CLASS-SPECIFIC SPARSE CODES FOR REPRESENTING ACTIVITIES

Sabanadesan Umakanthan, Simon Denman

Image and Video Research Laboratory
Queensland University of Technology
Brisbane, Queensland 4001

Clinton Fookes, Sridha Sridharan

Image and Video Research Laboratory
Queensland University of Technology
Brisbane, Queensland 4001

ABSTRACT

In this paper we investigate the effectiveness of class specific sparse codes in the context of discriminative action classification. The bag-of-words representation is widely used in activity recognition to encode features, and although it yields state-of-the art performance with several feature descriptors it still suffers from large quantization errors and reduces the overall performance. Recently proposed sparse representation methods have been shown to effectively represent features as a linear combination of an over complete dictionary by minimizing the reconstruction error. In contrast to most of the sparse representation methods which focus on Sparse-Reconstruction based Classification (SRC), this paper focuses on a discriminative classification using a SVM by constructing class-specific sparse codes for motion and appearance separately. Experimental results demonstrates that separate motion and appearance specific sparse coefficients provide the most effective and discriminative representation for each class compared to a single class-specific sparse coefficients.

Index Terms— Activity representation, sparse codes, bag-of-words

1. INTRODUCTION

Successfully recognizing human activities from complex video sequences has lots of potential applications such as intelligent video surveillance, automatic video annotation, smart homes and elderly patient monitoring. However, activity recognition faces several challenges due to occlusion, clutter, inter and intra class variations, complex and moving background, camera motion etc. Several activity recognition frameworks have been proposed in literature to address these challenges and improve the overall recognition performance such as Holistic features, space-time templates and tracking interest points in video sequences [1, 2, 3]. These approaches demand high computational requirements and are severely affected by the above mentioned challenges.

Recent approaches using spatio-temporal features demonstrate significant performance improvement compared to other methods. These methods usually incorporate the following procedure: sample the video (densely or sparsely) to detect the interest points, and use spatio-temporal descriptors to encode the appearance and motion information present around the detected points. Then a Bag of Words (BoW) representation is applied to quantize each feature to the closest visual word in the dictionary and video is represented as a histogram of visual word occurrences. Hard vector assignment generates more quantization errors and necessitates the use of a complex kernel function to improve performance. Recently several methods have been proposed to improve the feature representation such as hierarchical dictionaries, spatio-temporal feature modelling, sparse representations etc.

This paper focuses on sparse representation to improve the discriminative action classification. In recent years sparse representation has received lot of attention in a wide range of applications such as image denoising, image restoration, texture classification, face recognition, object recognition and action recognition [4, 5, 6, 7, 8]. Although sparse representation mainly focuses on learning an over complete dictionary to represent the signal with only few elements from the dictionary to minimize the reconstruction error, recently several approaches have been proposed in object recognition that not only minimize the reconstruction error, but also to improve the discriminative power of the sparse coefficients to improve the overall classification performance. Ramirez *et al*. [6] incorporates an incoherence promoting term to make the dictionaries for different classes as independent as possible. Mairal *et al*. [9] proposes to simultaneously learn a classifier by embedding a logistic loss function. Discriminative K-SVD [10] and label consistent K-SVD [11] focused on improving the discriminatory power of the spare codes with a good representation.

Several sparse representation methods have been proposed to solve the action classification problem. Zhu *et al*. [12] introduced sparse representation to classify actions with a shared dictionary with single scale max-pooling and a linear SVM classifier. Guha *et al*. [13] explored shared, class-specific and concatenated dictionaries with different reconstruction error based classification. Sparse Reconstruction-based Classification (SRC) with different features has been

explored in [14, 15, 7]. SRC with L_1 and L_2 regulation (SR-L12) was proposed by Gao *et al*. [16].

Amongst the various sparse coding methods proposed for action recognition, our method differs in two key ways: 1) Unlike other methods, where a single dictionary for a class is built, we build separate dictionaries for motion and appearance features; 2) In this paper we focus on discriminative classification and demonstrate better results compared to the SRC method. In our proposed method we first extract dense Histogram of Gradient (HOG) features to represent appearance and Motion Boundary Histogram (MBH) [3] features to represent motion information at different scales. Then we learn a separate over complete dictionary for appearance and motion vectors to approximately represent them as a weighted sum of sparse coefficients. These appearance and motion sparse coefficient vectors are concatenated and max-pooled separately for each action class at different spatio-temporal scales. A final vector is created by concatenating all the max-pooled sparse coefficients and a linear SVM is used for classification.

The rest of the paper is organized as follows. Section 2 describes our proposed approach in detail and Section 3 presents the experimental setup and results. Finally, Section 4 concludes the paper.

