term in their proposed model to obtain good classification ability. Shen et al. [Shen, Wang, Sun et al. (2013)] proposed to build hierarchical category structure to obtain better performance. Yang et al. [Yang, Zhang, Feng et al. (2014)] used a fisher penalty term to improve the discriminant model.

Although the above-mentioned methods could improve dictionary learning efficiency, learning a discriminative and representative dictionary for classification is still a challenging task.

3 Proposed novel sparse representation model

3.1 Modelling

In class-specific dictionary learning, the atoms of class labels in the learned dictionary D are represented as $[D_1, \cdots, D_K]$, where $D_i, i = 1, \cdots, K$ represents the sub-dictionary in class i. Once the representation vector $\widehat{\alpha} = [\widehat{\alpha}_1, \cdots, \widehat{\alpha}_K]$ is calculated, the corresponding representation residual $\|y - D_i\hat{\alpha}_i\|_2$ could be used for classification, where y denotes a query sample, $\hat{\alpha}_i, i = 1, \cdots, K$ is the sub-vector associated with class i. Let $a_{i,j}$, $i =$

$1, \cdots, K, j = 1, \cdots, n_i$ denotes a training sample that is introduced by deep features in class i, we can form the training sample set $A_i = [a_{i,1}, \cdots, a_{i,n_i}]$, $i = 1, \cdots, K$.

Then, we can learn the dictionary D from the following extended sparse model:

$$\langle D, Z\rangle = \operatorname{argmin}_{D,Z} \sum_{i=1}^{K} \Big\{\|A_i - DZ_i\|_F^2 + \lambda_1\|Z_i\|_1 + \lambda_2\left\|A_i - D_i Z_i^i\right\|_F^2 + k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2\Big\}$$
$$\text{s.t.}\ \ \|d_n\|_2 = 1, \forall n \tag{1}$$

where, Z is the sparse code, Z_i is the sub-matrix with respect to A_i over D, λ_1, λ_2 and k are weight factors.

Once Eq. (1) is solved, we can obtain $Z_i = [Z_i^1; \cdots; Z_i^j; \cdots; Z_i^K]$, where Z_i^j indicates the coefficients of A_i on D_j; $\tilde{Z}_j = [\tilde{z}_{j,1}, \cdots, \tilde{z}_{j,n_j}]$ where $\tilde{z}_{j,i} = z_{j,i}/\|z_{j,i}\|$ is normalized coefficients and there are n_j samples in A_j.

Different from the traditional sparse representation-based classification (SRC) model [Wright, Yang, Ganesh et al. (2009)], we introduce the representation-constrained term $\lambda_2\left\|A_i - D_i Z_i^i\right\|_F^2$ and the coefficients incoherence term $k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2$ in Eq. (1).

3.1.1 The representation-constraint term

For A_i from class i, we have $A_i \approx DZ_i$. Since A_i and class i are closely related, it is naturally possible that A_i can be represented by only using D_i. This indicates that there should be appropriate representation Z_i^i in Z_i so that $\left\|A_i - D_i Z_i^i\right\|_F^2$ is small enough.

3.1.2 The coefficients incoherent term

Wright et al. [Wright, Yang, Ganesh et al. (2009)] found that the largest coefficients in the SRC model are associated with the training samples, which have the same class labels as the test samples. This means that we can reconstruct the test samples by a linear weighted combination of its own training samples with their corresponding largest coefficients. Similarly, the largest coefficients of A_i are expected to be related to the sub-dictionary D_i. Therefore, in Eq. (1), minimizing the coefficients incoherence term $k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2$ can ensure that different dictionaries are independent of each other. This also means that the training samples of the same class will have similar coefficients vector over the learned dictionary D.

Overall, minimizing the representation-constrained term $\left\|A_i - D_i Z_i^i\right\|_F^2$ can ensure that the learned sub-dictionary has powerful reconstruction ability for the training samples, and minimizing the coefficients incoherence term $k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2$ can encourage that for A_i and A_j, the largest coefficients are associated with their corresponding different sub-dictionaries (i.e., D_i and D_j, respectively), as shown in Fig. 2.

In Fig. 2, the training samples marked with black and purple colors are coming from different class i and class j; the black and purple atoms in the learned dictionary D have different class label i and class label j; the sparse coefficients of the training samples marked with black and purple colors are plotted in the coefficients matrix with their corresponding largest values associated with the black and purple atoms in D.

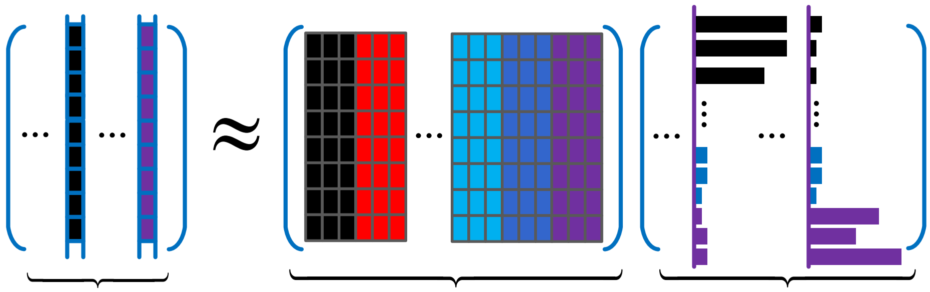


Figure 2: Sparse representation illustration of the training samples over the learned dictionary D

Therefore, these corresponding largest coefficients D_i will not only ensure the minimum but also have rather small representation residual for A_i. Instead, without these largest coefficients, other sub-dictionaries will have big representation residuals of A_i. This also means that, over the learned dictionary D, the training samples which belong to the same class will have similar coefficients vector. Conversely, the training samples which belong to different classes will have completely different coefficients vector. Hence, the value of the object function in Eq. (1) will be minimized if the training samples can be sparsely represented by the dictionary atoms in their own sub-dictionaries. In summary, by combining the two introduced constraint terms, our modified model is expected to be more effective for classification.

