ICIC Express Letters ICIC International ©2016 ISSN 1881-803X
Volume **10**, Number **1**, January **2016** pp. **153–160**

DETECTING DISCRIMINATIVE SPATIO-TEMPORAL PARTS USING SIMILARITY-CONSTRAINED LATENT STRUCTURAL SUPPORT VECTOR MACHINE FOR ACTION RECOGNITION

Shijian Huang[1,2], Junyong Ye[1], Tongqing Wang[1], Li Jiang[3] and Yang Li[1]

[1]Key Laboratory of Optoelectronic Technology and Systems of the Ministry of Education
Chongqing University
No. 174, Shazheng Street, Shapingba District, Chongqing 400044, P. R. China
{ ygyocr; huangshijian }@cqu.edu.cn

[2]School of Electronic Information Engineering
Yangtze Normal University
No. 98, Lidujulong Ave., Fuling District, Chongqing 408100, P. R. China

[3]College of Automation
Chongqing University of Posts and Telecommunications
No. 2, Chongwen Road, Nan'an District, Chongqing 400065, P. R. China

Received July 2015; accepted October 2015

Abstract. *Recently, representing videos using spatio-temporal parts becomes popular for action recognition. It can help localise which parts of a video are significant and discriminative. To find the discriminative parts, these methods usually first learn a set of candidate spatio-temporal part detectors and then heuristically select a portion of them. However, it is difficult to determine a good criteria for selection. Moreover, they employ two independent processes, i.e., learning and selection, neglecting the influences between each other. In this letter, we introduce group sparse regularizer into latent structural support vector machine to automatically learn and select a set of discriminative part detectors in a unified framework. We further employ similarity constraints to force the detected spatio-temporal parts from the same class to be more similar and consistent. We also propose an iterative method to compute the spatio-temporal parts with similarity constraints. At last, the max-pooled responses to learned part detectors are normalized and form the final action representation. We validate our method on two public datasets, and experiment results show the promising results of our method.*
Keywords: Similarity constraints, Latent structural support vector machine, Group sparse regularization, Action recognition

1. **Introduction.** Recognizing human actions is widely studied in computer vision field with many important applications, such as video surveillance, video retrieving, and human-computer interaction.

Many popular methods have been proposed for action recognition. For instance, Zhu et al. [1] present sparse coding on local spatial-temporal volumes to recognize actions. Sheng et al. [2] use direction-dependent feature pairs to represent actions. In [3, 4] contextual representation is employed for action recognition. Wang et al. [5] extract dense trajectories of action sequences to represent actions. Sun et al. [6] present sparseness and self-similarity to recognize actions. Despite that these techniques have made many achievements, there remain some challenging and complex problems due to occlusion, cluttered background and other factors.

Recently, representing videos as part-based model has attracted much attention and performed well on some benchmark datasets. Xie et al. [7] employed deformable part model for action representation. Sapienza et al. [8] learn the most discriminative spatio-temporal part for each action. However, an action usually has multiple spatio-temporal parts which can jointly separate it from other actions. In [9, 10, 11], to learn a set of discriminative spatio-temporal parts, they usually first learn spatio-temporal part detectors separately from spatio-temporal part clusters by discriminative learning method, such as support vector machine. Then a fixed number of them are selected and unimportant ones are discarded according to some heuristic criteria. However, these methods learn discriminative spatio-temporal part detectors in two independent steps neglecting the influences between each other. Moreover, it is heuristic to define a criteria for ranking and selecting spatio-temporal part detectors and difficult to determine a best number of them. This will lead to bad generalization performance for new scenarios.

In this letter, we propose a similarity-constrained latent structural support vector machine (SCLSSVM) model. Different from the part-based methods mentioned above, our method automatically learns and selects a set of discriminative spatio-temporal part detectors in a single framework. In our method, we unify learning and selection into a single process by incorporating group sparse regularizer into latent structural support vector machine (LSSVM) model [12]. Considering each spatio-temporal part detector as a group, group sparse regularizer forces the model to automatically learn and select a set of important part detectors by setting unimportant ones to zero in a max-margin framework. Moreover, to make the detected parts more consistent, we introduce the similarity constraints on detected parts into LSSVM model, i.e., we expect the detected parts corresponding to the same part detector in each action are similar and consistent as much as possible. Furthermore, we propose an iterative method to fast compute the spatio-temporal parts with similarity constraints. With learned spatio-temporal part detectors, a video can be encoded by max-pooling over the responses of it to the discriminative spatio-temporal part detectors. The max-pooled responses are further normalized and form the final representation for action classification. To validate the effectiveness of our method, we test it on two benchmarks and experiment results show its effectiveness.

