VIOLENT VIDEO DETECTION BASED ON MoSIFT FEATURE AND SPARSE CODING

Long Xu[1], *Chen Gong*[1], *Jie Yang*[1*], *Qiang Wu*[2], *Lixiu Yao*[1]

[1]Institution of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[2]School of Computing and Communications, University of Technology, Sydney, Australia

ABSTRACT

To detect violence in a video, a common video description method is to apply local spatio-temporal description on the query video. Then, the low-level description is further summarized onto the high-level feature based on Bag-of-Words (BoW) model. However, traditional spatio-temporal descriptors are not discriminative enough. Moreover, BoW model roughly assigns each feature vector to only one visual word, therefore inevitably causing quantization error. To tackle the constrains, this paper employs Motion SIFT (MoSIFT) algorithm to extract the low-level description of a query video. To eliminate the feature noise, Kernel Density Estimation (KDE) is exploited for feature selection on the MoSIFT descriptor. In order to obtain the highly discriminative video feature, this paper adopts sparse coding scheme to further process the selected MoSIFTs. Encouraging experimental results are obtained based on two challenging datasets which record both crowded scenes and non-crowded scenes.

Index Terms— violent video detection, Motion SIFT, kernel density estimation, sparse coding, max pooling

1. INTRODUCTION

Computer vision techniques are highly demanded for intelligent surveillance and automatic video annotation. In this paper, we focus on the challenging task of detecting violence in videos, which is insufficiently studied but really useful in rating/tagging video content and video surveillance. Any videos containing human fighting are defined as violent videos. The intra-class variations of human motion caused by scale, occlusion, viewpoint, and the clutter background make violence detection difficult.

To detect violent video, the visual feature is constructed based on either local spatio-temporal descriptors or global features. Global feature represents an action as a whole. For example, space-time shape templates from image sequences were used in [1, 2] to describe an action. This method requires foreground segmentation to extract precise silhouettes, which is difficult in a real environment. The Violent Flow (ViF) descriptor [3] is another global feature. It represents the statistics of flow-vector magnitudes changing over time. However, ViF is designed for crowded violent scenes and is not suitable for the scenes of less people.

Approaches based on local spatio-temporal descriptors are commonly combined with Bag-of-Words (BoW) model and have achieved promising performance in violence detection [4, 5]. Compared with the space-time shape and tracking based approaches, these methods do not require foreground segmentation or body parts tracking. Thus they are more robust to camera movement, illumination, occlusion and even low resolution video. These methods first detect spatio-temporal interest points [6, 7, 8] from video clips and then describe cuboids around the interest points using different spatio-temporal descriptors like Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF). Then, each local feature vector is quantized to its closest visual word, and a histogram of visual words occurrence is generated as the video level representation. These fixed-dimensional histogram vectors can then be fed into the standard classifier such as support vector machine (SVM) [9]. The visual word dictionary is typically constructed through K-means clustering over the sampled local descriptors. Each word in the dictionary is the cluster center obtained by K-means.

The conventional BoW methods rely on the discriminative power of local spatio-temporal descriptors, and focus on how often they occur in the video. However, traditional descriptors like HOG and HOF are not descriptive enough to capture both local appearance and motion information. To tackle this problem, Motion SIFT (MoSIFT) algorithm [10] was proposed to detect distinctive local features through local appearance and motion. Moreover, the performance of BoW model is impaired because of high quantization error. Recently, the sparse coding based method has been successfully utilized in image classification task [11, 12, 13] and action classification domain [14, 15]. Sparse coding method transforms each low-level descriptor to a linear combination of a few "atoms" in a well-trained dictionary. Compared with BoW model, it can achieve a much lower reconstruction error and generate a more discriminative video representation.

Motivated by the above insights, we take advantage of the robust MoSIFT descriptor and sparse coding method to generate a better representation of violent video. The framework of our approach is illustrated in Fig. 1. Firstly, we extract MoSIFT features from video clips. Secondly, we em-