3.2 Supervised class-specific dictionary learning

The object function in Eq. (1) is not convex to (D, X), but when the other one is fixed, it is convex for D and X. Therefore, by alternatively optimizing D and X, we can easily obtain the optimal solution of the objective function in Eq. (1).

3.2.1 Update of Z

Once the dictionary D is fixed, Eq. (1) can be treated as a sparse representation problem. This means that $Z = [Z_1, \cdots, Z_K]$ can be easily computed and $Z_i, i = 1, \cdots, K$ can be computed class by class. Note that all $Z_j, j \neq i$, are fixed when computing Z_i.

Hence, the objective function in Eq. (1) can be changed into the following form:

$$\min_{Z_i}\left\{\|A_i - DZ_i\|_F^2 + \lambda_1\|Z_i\|_1 + \lambda_2\left\|A_i - D_i Z_i^i\right\|_F^2 + k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2\right\} \tag{2}$$

Eq. (2) can be rewritten as:

$$\min_{Z_i}\{\varphi_i(Z_i) + \lambda_1\|Z_i\|_1\} \tag{3}$$

where $\varphi_i(Z_i) = \|A_i - DZ_i\|_F^2 + \lambda_2\left\|A_i - D_i Z_i^i\right\|_F^2 + k\sum_{j\neq i}\left\|\tilde{Z}_j^T Z_i\right\|_F^2$.

It can be proved that $\varphi_i(Z_i)$ is convex with Lipschitz continuous gradient. The detailed proof is omitted here. The fast iterative shrinkage-thresholding algorithm (FISTA) [Beck and Teboulle (2009)] is utilized to solve Eq. (3), as shown in Tab. 1.

Table 1: Learning sparse code Z_i

Algorithm of obtaining sparse codes $\boldsymbol{Z_i}$
1. Input: the training sample set A_i with label i; D denotes dictionary; the parameters $\rho, \tau > 0$.
2. Initialization: $\hat{Z}_i^{(1)} \leftarrow 0$ and $t \leftarrow 1$.
3. do
$t \leftarrow t+1$
$u^{(t-1)} \leftarrow \hat{Z}_i^{(t-1)} - {}^1/_{2\rho} \nabla\varphi_i\left(\hat{Z}_i^{(t-1)}\right)$, where $\nabla\varphi_i\left(\hat{Z}_i^{(t-1)}\right)$ is the derivative of $\varphi_i\left(\hat{Z}_i^{(t-1)}\right)$ w.r.t. $\hat{Z}_i^{(t-1)}$.
$\hat{Z}_i^{(t)} \leftarrow soft(u^{(t-1)}, {}^\tau/_\rho)$, where $soft(u, {}^\tau/_\rho)$ is defined in [Wright, Nowak and Figueiredo (2009)]:
$\left[soft(u, {}^\tau/_\rho)\right]_j = \begin{cases} 0, & \lvert u_j \rvert \le {}^\tau/_\rho \\ u_j - sign(u_j)\,{}^\tau/_\rho, & otherwise \end{cases}$
while convergence or the predefined iterations are not reached
4. Output: $\hat{Z}_i = Z_i^{(t)}$.

3.2.1 Update of D

Similarly, the dictionary $D = [D_1, \cdots, D_K]$ is updated when Z is fixed. $D_i = [d_1, \cdots, d_{p_i}]$ is updated class by class. Note that when D_i is updating, all $D_j, j \neq i$ are fixed. Therefore, the objective function in Eq. (1) can be reduced as:

$$\min_{D_i}\left\{\left\|\bar{A} - D_i Z^i\right\|_F^2 + \lambda_2\left\|A_i - D_i Z_i^i\right\|_F^2\right\}, s.t.\ \|d_l\|_2 = 1, l = 1, \cdots, p_i \tag{4}$$

where $\bar{A} = A - \sum_{j=1, j\neq i}^{K} D_j Z^j$; Z^i represents the coefficient matrix of A on D_i.

Eq. (4) can be rewritten as:

$$\min_{D_i}\left\|\bar{\bar{A}}_i - D_i X_i\right\|_F^2\ ,\ s.t.\ \|d_l\|_2 = 1, l = 1, \cdots, p_i \tag{5}$$

where $\bar{\bar{A}}_i = [\bar{A}A_i]$, $X_i = \left[Z^i Z_i^i\right]$. We can solve Eq. (4) by the dictionary learning algorithm proposed by Yang et al. [Yang, Zhang, Yang et al. (2010)], as described in Tab. 2.