The remainder of the letter is organized as follows. Section 2 presents our proposed SCLSSVM model and its solution. Section 3 reports the experimental results on two benchmarks. Finally, we conclude the letter in Section 4.

2. **Proposed Approach.** The flowchart of our proposed approach for action recognition is illustrated in Figure 1. Specifically, we first dense sample spatio-temporal parts and represent them into feature vectors. Then discriminative spatio-temporal part detectors are learned and selected by our proposed SCLSSVM model. With learned part detectors, each video is represented by max-pooled responses which are further normalized and form the final representation for action classification. In this section, we first introduce the description of spatio-temporal parts and then present our SCLSSVM model. At last, we show an iterative method to solve our model.

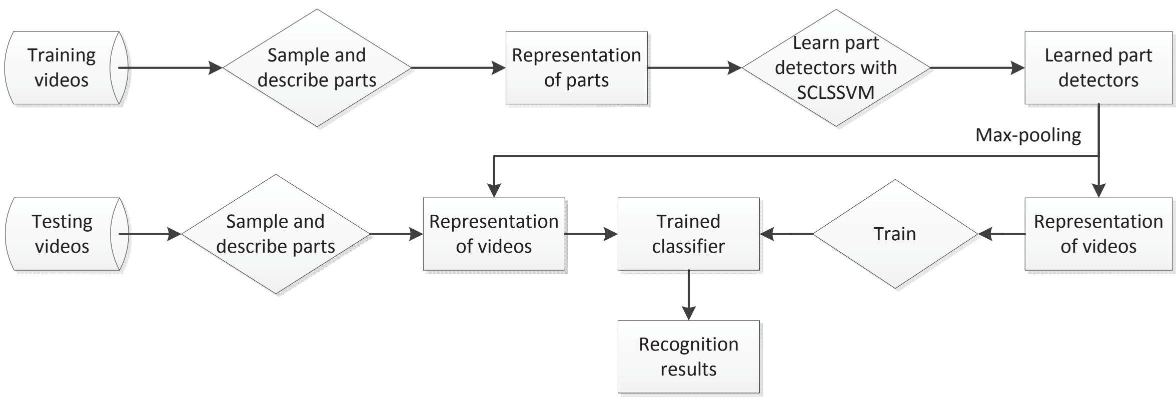

Figure 1. Flowchart of human action recognition with SCLSSVM model

2.1. **Video representation.** Given a video, we first extract and describe dense multi-scale spatio-temporal parts, each of which is a cuboid centred at $z = \{x, y, t\}$, where x, y, t represent the spatio-temporal coordinates respectively. Subsequently, each part is characterized by dense trajectory descriptor [13] and histogram of 3 dimensional oriented gradients (HOG3D) descriptor [14]. For dense trajectory descriptor, there are three channel features in total (i.e., trajectory, histogram of oriented gradients (HOG)/histogram of optical flow (HOF), and motion boundary histograms (MBH)). Following [13], we use the bag-of-features (BoF) representation and the three channel features are clustered by k-means algorithm into W words respectively. Then each channel is represented by a histogram obtained by aggregating the quantized dense trajectory features in the corresponding cuboid and further L1 normalized. The histogram features of three channels are concatenated together, further combined with HOG3D features, and form the final representation for spatio-temporal parts.