*Corresponding author: Jie Yang, jieyang@sjtu.edu.cn

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

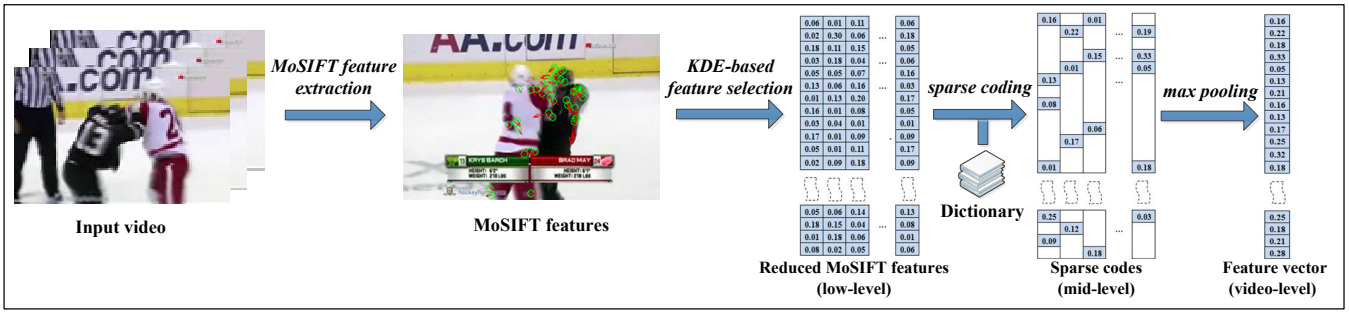

Fig. 1. Framework of the proposed violence detection approach.

ploy Kernel Density Estimation (KDE) based feature selection method to select the most representative features from the original 256-dimensional MoSIFT descriptor. Subsequently, sparse coding is adopted to transform the reduced low-level descriptors into compact mid-level features. To obtain a highly discriminative representation of the whole video, max pooling process is operated over the whole sparse code set of the query video. Finally, a SVM classifier is trained using these video level feature vectors.

2. OUR APPROACH

2.1. MoSIFT algorithm

The MoSIFT algorithm [10] was inspired by the highly successful Scale-Invariant Feature Transform (SIFT) [16] for object recognition. First, the standard SIFT algorithm is applied to find visually distinctive interest points in the spatial domain. Then the candidate points with insufficient optical flow around the neighborhood are rejected, leaving only spatio-temporal interest points with strong motion. The MoSIFT descriptor was designed to represent the feature point in two parts: a standard SIFT image descriptor and an analogous histogram of optical flows. The final MoSIFT feature is a 256 dimensional vector: the first 128 dimensions are the standard SIFT features and the remaining 128 dimensions are the aggregated histogram of optical flow.

MoSIFT algorithm detects interest points from a video clip. Then it not only encodes their local appearance but also explicitly models local motion. Compared with the popular spatio-temporal descriptors such as HOG [17] and HOF [17], the MoSIFT descriptor is more descriptive and more robust to deformation.

2.2. KDE-based feature selection

The original 256-dimensional MoSIFT descriptor may contain some irrelevant and redundant features. To improve performance and computational efficiency, we employ the KDE-based feature selection method [18] to select the most representative features from the original MoSIFT descriptor. KDE is a traditional non-parametric method for inferring the underlying probability density function (PDF). Suppose

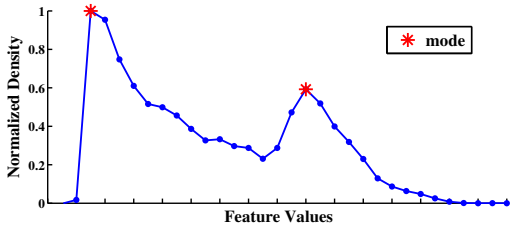

Fig. 2. Normalized probability density function estimated by KDE method.

$x_1, x_2, ..., x_N$ is N independent identically distributed observed data of a one-dimensional random variable x. KDE infers the probability density function of x by centering a kernel function $K(x)$ at each data point x_i:

$$\hat{f}_h(x) = \frac{1}{hN} \sum_{i=1}^{N} K(\frac{x - x_i}{h}), \quad (1)$$

where $h > 0$ is a smoothing parameter called bandwidth.

For j-th feature of MoSIFT descriptor, we can use KDE to obtain a smooth probability density function based on the training data. $K(x)$ is chosen to be a Gaussian kernel: $K(x) = (1/\sqrt{2\pi})e^{-(1/2)x^2}$. The bandwidth h can be adaptively chosen using the method proposed in [19]. If the probability density function of a feature is bimodal or multimodal, this feature is considered to be more discriminative than those with only a single mode. Fig. 2 shows a typical PDF of a feature with two modes. On the original 256 features of MoSIFT, we estimate the PDF of each feature. According to the number of modes, we sort the 256 features of MoSIFT descriptor in descending order. Then the first 150 features are selected to form the reduced MoSIFT descriptor which is more effective than the original one.