Table 2: Learning dictionary D_i

Algorithm of obtaining sparse codes $\boldsymbol{D_i}$
1. Input: the training subset A_i with class i; the initial dictionary D_i; the coefficients X_i.
2. Let $X_i = \left[x_1; \cdots; x_{p_i}\right]$ and $D_i = \left[d_1, \cdots, d_{p_i}\right]$, where $x_j, j = 1, \cdots, p_i$, is the j^{th} vector of x_i and $d_j, j = 1, \cdots, p_i$, is the j^{th} vector of D_i.
3. For $j = 1$ to p_i **do**
Fix $d_l, l \neq j$ while update d_j. Let $Y = \bar{\bar{A}}_i - \sum_{l\neq j} d_l x_l$. The minimization of Eq. (5) becomes: $\min_{d_j}\left\|Y - d_j x_j\right\|_F^2$ s.t. $\left\|d_j\right\|_2 = 1$;
Using the method proposed by Yang et al. [Yang, Zhang, Yang et al. (2010)], we could obtain the solution $d_j = Y x_j^T / \left\|Y x_j^T\right\|_2$.
4. Output: updated D_i.

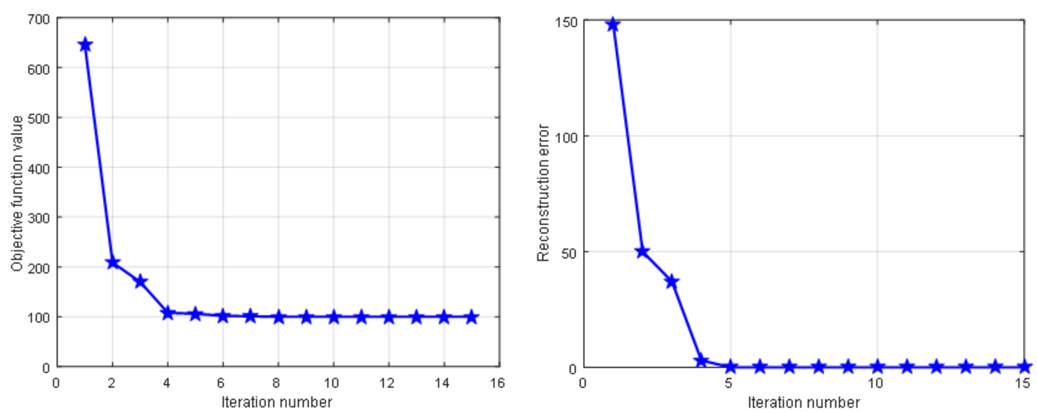
3.2.3 The whole dictionary D learning algorithm

The whole algorithm of dictionary learning can be found in Tab. 3.

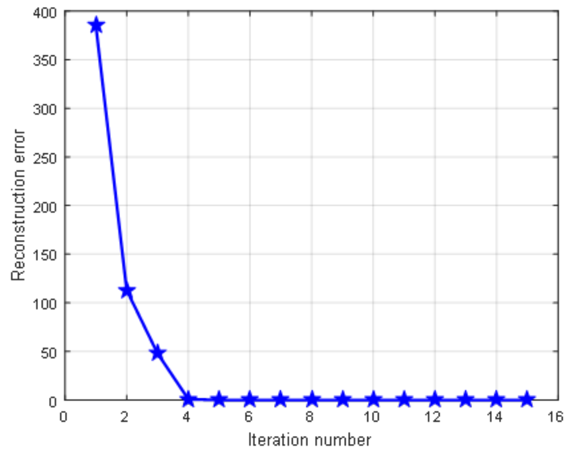
Table 3: The whole algorithm of dictionary D learning

The whole algorithm of obtaining dictionary *D*
1. Initialize *D*.   Initialize the D_i with the eigenvectors of A_i.
2. Update coefficients *Z*.   Fix $\boldsymbol{D}$ and compute $Z_i, i = 1, \cdots, K,$ solve Eq. (1) using the method described in Tab. 1.
3. Update dictionary *D*.   Fix $\boldsymbol{Z}$ and update each $D_i, i = 1, \cdots, K$, solve Eq. (4) using the method described in Tab. 2.
4. Go to step 2 until the value of the objective function is small enough.
5. Output: $\boldsymbol{Z}$ and $\boldsymbol{D}$.

Fig. 3 illustrates the minimization process on the Weizmann dataset [Gorelick, Blank, Shechtman et al. (2007)]. Fig. 3(a) presents the convergence process of Eq. (1). Fig. 3(b) plots the curve of $\sum_{i=1}^{K} \lambda_2 \left\| A_i - D_i Z_i^i \right\|_F^2$, showing that D_i represents A_i well. Because Z_i represents the sparse coefficients of A_i over dictionary D, so $A_i \approx D Z_i$. Z_i can be well represented by only A_i in class i because Z_i is related to class i, which is in natural expectation. Therefore, there should exist a Z_i^i which makes $\left\| A_i - D_i Z_i^i \right\|_F^2$ small enough. This term is able to keep the reconstruction error of coefficients Z_i^i under control. In addition, λ_2 is the scalar controlling the relative contribution of the corresponding term, so we can control the reconstruction error by adjusting λ_2. Fig. 3(c) plots the curve of $\sum_{i=1}^{K} k \sum_{j \neq i} \left\| \tilde{Z}_j^T Z_i \right\|_F^2$, showing that the coefficients of different training samples are related with the corresponding sub-dictionaries.



(a) (b)



(c)

Figure 3: The minimization process on the Weizmann dataset. (a) The iteration process of the proposed sparse model; (b) The curve of $\sum_{i=1}^{K} \lambda_2 \left\| A_i - D_i Z_i^i \right\|_F^2$ *vs.* the iteration number; (c) The curve of $\sum_{i=1}^{K} k \sum_{j \neq i} \left\| \tilde{Z}_j^{\mathrm{T}} Z_i \right\|_F^2$ *vs.* the iteration number

3.3 The classification scheme

When training, the dictionary D can be used to denote the query sample y and complete the classification task. On the basis of different ways to learn the dictionary D, we can use different information to carry out the classification task.