2. CLASS-SPECIFIC DICTIONARY LEARNING

Sparse coding is popularly used to represent a signal as a linear combination of an over complete basis using a few elements of the dictionary. Sparse representation is defined as follows: for a given signal $x \in \mathbb{R}^n$ and a dictionary $D \in \mathbb{R}^{n \times k}$, $min_a \|a\|_0$, s.t. $x = Da$, where $\|a\|_0$ is the l_0 norm of the coefficient vector $a \in \mathbb{R}^k$. *i.e*. minimizing the number of non-zero elements present in the coefficient vector. Since minimizing the l_0 norm is an NP-hard problem and greedy algorithms don't guarantee an optimal solution it is replaced with an l_1 norm and the following optimization problem is solved instead:

$$\min_a \|x - Da\|_2^2 + \lambda |a|_1, \tag{1}$$

where the parameter λ is used to establish balance between the sparsity and reconstruction error. The above optimization problem becomes convex and can be solved using the popular LASSO algorithm. The l_1 norm induces a sparse solution for the code vector a.

In our proposed method, the appearance feature vector $X_A = [x^1, ..., x^m] \in \mathbb{R}^{n_A \times m_A}$, where n_A is the dimension of the appearance vector extracted from a given class and m_A is the number of the appearance vectors, is sparsely represented by minimizing the equation:

$$\min_{D_A, C_A} \|X_A - D_A C_A\|_2^2 + \lambda |C_A|_1 \tag{2}$$

where, the class-specific appearance dictionary is $D_A \in \mathbb{R}^{n_A \times d_A}$ with the size of the dictionary set to d_A and corresponding sparse coefficients are $C_A \in \mathbb{R}^{d_A \times m_A}$. The appearance feature vector x^i can be approximated as $x^i \approx D_A c_A^i$. *i.e*. c_A^i is the sparse coefficient vector corresponding to the appearance feature vector x^i.

Similar to the appearance encoding, the motion vector $Y_M = [y^1, ..., y^m] \in \mathbb{R}^{n_M \times m_M}$, where n_M is the dimension of the motion vector extracted from a given class and m_M is the number of the motion vectors, is sparsely represented by minimizing the equation:

$$\min_{D_M, C_M} \|Y_M - D_M C_M\|_2^2 + \lambda |C_M|_1 \tag{3}$$

where, the class-specific motion dictionary is $D_M \in \mathbb{R}^{n_M \times d_M}$ with the size of the dictionary of d_M and the corresponding sparse coefficients are $C_M \in \mathbb{R}^{d_M \times m_M}$. The motion feature vector y^i can be approximated as $y^i \approx D_M c_M^i$. *i.e*. c_M^i is the sparse coefficient vector corresponding to the motion feature vector y^i.

The final representation of an interest point (I^i) is given by a concatenation of motion (c_M^i) and appearance (c_A^i) sparse coefficient vectors. *i.e*. $I^i = [c_M^i c_A^i]$.

Spatio-temporal characteristics are captured using max-pooling at different spatio-temporal scales, which is shown to be effective with sparse coding [17]. The pooled features at different scales are concatenated to form the final spatio-temporal pyramid representation. Finally, a linear SVM is used for classification.

3. EXPERIMENTS AND RESULTS

In the experiments, we evaluated our proposed representation against the following 3 methods with two popular action recognition datasets with varying complexity: KTH [18] and UCF sports [19].

Sparse Representation-based classification (SRC): The SRC method [8, 20] assigns each feature to the action class based on the reconstruction error: $R(x, D) = \|x - Da\|_2^2$, where $x \in \mathbb{R}^n$ is the feature vector, D is the dictionary and the sparse code vector, $a \in \mathbb{R}^k$, is calculated from Equation 1. For a K class classification problem, each class i has a dictionary D^i and a code a^i is calculated for each dictionary. Finally the feature vector x is assigned to the class i^* which minimizes the reconstruction error R:

$$i^* = \underset{i}{\operatorname{argmin}} R(x, D^i) \tag{4}$$

Shared dictionary with an SVM classifier: A single shared dictionary Φ is learned to sparsely encode each feature vector followed by spatio-temporal max-pooling and a linear SVM classifier is applied for classification.

Experimental setup	Average accuracy (%)
SRC	86%
Shared Dictionary + SVM	92%
Class Dictionary + SVM	94.5%
Our method	96.8%

Table 2: Average Accuracy on the **KTH** Dataset using the four different experimental setups

Class-specific dictionary with SVM classifier: We learn K separate dictionaries $\Phi_1, \Phi_2, ..., \Phi_K$ for each class followed by spatio-temporal max-pooling and linear SVM classification.

In feature extraction we densely sample each video and extract Histogram Oriented Gradients (HOG) and Motion Boundary Histogram (MBH) features to represent each video. For each cell, an 8-bin HOG histogram is calculated and normalised into a HOG descriptor. The robust optical flow based MBH [3] descriptor is used to capture the motion information present in the spatio-temporal volume.

The parameter λ in Equation 2 and 3 controls the sparsity of the the sparse coefficient vector while minimizing the reconstruction error. We set the λ parameter to 10% in all experimental settings which yields better results. We use randomly selected HOG and MBH features from each class to generate the appearance and motion specific dictionaries. Once we learn D_A and D_M for all classes each feature vector is mapped to the sparse coefficient vector.