2.2. **SCLSSVM model.** The training process is weakly supervised, i.e., we know nothing about the discriminative parts and only have a set of training videos with action labels. We employ latent variables $h \in H$ to indicate the spatio-temporal parts. Given the training video set $\{x_i, y_i\}_{i=1}^{N} \in X \times Y$, the feature vector $\phi(x_i, y, h)$ describes the spatio-temporal part in video x_i with predicted label y under latent variable h. For each action class, we assume that there are K parts which can jointly separate it from other actions. Then the part detectors D for a task with C action classes can be represented as:

$$D = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,K} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ d_{C,1} & d_{C,2} & \cdots & d_{C,K} \end{pmatrix} \tag{1}$$

and the corresponding joint feature vector can be formulated as $(0, \ldots, \phi_{x_i,y,h}, 0, \ldots, 0)$, where most of the entries are zero except the ones at the interval of predicted label y. We further define the response of video x_i associated with part detector $d_{c,k}$ as $f_{d_{c,k}}(x_i, y) = \max_h d_{c,k}^T \phi_{x_i,y,h}$.

Then for a multi-class task, our model can be formulated into a single framework with latent variables:

$$D^* = \arg\min_D \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_i^k + \lambda g(D)$$

$$s.t. \quad \max_h d_{y_i,k}^T \phi_{x_i,y_i,h} - \max_h d_{y,m}^T \phi_{x_i,y,h} \geq \Delta(y_i, y) - \xi_i^k, \quad \xi_i^k \geq 0, \quad \forall i, y, k, m \tag{2}$$

where $\Delta(y_i, y)$ is the loss function, $g(D)$ is the regularizer on part detectors D, and λ is the weight coefficient of $g(D)$. The loss function $\Delta(y_i, y)$ measures the cost of predicting ground-truth label y_i as y. In our work, we set it to be $0-1$ loss function, i.e., $\Delta(y_i, y) = 1$ if $y_i \neq y$ and $\Delta(y_i, y) = 0$ otherwise. Regularizer $g(D)$ forces the model to select a portion of important part detectors and discard non-discriminative ones by setting them to zero. We employ the group sparse regularization technique [15]. Usually, $l_{1,2}$ and $l_{1,inf}$ mixed norms are used in practice and we select $l_{1,2}$ regularizer as the regularization term. Taking each part detector as a group, $g(D)$ can be formulated as $\sum_{c=1}^{C} \sum_{k=1}^{K} \|d_{c,k}\|_2$. The constraints in Equation (2) force that the response $f(x_i, y, h)$ associated with ground-truth label y_i should be larger than that associated with any other predicted labels.

Our proposed model can be solved in a two-step iterative method. Firstly, the latent variables h are determined with part detectors fixed. Secondly, a structural support vector machine with group sparse regularizer without latent variables is solved. Traditional LSSVM model computes latent variables by maximizing responses of videos associated with part detectors neglecting any relationship between detected spatio-temporal parts.

This may lead to large variation and inconsistency in the detected parts from the same class. We expect them not only to be discriminative but also similar and consistent as much as possible. To achieve this goal, we introduce a pairwise similarity-constrained term during computing latent variables.

In our work, we just consider the similarity of latent variables corresponding to ground-truth labels under the ground-truth part detector. We measure the similarity of detected parts using the Euclidean distance function, and the similarity-constrained term can be defined as:

$$S = \sum_i \sum_j d(\phi(x_i, y_i, h), \phi(x_j, y_j, h)) = \sum_i \sum_j -\|\phi(x_i, y_i, h) - \phi(x_j, y_j, h)\|_2^2 \tag{3}$$

Then in our model, for part detector $d_{c,k}$, similarity-constrained latent variables are determined by:

$$h^* = \begin{cases} \arg\max_{h\in H} \left(d_{c,k}^T \phi(x_i, y_i, h) + \alpha S\right), & \text{if } y_i = c \\ \arg\max_{h\in H} d_{c,k}^T \phi\left(x_i, y_i, h\right), & \text{if } y_i \neq c \end{cases} \tag{4}$$

where α is the tradeoff between similarity and response. In the following section, we introduce how to solve our model in detail.

2.3. **Optimization method.** To compute latent variables, we fix spatio-temporal part detectors D. From Equation (4), for part detector $d_{c,k}$, we can observe that it is easy to compute the latent variables for videos whose labels are not equal to c. However, it is inhibitive to directly compute the latent variables for videos whose labels are equal to c. There will be $\prod_{i=1}^{N_p} n_i$ combinations if there are n_i possible latent values for video x_i, where N_p is the number of videos whose labels are equal to c. To efficiently compute latent variables, we reformulate Equation (3) as $S = -\frac{1}{N_p}\sum_{y_i=c} \|\phi(x_i, y_i, h) - \overline{\phi})\|_2^2$, where $\overline{\phi}$ is the mean value of latent variables. Then latent variables can be determined in an iterative manner. Firstly, we compute the mean value of latent variables $\overline{\phi}$. Secondly, we compute S and update the latent variables using Equation (4). We run the two steps iteratively until convergence.