From the sparse classification model, we can learn the dictionary D from the training dataset A. Thus, we propose the following model:

$$\hat{\alpha} = \mathrm{argmin}_{\alpha}\{\|y - D\alpha\|_2^2 + \gamma\|\alpha\|_1\} \tag{6}$$

where γ is a fixed value.

Denote $\hat{\alpha} = [\hat{\alpha}^1, \cdots, \hat{\alpha}^K]^{\mathrm{T}}$, where $\hat{\alpha}^i$ is the coefficient sub-vector associated with sub-dictionary D_i. During the learning stage, we have enforced the class-specific dictionary learning algorithm. Therefore, if y belongs to class i, the term $\left\| y - D_i \hat{\alpha}^i \right\|_2^2$ may be small, while the term $\left\| y - D_j \hat{\alpha}^j \right\|_2^2, j \neq i$, is a big value. Finally, taking the discriminative ability of two added terms into account, the following metric can be defined to classify:

$$e_i = \left\| y - D_i \hat{\alpha}^i \right\|_2^2 + w \sum_{j \neq i} \left\| \tilde{Z}_j^T \hat{\alpha} \right\|_F^2 / n_j \tag{7}$$

where w is a preset value. The classification can be completed by setting $\mathrm{identity}(y) = \mathrm{argmin}_i\{e_i\}$.

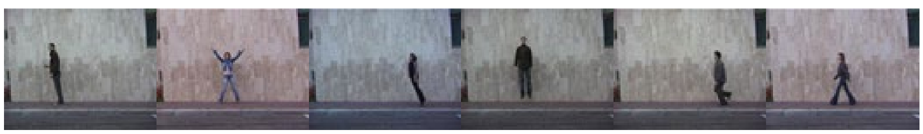
4 Experimental results and discussions

In order to evaluate the performance of our sparse model, and to distinguish the effectiveness of our sparse model in terms of recognition accuracy, we conduct a detailed experimental research on benchmark datasets. The benchmark datasets used in our experiments are Weizmann dataset [Gorelick, Blank M, Shechtman et al. (2007)], KTH dataset [Kim and Cipolla (2009)], and Cambridge-Gesture dataset [Dalal and Triggs (2005)]. We illustrate some human action examples of the three benchmark datasets in Fig. 4.

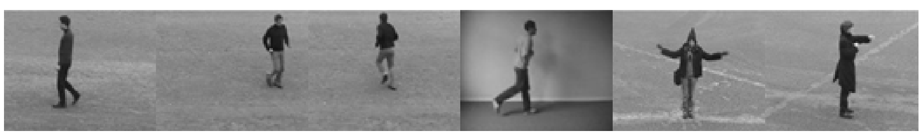
4.1 Parameter settings

In the all datasets, bounding boxes can be obtained from the method in Caetano et al. [Caetano, Santos and Schwartz (2016)], and trackers comes from the method in Fanello et al. [Fanello, Gori, Metta et al. (2013)], and the size of bounding boxes is set as $M \times N$ (M=80, N=64). Then, we adjust all sequences to 32 frames for all datasets that is similar to the scheme in Ali et al. [Ali and Shah (2010)]. To evaluate the classification accuracy, we employ the 5-fold cross-validation test on each dataset. The results are averaged under 10 independent trials.

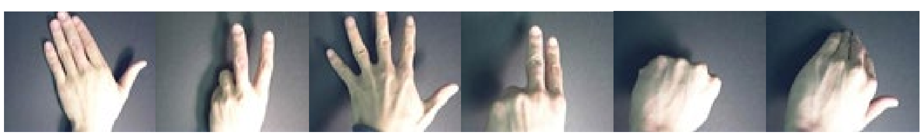
Our model includes two stages. The former stage is dictionary learning, the latter is classification. In the former stage, we set $\lambda_1 = 0.005$, $\lambda_2 = 1$, $k = 0.01$; while in the latter stage, we set $\gamma = 0.01$, $w = 0.05$.



(a)



(b)



(c)

Figure 4: Examples of three benchmark datasets: (a) The Weizmann dataset; (b) The KTH dataset; (c) The Cambridge-Gesture dataset

4.2 Experiments on the Weizmann dataset

Weizmann dataset consists of 93 video clips coming from nine different cases, and it has different forms of actions like one-hand-waving (Wave1), two-hands-waving (Wave2), galloping sideways (Side), walking (Walk), running (Run), bending (Bend), jumping (Jump), jumping jack (JumpJ), jumping in place (JumpP). The camera is fixed, and the background is simple and there is no occlusion of actions. Some action examples are given in Fig. 4(a).

Eight subjects are used for training, and the remaining subjects are used for testing. The average confusion matrix of nine different actions is presented in Fig. 5. Fig. 5 shows that our sparse model performs well. For instance, the recognition rate of some actions is absolutely 100%, such as “Bend” and “Wave2”. While for other actions, for example “Side”, “Run” and “Walk”, they are relatively complex. Unfortunately, the recognition accuracy of “Side” falls to 89%. A possible reason is that this behavior relies heavily on contextual information.