2.3. Sparse coding scheme for violence detection

In our violence detection framework, sparse coding instead of BoW model is adopted to provide a more accurate and discriminative intermediate representation for human action. Let $\mathbf{X}$ be a set of reduced MoSIFT feature vectors extracted from a query video clip, i.e. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{d \times N}$, where $\mathbf{x}_i$ denotes i-th feature vector of the total N data samples. A

sparse coding problem can be formulated as

$$\mathbf{Z} = \underset{\mathbf{Z}\in\mathbb{R}^{k\times N}}{\arg\min} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_{\ell_2}^2 + \lambda\|\mathbf{Z}\|_{\ell_1}, \quad (2)$$

where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N] \in \mathbb{R}^{k\times N}$ and $\mathbf{z}_i$ is the sparse representation of the feature vector $\mathbf{x}_i$. $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k] \in \mathbb{R}^{d\times k}$ is a pre-trained dictionary, which is an overcomplete basis set, i.e. $k > d$. λ is a positive regularization parameter to control the tradeoff between reconstruction error and sparseness. When the dictionary $\mathbf{D}$ is fixed, the optimization over $\mathbf{Z}$ alone is convex. The LARS-lasso method [20] is utilized to solve Eq. (2) to get the set of sparse codes $\mathbf{Z}$. In this way, the original query video representation in $\mathbf{X}$ is converted to the corresponding spare code representation $\mathbf{Z}$. Then, the video analysis/recognition is carried out on $\mathbf{Z}$ domain.

The dictionary $\mathbf{D}$ contains k atoms representing basic patterns of the specific data distribution in feature space. Given a large collection of the reduced MoSIFT features extracted from training video clips $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M] \in \mathbb{R}^{d\times M}$, the dictionary learning problem in sparse coding scheme can be defined by

$$\underset{\mathbf{U}\in\mathbb{R}^{k\times M}, \mathbf{D}\in\mathcal{C}}{\arg\min} \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\|\mathbf{y}_i - \mathbf{D}\mathbf{u}_i\|_{\ell_2}^2 + \lambda\|\mathbf{u}_i\|_{\ell_1}, \quad (3)$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M] \in \mathbb{R}^{k\times M}$ is the coefficients set and $\mathcal{C}$ is a convex set

$$\mathcal{C} \triangleq \left\{\mathbf{D} \in \mathbb{R}^{d\times k}, s.t. \|\mathbf{d}_i\|_{\ell_2} \leqslant 1, i \in \{1, ..., k\}\right\}.$$

The formulation is not convex with respect to $\mathbf{D}$ and $\mathbf{U}$ simultaneously. We adopt the online dictionary learning algorithm [21] to solve this joint optimization problem, which has been proven to be more suitable for large training sets.

2.4. Max pooling over motion features

To capture the global statistics of the whole video, max pooling is applied over sparse code set $\mathbf{Z} \in \mathbb{R}^{k\times N}$ to get a video level feature,

$$\boldsymbol{\beta} = \mathcal{F}(\mathbf{Z}), \quad (4)$$

where $\boldsymbol{\beta}$ is a vector with k dimensions and $\mathcal{F}$ is a pooling function defined on each row of $\mathbf{Z}$. Different pooling functions construct different video statistics [14, 15]. It has been reported empirically and also theoretically that max pooling outperforms the average pooling [11, 22]. In this work, we adopt the max pooling function defined as

$$\beta_i = max\{|Z_{i1}|, |Z_{i2}|, ..., |Z_{iN}|\}, \quad (5)$$

where β_i is the i-th element of $\boldsymbol{\beta}$, Z_{ij} denotes the (i, j)-th entry of the matrix $\mathbf{Z}$.

Compared with the BoW model, sparse coding method achieves a much lower reconstruction error and captures the salient properties of human actions. By max pooling procedure over the sparse code set, the irrelevant information is discarded. Only the strongest response to each particular atom in dictionary is preserved. It generates a compact and discriminative video feature $\boldsymbol{\beta}$ for our violence detection task. SVM is then employed to classify $\boldsymbol{\beta}$ as either violent or non-violent.

Fig. 3. Sample frames from Hockey Fight dataset (first row) and Crowd Violence dataset (second row). The left three columns are violent scenes while the right three columns are non-violent scenes.

3. EXPERIMENTS

3.1. Datasets

We carry out the experiments on two challenging datasets created specifically for violent video detection: Hockey Fight [5] and Crowd Violence [3]. Fig. 3 shows a few sample frames from each dataset.