To compute part detectors, we update them one by one with latent variables for all videos fixed. To update $d_{c,k}$, we fix the other part detectors. Then, the object function can be reformulated as:

$$d_{c,k}^* = \arg\min_{d_{c,k}} \frac{1}{N} \sum_{i=1}^{N} \xi_i + \lambda \|d_{c,k}\|_2$$
$$s.t. \quad \max_h d_{y_i,k}^T \phi_{x_i,y_i,h} - \max_h d_{y,m}^T \phi_{x_i,y,h} \geq \Delta(y_i, y) - \xi_i, \quad \xi_i \geq 0, \quad \forall i, y, k, m \tag{5}$$

From Equation (5), we can observe that there are $K(C-1)$ constraints for each video. For the video x_i whose label y_i is equal to c, the constraints can be reformulated as:

$$\xi_i \geq \left[1 + \max_h d_{y,m}^T \phi_{x_i,y,h} - \max_h d_{c,k}^T \phi_{x_i,y_i,h}\right]_+, \quad \forall y \in Y \backslash y_i, m \tag{6}$$

where $[x]_+$ means $\max(x, 0)$. Because the other part detectors are fixed, then $\max_h d_{y,m}^T$ $\phi_{x_i,y,h}$ are constants. The constraints can be rewritten as one constraint:

$$\xi_i \geq \left[1 + \max_{y,m} \max_h d_{y,m}^T \phi_{x_i,y,h} - \max_h d_{c,k}^T \phi_{x_i,y_i,h}\right]_+, \quad \forall y \in Y \backslash y_i, m \tag{7}$$

With the similar analysis, the constraints on the videos whose labels are not equal to c can be rewritten as one constraint as follows:

$$\xi_i \geq \left[1 + \max_h d_{c,k}^T \phi_{x_i,c,h} - \min_m \max_h d_{y_i,m}^T \phi_{x_i,y_i,h}\right]_+, \quad \forall m \tag{8}$$

With the above simplification, there are N constraints in total. Then we can reformulate the object function as an unconstrained problem:

$$d_{c,k}^* = \arg\min_{d_{c,k}} \frac{1}{N} f(d_{c,k}) + \lambda \|d_{c,k}\|_2 \tag{9}$$

where

$$\begin{aligned} f(d_{c,k}) = & \sum_{y_i=c} \left[1 + \max_{y,m} \max_h d_{y,m}^T \phi_{x_i,y,h} - \max_h d_{c,k}^T \phi_{x_i,y_i,h} \right]_+ \\ & + \sum_{y_i \neq c} \left[1 + \max_h d_{c,k}^T \phi_{x_i,c,h} - \min_m \max_h d_{y_i,m}^T \phi_{x_i,y_i,h} \right]_+ \end{aligned} \tag{10}$$

Due to the group sparse regularization on D, we employ a proximal method [16] to update $d_{c,k}$, i.e.,

$$d_{c,k}^{t+1} = soft\left(u^t, \lambda\alpha^t\right) \tag{11}$$

where $soft(u,a) = sign(u)\left[|u| - a\right]_+$, α^t is the step length, and $u^t = d_{c,k}^t - \alpha^t \frac{\partial f}{\partial d_{c,k}^t}$. The partial derivative is computed as $\frac{\partial f}{\partial d_{c,k}^t} = \sum_{y_i=c,\sigma>0} -\phi_{x_i,y_i,h^*} + \sum_{y_i \neq c,\eta>0} \phi_{x_i,c,h^*}$, where $\sigma = 1 + \max_{y,m} \max_h d_{y,m}^T \phi_{x_i,y,h} - \max_h d_{c,k}^T \phi_{x_i,y_i,h}$ and $\eta = 1 + \max_h d_{c,k}^T \phi_{x_i,c,h} - \min_m \max_h d_{y_i,m}^T \phi_{x_i,y_i,h}$.