The accuracy of three different action recognition descriptors, i.e., HOG (histogram of gradients) and HOF (histogram of optical flow), HWOM (histograms of weber orientation magnitude) and HOF, Deep CNN descriptors are presented in Tab. 4. To make the comparison fair, all the used features are integrated with the sparse model that we proposed for the following action classification. We can see from Tab. 4 that each of the action features offers discriminative ability for action classification. The HWOM and HOF descriptor outperforms the traditional HOG and HOF descriptor, the reason is that the form descriptor of HWOM and HOF descriptor is constructed on the basis of the original WLD map. The Deep CNN descriptor achieves the best accuracy (marked with bold font). Therefore, the sparse model that we proposed is more effective for action classification than others.

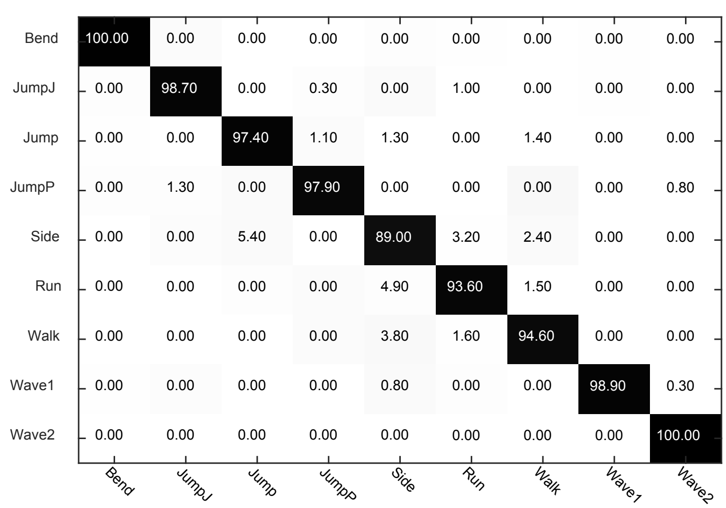


Figure 5: Average confusion matrix of different actions on the Weizmann dataset

Table 4: Accuracy of different descriptors on the Weizmann dataset

Descriptors	**Accuracy (%)**
HOG and HOF	90.2
HWOM and HOF	92.0
Deep CNN	**96.7**

To further evaluate the performance of our model, we compare the accuracy of different classifiers on the Weizmann dataset, and the results are plotted in Fig. 6. It can be seen from Fig. 6 that our model has stronger discriminative ability than the traditional SVM and SRC. When neither the representation-constrained term nor the coefficients incoherence term is removed, the recognition rate will decrease slightly (Note that WDR and WDCI stands for the sparse model without the dictionary representation-constrained term and the dictionary coefficients incoherence term, respectively). Tab. 5 lists the results of recognition accuracy using different methods on the Weizmann dataset, which further validates the effectiveness of our method.

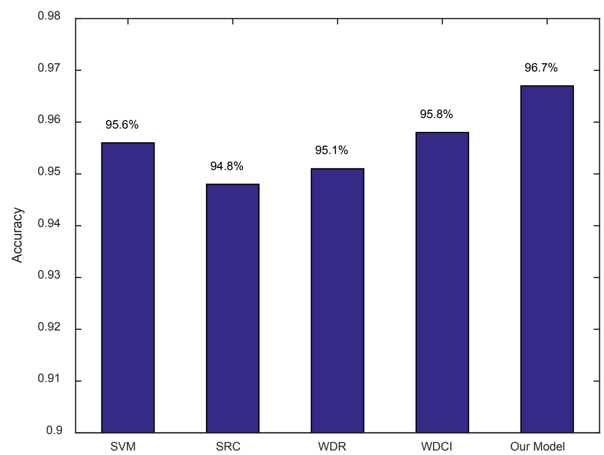


Figure 6: Accuracy of different classifiers on the Weizmann dataset

Table 5: Comparison of accuracy using different methods on the Weizmann dataset

Methods	Accuracy (%)
Our method	96.7
Ali and Shah (2010)	95.8
Junejo, Dexter, Laptev et al. (2010)	95.3
Yang, Zhang, Feng et al. (2011)	96.4
Wang, Yuan, Hu et al. (2012)	96.7
Castrodad and Sapiro (2012)	95.2
Jiang, Lin and Davis (2013)	95.4
Fanello, Gori, Metta et al. (2013)	96.7
Lu and Kudo (2014)	95.6
Zhang, Xu, Shi et al. (2015)	95.6
Caetano, Santos and Schwartz (2016)	96.3
Cherian, Fernando, Harandi et al. (2017)	**97.5**
Yang, Chang, Luo et al. (2017)	96.2

4.3 Expeiments on the KTH dataset

The videos background of the KTH dataset is relatively complex compared with the Weizmann dataset. There are six different kinds of actions including hand waving (HWav), hand clapping (HClap), boxing (Box), walking (Walk), jogging (Jog), and running (Run). This database includes indoor and outdoor cases. Totally it has 599 video clips, which is enough for training. Some action examples are given in Fig. 4(b).

Fig. 7 presents the average confusion matrix of six different actions. We can see from Fig. 7 that four actions can be detected by our model. Unfortunately, it is difficult to discriminate the "Jog" and "Run" actions. The reason is that "Jog" and "Run" are quite similar. Similarly, our model cannot discriminate the "HWave" and "HClap" actions, which also seems the same.