Hockey Fight dataset. This dataset contains 1000 video clips of action from hockey games of the National Hockey League (NHL). 500 videos in the dataset are manually labeled as fight and others are labeled as non-fight. Each clip consists of 50 frames with a resolution of 360×288 pixels.

Crowd Violence dataset. This dataset is assembled for violent crowd behavior detection. All video clips are collected from YouTube, presenting a wide range of scene types, video qualities and surveillance scenarios. The dataset consists of 246 video clips including 123 violent clips and 123 normal clips with a resolution of 320×240 pixels. The whole dataset is split into five sets for 5-fold cross validation. Half of the footages in each set present violent crowd behavior and the other half presents non-violent crowd behavior.

3.2. Experimental settings

The regularization parameter λ in Eq. (2) and Eq. (3) is set to $\frac{1.2}{\sqrt{m}}$ according to [21], where m is the dimension of the original feature. In our approach, the dimension of the reduced MoSIFT feature is 150. Hence $m = 150$ and $\lambda \approx 0.098$. To assess the impact of dictionary size, we learn dictionaries of different sizes. Both the MoSIFT feature and the final video level feature vector are ℓ_2 normalized. To evaluate the classification accuracy, we employ the 5-fold cross validation test on each dataset.

3.3. Results and discussions

We compare the proposed method against the state-of-the-art techniques including BoW based methods, Local Trinary Pat-

Table 1. Violence detection performance of various algorithms on Hockey Fight dataset (5-fold cross validation)

Dictionary	HOG + BoW [5]	HOF + BoW [5]	MoSIFT + BoW [5]	MoSIFT + Sparse Coding		MoSIFT + KDE + Sparse Coding	
				ACC $\pm$ SD	AUC	ACC $\pm$ SD	AUC
50 words	87.8%	83.5%	87.5%	88.3 $\pm$ 1.35%	0.9220	90.9 $\pm$ 1.82%	0.9512
100 words	89.1%	84.3%	89.4%	90.1 $\pm$ 0.89%	0.9410	92.6 $\pm$ 2.19%	0.9579
150 words	89.7%	85.9%	89.5%	91.9 $\pm$ 1.52%	0.9579	93.4 $\pm$ 1.85%	0.9630
200 words	89.4%	87.5%	90.4%	92.7 $\pm$ 1.92%	0.9670	94.1 $\pm$ 1.64%	**0.9713**
300 words	90.8%	87.2%	90.4%	93.1 $\pm$ 1.52%	0.9598	94.1 $\pm$ 1.71%	0.9682
500 words	91.4%	87.4%	90.5%	93.0 $\pm$ 1.27%	0.9661	**94.3 $\pm$ 1.68%**	0.9708
1000 words	**91.7%**	**88.6%**	**90.9%**	**93.6 $\pm$ 1.67%**	**0.9694**	94.0 $\pm$ 1.97%	0.9666

tern (LTP) [23] and ViF. SVM with RBF kernel is adopted as classifier in all the mentioned approaches. Results are reported with mean prediction accuracy (ACC) $\pm$ standard deviation (SD) as well as the area under the ROC curve (AUC).

3.3.1. Hockey Fight dataset

Table 1 shows the violence detection performance of various methods on the Hockey Fight dataset. The results on this dataset using BoW model paired with HOG, HOF and MoSIFT are reported in [5]. Among the BoW based methods, MoSIFT and HOG perform comparably, with a slight improvement over HOF. It indicates the MoSIFT descriptor is discriminative and effective. Our proposed method combines the MoSIFT algorithm and the sparse coding framework. The results show that this method can obtain a higher accuracy than BoW based approaches because the former encodes the local descriptor with less quantization error. The performance is further improved by adding the KDE-based feature selection procedure to our method. Reason for the improvement resides in the fact that the irrelevant and redundant features of MoSIFT are removed while leveraging feature selection, thus contributing to a more descriptive local descriptor.

In this experiment, the number of words in the dictionary of BoW equals to the size of sparse dictionary in sparse coding. With the increase of the size of dictionary, the performance will improve first and then stay stably if the size is large enough. This indicates that an appropriate size of dictionary contributes to both accuracy improvements and computational saving. Besides, the quantization process of BoW is very time consuming especially for large dictionary size. Our sparse coding based method performs much faster when LARS-lasso method is exploited to solve Eq. (2).