2.4. **Action classification.** After optimization, non-discriminative part detectors are set to be zero. Given C classes of actions, we will learn KC spatio-temporal part detectors. To efficiently employ the learned part detectors, we code each video by max-pooling the responses of it to the detected spatio-temporal part detectors and it will have KC responses which are further scaled by logistic function $\Psi(x) = \frac{1}{1+\exp(-ax)}$, where a is the smoothing factor and is fixed to be 0.5 in our experiments. Then each video can be characterized by KC scaled values. At last, we employ multi-class non-linear support vector machine with radial basis function kernel for final classification.

3. **Experiments.** In this section, we evaluate the performance of the proposed method on two public datasets: Kungliga Tekniska Hogskolan (KTH) dataset [17] and University of Central Florida (UCF) Sports dataset [18] in a leave-one-out cross validation manner.

The KTH dataset consists of six action categories (i.e., boxing, hand clapping, hand waving, jogging, running and walking,). The UCF Sports dataset contains 10 sport action categories (i.e., Dive, Golf, Kick, Lift, Ride, Run, Skate, BSwing, HSwing, Walk).

3.1. **Parameter settings.** In our work, we extract spatio-temporal parts from a regular grid spacing of 16 pixels in space and 5 pixels in time. The spatio-temporal parts allow for 5 scales in space (1, $\sqrt{2}$, 2, $2\sqrt{2}$, 4)[1] and 3 scales in time (1, $\sqrt{2}$, 2), and the smallest one has a size of $32 \times 32 \times 8$. We use the same scales for the two datasets by down-sampling UCF Sports dataset to half the spatial resolution. Then, spatio-temporal parts are described by HOG3D and dense trajectory features. HOG3D features are computed with parameters $5\times5\times4$ cells with 10 discrete orientations, resulting in the HOG3D features with 1000 dimensions. Dense trajectory features are grouped into 2000 words ($W = 2000$) for trajectory, HOG/HOF, MBH features respectively. To initialize part detectors, we group the dense features of each class videos into 300 ($K = 300$) clusters using k-means algorithm, and set the centres as the initial part detectors for the corresponding class. The coefficients α and λ are set to 0.3 and 1.5 respectively on the two datasets.

[1]Notice that we set the space scales in horizontal and vertical the same.

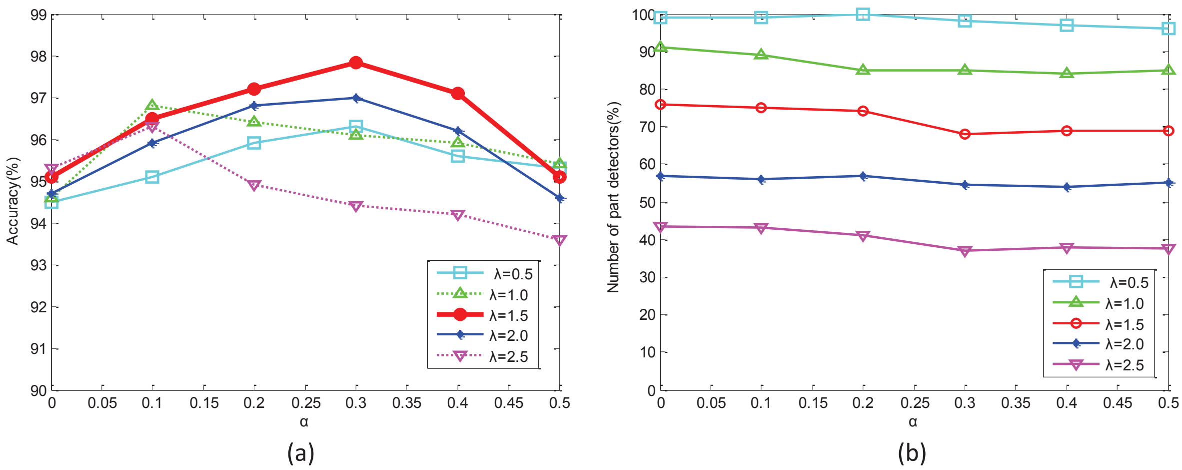

FIGURE 2. (a) Performances with different coefficients, (b) numbers of preserved part detectors with different coefficients