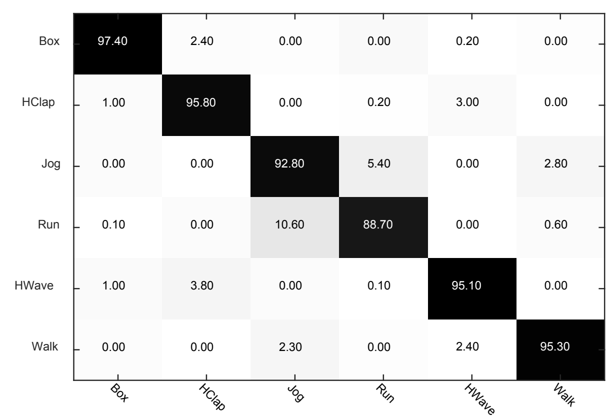


Figure 7: Average confusion matrix of different actions on the KTH dataset

Tab. 6 lists the accuracy of HOG and HOF, HWOM and HOF, and Deep CNN descriptor. The best accuracy is marked with bold font. Obviously, the Deep CNN descriptor can provide much better feature representation than others. The reason is that the motion context is merged into the Deep CNN descriptor. Fig. 8 presents the accuracy of different classifiers on the KTH dataset. It is clear that our method obtains higher recognition rate than the traditional SVM and SRC on the KTH dataset. Similarly, when neither the representation-constrained term nor the dictionary incoherence term is removed, the recognition rate of WDR or WDCI will be slightly lower than our method. The results of recognition accuracy using different methods on the KTH dataset is presented in Tab. 7. It can be seen from Tab. 7 that our model can obtain competitive results with other state-of-the-art algorithms.

Table 6: Accuracy of different descriptors on the KTH dataset

Descriptors	Accuracy (%)
HOG and HOF	88.4
HWOM and HOF	89.2
Deep CNN	**94.2**

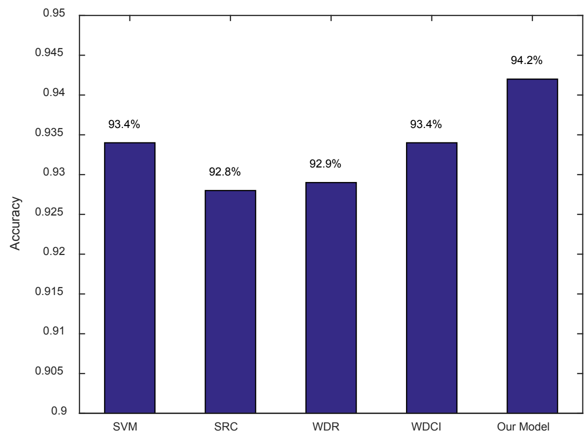


Figure 8: Accuracy of different classifiers on the KTH dataset

Table 7: Comparison of accuracy using different methods on the KTH dataset

Methods	**Accuracy (%)**
Our method	94.20
[Ali and Shah (2010)]	**94.22**
[Junejo, Dexter, Laptev et al. (2010)]	93.60
[Yang, Zhang, Feng et al. (2011)]	94.00
[Wang, Yuan, Hu et al. (2012)]	94.17
[Castrodad and Sapiro (2012)]	92.80
[Jiang, Lin and Davis (2013)]	93.10
[Fanello, Gori, Metta et al. (2013)]	93.17
[Wang, Sun, Liu et al. (2013)]	92.36
[Lu and Kudo (2014)]	93.30
[Zhang, Xu, Shi et al. (2015)]	93.23
[Caetano, Santos and Schwartz (2016)]	93.80
[Cherian, Fernando, Harandi et al. (2017)]	93.90
[Yang, Chang, Luo et al. (2017)]	93.50

4.4 Expeiments on the Cambridge-Gesture dataset

To further distinguish the effectiveness of our model, we conduct experiments on the Cambridge-Gesture dataset. The Cambridge-Gesture dataset consists of 900 image videos of nine different hand gestures including V-shape-leftward (VLeft), V-shape-rightward (VRight), V-shape-contract (VCont), flat-leftward (FLeft), flat-rightward (FRight), flat-contract (FCont), spread-leftward (SLeft), spread-rightward (SRight), and spread-contract (SCont). Some action examples are given in Fig. 4(c).

The videos under single plain illumination are used for training, and the videos under other illuminations are used for testing. The average confusion matrix on the Cambridge-Gesture dataset is presented in Fig. 9. It can be seen from Fig. 9 that different actions have different performances. Fig. 9 implies that the "FLeft", "FCont", and "VLeft" actions are more easily discriminated than other actions.

The accuracy of three different descriptors is shown in Tab. 8. Compared with the traditional HOG and HOF descriptor, both HOWM and HOG and deep CNN descriptors can offer much more discriminative ability. We can find from Tab. 8 that the accuracy of our selected features raises up to 82.11%. This is because the deep network information is fused with our model. Therefore, the Deep CNN descriptor will provide richer representation, which can greatly increase the classification rate. Similarly, our method performs better than SVM and SRC, as shown in Fig. 10. However, different from the performance on the Weizmann and KTH datasets, the accuracy using SRC is approximate to that based on SVM. Moreover, the accuracy using WDR is only slightly higher than that of SVM. When the dictionary incoherence term is removed, the accuracy using WDCI remains the same as that of our model. This means that on the Cambridge-Gesture dataset, WDCI has the same classification ability as our model. The overall average accuracy is 82.11%, which is comparable to several state-of-the-art methods, as shown in Tab. 9.