3.3.2. Crowd Violence dataset

This dataset is more challenging than the Hockey Fight dataset because it consists of videos in crowded scenes. Table 2 presents the results of various methods on the Crowd Violence dataset. The dictionary size is fixed to 500 in our experiments on this dataset. HOG, HOF and HNF (combination of HOG and HOF) [17] are spatio-temporal descriptors combined with BoW model while LTP and ViF are the approaches based on global representation.

Table 2. Violence detection performance of various algorithms on Crowd Violence dataset (5-fold cross validation)

Method	ACC $\pm$ SD	AUC
HOG + BoW [3]	57.43 $\pm$ 0.37%	0.6182
HOF + BoW [3]	58.53 $\pm$ 0.32%	0.5760
HNF + BoW [3]	56.52 $\pm$ 0.33%	0.5994
LTP [3]	71.53 $\pm$ 0.17%	0.7986
ViF [3]	81.30 $\pm$ 0.21%	0.8500
MoSIFT + BoW	83.42 $\pm$ 8.03%	0.8751
MoSIFT + Sparse Coding	86.60 $\pm$ 3.29%	0.8922
MoSIFT + KDE + Sparse Coding	**89.05 $\pm$ 3.26%**	**0.9357**

Our sparse coding based methods still outperform other approaches despite the challenging nature of this dataset. In this case, MoSIFT descriptor is significantly superior in performance to HOG, HOF and HNF. It proves that MoSIFT is a more effective descriptor for representing action. Consistent with the results on Hockey Fight dataset, MoSIFT combined with the sparse coding method outperforms BoW based method, and employing the KDE-based feature selection effectively facilitates the improvements of classification accuracy. Results on this dataset demonstrate that our method is also effective for detecting violence in crowded scenes.

4. CONCLUSION

This paper proposes an effective violent video detection approach based on the MoSIFT algorithm and the sparse coding scheme. Several procedures have been employed to generate a highly discriminative video representation: 1) MoSIFT captures distinctive local shape and motion patterns of an activity; 2) KDE-based feature selection method selects the most representative features of the MoSIFT descriptor; 3) sparse coding method paired with max pooling procedure generates a discriminative high-level video representation from local features. The proposed method outperforms the state-of-the-art techniques for violence detection in both crowded and non-crowded scenes. It demonstrates the effectiveness of the proposed video feature extraction framework, and whether this video feature can maintain effectiveness in other video analysis tasks is worthy of further research.

Acknowledgements. This research is partly supported by NSFC, China (No: 61273258), Ph.D. Programs Foundation of Ministry of Education of China (No. 20120073110018).

5. REFERENCES

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th ICCV*. IEEE, 2005, vol. 2, pp. 1395–1402.

[2] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proceedings of the 18th CVPR*. IEEE, 2005, vol. 1, pp. 984–989.

[3] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 1–6.

[4] F. D. M. de Souza, G. Ca. chávez, E. A. do Valle, and A. de A Araujo, "Violence detection in video using spatio-temporal features," in *Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2010, pp. 224–230.

[5] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[6] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Workshop on VS-PETS*. IEEE, 2005, pp. 65–72.

[8] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proceedings of the 22nd CVPR*. IEEE, 2009, pp. 1948–1955.

[9] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[10] M.-Y. Chen and A. Hauptmann, "Mosift: Recognizing human actions in surveillance videos," *CMU-CS-09-161. Carnegie Mellon University*, 2009.

[11] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the 22nd CVPR*. IEEE, 2009, pp. 1794–1801.

[12] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proceedings of the 23rd CVPR*. IEEE, 2010, pp. 3517–3524.

[13] B.-D. Liu, Y.-X. Wang, Y.-J. Zhang, and Y. Zheng, "Discriminant sparse coding for image classification," in *Proceedings of the 37th ICASSP*. IEEE, 2012, pp. 2193–2196.

[14] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Proceedings of the 10th ACCV*. Springer, 2011, pp. 660–671.

[15] S. Lu, J. Zhang, Z. Wang, and D. D. Feng, "Fast human action classification and voi localization with enhanced sparse coding," *Journal of Visual Communication and Image Representation*, 2012.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 21st CVPR*. IEEE, 2008, pp. 1–8.

[18] X. Geng and G. Hu, "Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting," *Biomedical Signal Processing and Control*, vol. 7, no. 2, pp. 112–117, 2012.

[19] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th ICML*. ACM, 2009, pp. 689–696.

[22] Y. L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th ICML*. ACM, 2010, pp. 111–118.

[23] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Proceedings of the 12th ICCV*. IEEE, 2009, pp. 492–497.