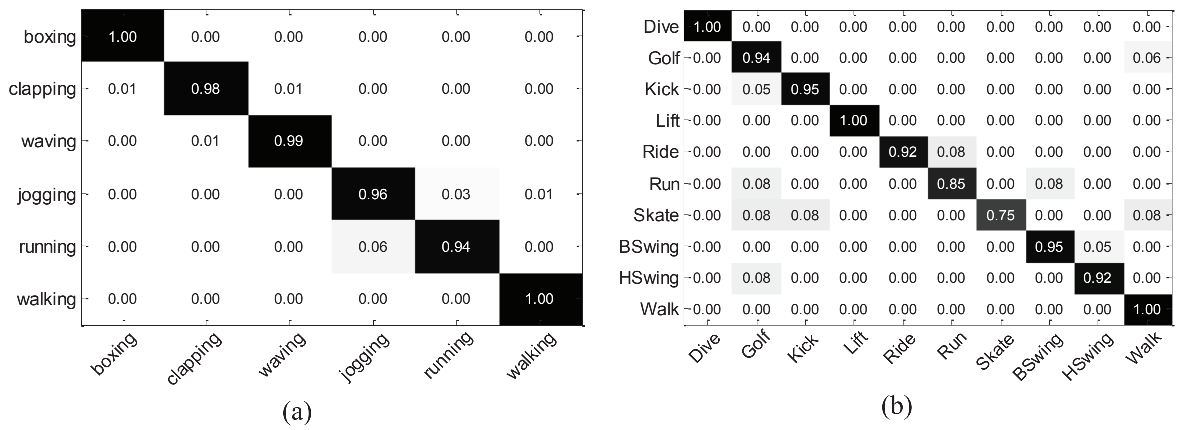

FIGURE 3. Confusion matrix: (a) On KTH datset, (b) On UCF sports datset

3.2. **Experimental results on KTH and UCF Sports datasets.** To investigate the effect of similarity constraints and group sparse regularizer on our model, we test different combinations of α and λ on the two datsets. The effects of the two terms on two datasets are similar. Due to limited space, we only show the comparison results on KTH datset. As illustrated in Figure 2(a), we can observe that the similarity constraints can affect and improve the performance of our model. We also investigate the effect of α and λ on numbers of preserved part detectors. In Figure 2(b), we can observe that the numbers of preserved part detectors are mainly dependent on group sparse regularizer.

Figure 3(a) shows the confusion matrix on KTH dataset. The main confusion occurs between jogging and running, which are performed similarly. Figure 3(b) shows the confusion matrix on UCF Sports dataset. The golf is relatively confused by other class. Finally, we achieve the average accuracies of 97.83% on KTH dataset and 92.67% on UCF Sports dataset.

Table 1 presents a comparison of our method with state-of-the-art methods on the two datasets, which indicate that our method achieves the competitive results.

4. **Conclusions.** In this letter, we represent videos using multi-scale dense spatio-temporal parts which are described by dense trajectory features and HOG3D features and present a similarity-constrained latent structural support vector machine model to learn and select a set of discriminative spatio-temporal part detectors in a weakly supervised

Table 1. Recognition results on two benchmark datasets

Method	Year	KTH (%)	UCF Sports (%)
Zhu et al. [1]	2010	94.92	84.33
Xie et al. [7]	2011	87.33	N/A
Zhang et al. [3]	2012	95.6	87.33
Wang et al. [5]	2013	94.2	88.0
Li et al. [4]	2014	96.33	92
Sapienza et al. [8]	2014	96.73	N/A
Sheng et al. [2]	2015	94.99	87.33
Sun et al. [6]	2015	96.5	88.5
Ours		**97.83**	**92.67**

setting. By introducing group sparse regularizer, non-discriminative part detctors are set to be zero automatically in a single framework. Pairwise similarity constraints are employed to force the detected parts from the same action to be more similar and consistent. Our method is tested on two benchmarks and experiment results show its effectiveness in action recognition. In the future, we will investigate novel fusion strategies to combine part-based features with other outstanding features for recognizing more complex actions.