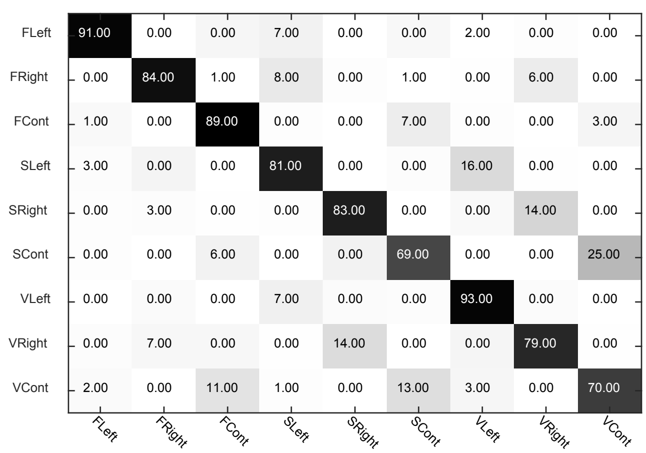


Figure 9: Average confusion matrix of different actions on the Cambridge-Gesture dataset

Table 8: Accuracy of different descriptors on the Cambridge-Gesture dataset

Descriptors	Accuracy (%)
HOG and HOF	76.12
HOWM and HOG	79.18
Deep CNN	**82.11**

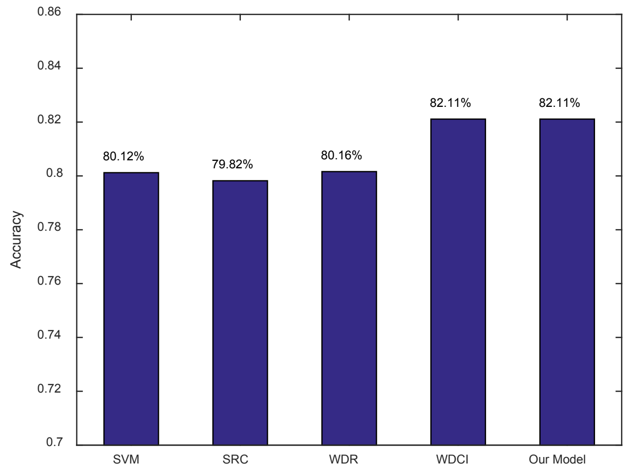


Figure 10: Accuracy of different classifiers on the Cambridge-Gesture dataset

Table 9: Comparison of accuracy using different methods on the Cambridge-Gesture dataset

Methods	Accuracy (%)
Our method	82.11
[Kim and Cipolla (2009)]	82.00
[Yang, Zhang, Feng et al. (2011)]	**82.19**
[Castrodad and Sapiro (2012)]	80.56
[Jiang, Lin and Davis (2013)]	81.06
[Zhang, Xu, Shi et al. (2015)]	82.00
[Yang, Chang, Luo et al. (2017)]	81.57
[Lu, Wang and Zhou (2017)]	65.00
[Tu, Yue, Zhou et al. (2017)]	66.00
[Hou, Li, Wang et al. (2018)]	82.14

5 Conclusion

In this paper, we present a novel sparse representation model for recognizing human actions under complex environments. Following the popular feature extraction approach, the deep CNN approach is firstly applied to action recognition area. Then, we design a modified sparse model to learn a dictionary used for classification. The two terms introduced in our sparse model, i.e., the representation-constrained term and the coefficient incoherence term, can ensure that the learned dictionary has stronger discriminative ability than other state-of-the-art models. Finally, a corresponding classification scheme is presented on the basis of the proposed sparse model. Experiment results on three benchmark datasets verified that our framework works well, and the proposed sparse model can make classification more effective.

Acknowledgement: This research was funded by the National Natural Science Foundation of China (21878124, 31771680 and 61773182).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Ali, S.; Shah, M. (2010): Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288-303.

Beck, A.; Teboulle, M. (2009): A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Science*, vol. 2, no. 1, pp. 183-202.

Bian, W.; Tao, D.; Rui, Y. (2012): Cross-domain human action recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B* (*Cybernetics*), vol. 42, no. 2, pp. 298-307.

Caetano, C.; Santos, J. A. D.; Schwartz, W. R. (2016): Optical flow co-occurrence matrices: a novel spatiotemporal feature descriptor. *23rd International Conference on Pattern Recognition*, pp. 1947-1952.

Castrodad, A.; Sapiro, G. (2012): Sparse modeling of human actions from motion imagery. *International Journal of Computer Vision*, vol. 100, no. 1, pp. 1-15.

Chen, J.; Shan, S.; He, C.; Zhao, G.; Pietikainen, M. et al. (2010): WLD: a robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705-1720.

Cherian, A.; Fernando, B.; Harandi, M.; Gould, S. (2017): Generalized rank pooling for activity recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1581-1590.

Comaniciu, D.; Ramesh, V.; Meer, P. (2003): Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564-575.

Dalal, N.; Triggs, B. (2005): Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Fahad, S. K.; Joost, V. W.; Rao, M. A.; Andrew, D. B.; Michael, F. et al. (2018): Scale coding bag of deep features for human attribute and action recognition. *Machine Vision & Applications*, vol. 29, no. 1, pp. 55-71.

Fanello, S. R.; Gori, I.; Metta, G.; Odone, F. (2013): Keep it simple and sparse: real-time action recognition. *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2617-2640.

Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. (2007): Action as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253.