Acknowledgment. This work was supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission of China under Grant KJ1401207, and Fundamental Research Funds for the Central Universities of China under Grant 106112013CDJZR120014.

REFERENCES

[1] Y. Zhu, X. Zhao, Y. Fu and Y. Liu, Sparse coding on local spatial-temporal volumes for human action recognition, *Proc. of the 10th Asian Conf. on Comput. Vis.*, Queenstown, New Zealand, pp.660-671, 2010.

[2] B. Sheng, W. Yang and C. Sun, Action recognition using direction-dependent feature pairs and non-negative low rank sparse model, *Neurocomputing*, vol.158, pp.73-80, 2015.

[3] Z. Zhang, C. Wang, B. Xiao, W. Zhou and S. Liu, Action recognition using context-constrained linear coding, *IEEE Signal Process. Lett.*, vol.19, no.7, pp.439-442, 2012.

[4] Y. Li, J. Ye, T. Wang and S. Huang, Augmenting bag-of-words: A robust contextual representation of spatiotemporal interest points for action recognition, *Vis. Comput.*, 2014.

[5] H. Wang, A. Klaser, C. Schmid and C. L. Liu, Dense trajectories and motion boundary descriptors for action recognition, *Int. J. Comput. Vis.*, vol.103, no.1, pp.60-79, 2013.

[6] C. Sun, I. N. Junejo, M. Tappen and H. Foroosh, Exploring sparseness and self-similarity for action recognition, *IEEE Trans. Image Process.*, vol.24, no.8, pp.2488-2510, 2015.

[7] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen and D. Zhao, A unified framework for locating and recognizing human actions, *Proc. of 2011 IEEE Int. Conf. on Comput. Vis. and Pattern Recognit.*, Springs, Colorado, pp.25-32, 2011.

[8] M. Sapienza, F. Cuzzolin and P. Torr, Learning discriminative space-time actions from weakly labelled videos, *Int. J. Comput. Vis.*, vol.110, no.1, pp.30-47, 2014.

[9] L. M. Wang, Y. Qiao and X. Tang, Motionlets: Mid-level 3d parts for human motion recognition, *Proc. of IEEE Int. Conf. Comput. Vis. and Pattern Recognit.*, Portland, Oregon, pp.2674-2681, 2013.

[10] W. Zhang, M. Zhu and K. G. Derpanis, From actemes to action: A strongly-supervised representation for detailed action understanding, *Proc. of IEEE Int. Conf. on Comput. Vis.*, Darling Harbour, Sydney, pp.2248-2255, 2013.

[11] A. Jain, A. Gupta, M. Rodriguez and L. S. Davis, Representing videos using mid-level discriminative patches, *Proc. of IEEE Int. Conf. Comput. Vis. and Pattern Recognit.*, Portland, Oregon, pp.2571-2578, 2013.

[12] C. N. J. Yu and T. Joachims, Learning structural svms with latent variables, *Proc. of the 26th Annu. Int. Conf. Mach. Learn.*, pp.1169-1176, 2009.

[13] H. Wang, A. Klaser, C. Schmid and C. L. Liu, Action recognition by dense trajectories, *Proc. of IEEE Int. Conf. on Comput. Vis. and Pattern Recognit.*, Colorado Springs, CO, pp.3169-3176, 2011.

[14] A. Klaser, M. Marszalek and C. Schmid, A spatio-temporal descriptor based on 3d-gradients, *Proc. of Br. Mach. Vis. Conf.*, Leeds, UK, pp.1-8, 2008.

[15] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. Series B Stat. Method.*, vol.68, no.1, pp.49-67, 2006.

[16] S. J. Wright, R. D. Nowak and M. A. T. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Trans. Signal Process.*, vol.57, no.7, pp.2479-2493, 2009.

[17] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local svm approach, *Proc. of the 17th Int. Conf. Pattern Recognit.*, Cambridge, England, pp.32-36, 2004.

[18] M. D. Rodriguez, J. Ahmed and M. Shah, Action mach: A spatiotemporal maximum average correlation height filter for action recognition, *Proc. of IEEE Int. Conf. Comput. Vis. and Pattern Recognit.*, Anchorage, AK, pp.1-8, 2008.