Hara, K.; Kataoka, H.; Satoh, Y. (2017): Learning spatio-temporal features with 3D residual networks for action recognition. *IEEE International Conference on Computer Vision Workshops*, pp. 3154-3160.

Hou, H.; Li, Z.; Wang, P.; Li, W. (2018): Skeleton optical spectra based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807-811.

Jiang, Z.; Lin, Z.; Davis, L. S. (2013): Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651-2664.

Junejo, I. N.; Dexter, E.; Laptev, I.; Pérez, P. (2010): View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172-185.

Kim, T. K.; Cipolla, R. (2009): Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415-1428.

Kong, S.; Wang, D. (2012): A dictionary learning approach for classification: separating the particularity and the commonality. *12th European Conference on Computer Vision*, pp. 186-199.

Laptev, I. (2005): On space-time interest points. *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123.

Li, S.; Gong, D.; Yuan, Y. (2013): Face recognition using Weber local descriptors. *Neurocomputing*, vol. 122, pp. 272-283.

Lu, G.; Kudo, M. (2014): Learning action patterns in difference images for efficient action recognition. *Neurocomputing*, vol. 123, pp. 328-336.

Lu, H.; Fang, G.; Shao, X.; Li, X. (2012): Segmenting human from photo images based on a coarse-to-fine scheme. *IEEE Transactions on Systems, Man and Cybernetics, Part B* (*Cybernetics*), vol. 42, no. 3, pp. 889-899.

Lu, J.; Wang, G.; Zhou, J. (2017): Simultaneous feature and dictionary learning for image set based face recognition. *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4042-4054.

Mairal, J.; Bach, F.; Ponce, J. (2012): Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791-804.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zissserman, A. (2008): Discriminative learned dictionaries for local image analysis. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. (2008): Supervised dictionary learning. *21th Annual Conference on Neural Information Processing Systems*, pp. 1-8.

Minhas, R.; Mohanmmed, A. A.; Wu, Q. M. (2012): Incremental learning in human action recognition based on snippets. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 11, pp. 1529-1541.

Oikonomopoulos, A.; Patras, I.; Pantic, M. (2006): Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man and Cybernetics, Part B* (*Cybernetics*), vol. 36, no. 3, pp. 710-719.

Ramirez, I.; Sprechmann, P.; Sapiro, G. (2010): Classification and clustering via dictionary learning with structured incoherence and shared features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3501-3508.

Sapuppo, F.; Umana, E.; Frasca, M.; Rosa, M. L.; David, S. et al. (2006): Complex spatio-temporal features in meg data. *Mathematical Biosciences & Engineering*, vol. 3, no. 4, pp. 1547-1063.

Schindler, K.; Gool, L. V. (2008): Action snippets: how many frames does human action recognition require? *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.

Shao, L.; Zhen, X.; Liu, Y.; Ji, L. (2011): Human action representation using pyramid correlogram of oriented gradients on motion history images. *International Journal of Computer Mathematics*, vol. 88, no. 18, pp. 3882-3895.

Shen, L.; Wang, S.; Sun, G.; Jiang, S.; Huang, Q. (2013): Multi-level discriminative dictionary learning towards hierarchical visual categorization. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 383-390.

Tu, H.; Yue, Y.; Zhou, X.; Luo, K. (2017): Novel action recognition via improved PLSA and CBR. *Computer Science*, vol. 44, no. 6, pp. 283-289.

Wang, B.; Li, W.; Yang, W.; Liao, Q. (2011): Illumination normalization based on Weber's law with application to face recognition. *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 462-465.

Wang, H.; Yuan, C.; Hu, W.; Sun, C. (2012): Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognition*, vol. 45, no. 11, pp. 3902-3911.

Wang, J.; Sun, X.; Liu, P.; She, M. F. H.; Kong, L. (2013): Sparse representation of local spatial-temporal features with dimensionality reduction for motion recognition. *Neurocomputing*, vol. 115, pp. 150-160.

Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. (2016): A discriminative learning approach for deep face recognition. *14th European Conference on Computer Vision*, pp. 499-515.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; Ma, Y. (2009): Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227.

Wright, J.; Nowak, R. D.; Figueiredo, M. A. T. (2009): Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479-2493.

Yang, M.; Chang, H.; Luo, W.; Yang, J. (2017): Fisher discrimination dictionary pair learning for image classification. *Neurocomputing*, vol. 269, pp. 13-20.

Yang, M.; Zhang, L.; Feng, X.; Zhang, D. (2014): Sparse representation based Fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209-232.

Yang, M.; Zhang, L.; Feng, X.; Zhang, D. (2011): Fisher discrimination dictionary learning for sparse representation. *IEEE International Conference on Computer Vision*, pp. 543-550.

Yang, M.; Zhang, L.; Yang, J.; Zhang, D. (2010): Metaface learning for sparse representation based face recognition. *IEEE International Conference on Image Processing*, pp. 1601-1604.

Zhang, Q.; Li, B. X. (2010): Discriminative K-SVD for dictionary learning in face recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2691-2698.

Zhang, T; Xu, L.; Shi, P.; Jia, W. (2015): Sparse coding-based spatiotemporal saliency for action recognition. *IEEE International Conference on Image Processing*, pp. 2045-2049.

Zhen, X.; Shao, L.; Tao, D.; Li, X. (2013): Embedding motion and structure features for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1182-1190.

Zhou, N.; Shen, Y.; Peng, J. Y.; Fan, J. P. (2012): Learning inter-related visual dictionary for object recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3490-3497.