3.1. KTH Dataset

The KTH dataset consists of 6 different activities such as clapping, boxing, jogging, waving, walking and running performed by 25 subjects under 4 different environmental settings: indoors, outdoors, outdoors with different clothing and outdoor with camera motion. We use the same experimental setting proposed by Schuldt *et al*. [18]. Table 2 shows the average accuracy obtained with 4 different experimental setups. Our proposed sparse representation outperforms the class-specific dictionary by 2.3%. Confusion matrices for class-specific representation and our proposed method are shown in Table 1. Our representation not only performs well across all the classes but also reduces the confusion amongst closely related classes by increasing the discriminatory power.

3.2. UCF Sports Dataset

UCF-Sports dataset [19] consists of approximately 200 videos sequences at a resolution of 720×480. It consists of 9 different action classes such as driving (S1), golf swinging (S2), kicking (S3), lifting (S4), horseback riding (S5),

Experimental setup	Average accuracy (%)
SRC	84%
Shared Dictionary + SVM	87%
Class Dictionary + SVM	89 %
Our method	92.3%

Table 3: Average Accuracy on the **UCF-Sports** Dataset using the four different experimental setups

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1	.97	.00	.00	.00	.03	.00	.00	.00	.00
S2	.00	.93	.03	.00	.00	.00	.00	.04	.00
S3	.03	.00	.90	.00	.00	.00	.00	.00	.07
S4	.00	.00	.00	.90	.00	.00	.00	.00	.10
S5	.00	.00	.00	.00	.82	.05	.00	.00	.13
S6	.00	.00	.00	.00	.00	.82	.05	.00	.13
S7	.00	.00	.12	.00	.00	.00	.88	.00	.00
S8		.05	.08	.00	.00	.00	.00	.87	.00
S9	.00	.05	.00	.00	.00	.10	.00	.00	.85

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1	.99	.00	.00	.00	.01	.00	.00	.00	.00
S2	.00	.96	.00	.00	.00	.00	.00	.00	.04
S3	.00	.00	.94	.00	.00	.00	.00	.04	.02
S4	.00	.00	.00	.96	.00	.00	.00	.00	.04
S5	.00	.00	.00	.00	.85	.00	.07	.00	.08
S6	.00	.00	.02	.00	.00	.88	.00	.00	.1
S7	.00	.00	.03	.00	.00	.00	.94	.00	.03
S8	.00	.05	.02	.00	.00	.00	.00	.93	.00
S9	.00	.05	.00	.00	.00	.05	.00	.00	.90

Table 4: Confusion matrices for the **UCF-Sports** Dataset using class-specific dictionary (Top) and our proposed representation (Bottom).

	Running	Boxing	Walking	Jogging	Waiving	Clapping
Running	0.91	0.00	0.02	0.07	0.00	0.00
Boxing	0.00	0.96	0.00	0.00	0.00	0.04
Walking	0.00	0.00	0.97	0.03	0.00	0.00
Jogging	0.03	0.00	0.04	0.93	0.00	0.00
Waiving	0.00	0.00	0.00	0.00	0.95	0.05
Clapping	0.00	0.03	0.00	0.00	0.02	0.95

	Running	Boxing	Walking	Jogging	Waiving	Clapping
Running	0.92	0.00	0.00	0.08	0.00	0.00
Boxing	0.00	0.98	0.00	0.00	0.00	0.02
Walking	0.00	0.00	1.00	0.00	0.00	0.00
Jogging	0.02	0.00	0.05	0.93	0.00	0.00
Waiving	0.00	0.00	0.00	0.00	1.00	0.00
Clapping	0.00	0.00	0.00	0.00	0.02	0.98

Table 1: Confusion matrices for the **KTH** Dataset using class-specific dictionary (Left) and our proposed representation (Right).

running (S6), skating (S7), swinging (S8) and walking (S9). These videos are obtained from video broadcast television channels such as BBC and ESPN. Unlike the KTH dataset, which is recorded in a static environment, this dataset consists of videos with dynamic and cluttered backgrounds, camera motion, viewpoint changes and illumination changes. We used the Leave-one-out cross validation and average accuracy is reported in Table 3. We obtain an overall classification rate of 92.3% which is 3.3% higher compared to the class-specific dictionary. Confusion matrices for class-specific sparse codes and our proposed method is shown in Table 4.

Experimental results in two datasets demonstrate that class-specific dictionaries provide a better, sparse and discriminative representation for their own class compared to other classes. Further, the shared nature of the class-specific appearance and motion dictionaries allow other classes to effectively capture common spatio-temporal elements present in their action sequences. For example, some atoms in the motion dictionary built for the running class can be used to represent temporal elements of walking or jogging class and atoms in the appearance dictionary built for the boxing class can be used to represent some spatial elements in hand clapping class. This rich dictionary structure allows us to focus on and capture minor spatio-temporal elements which are important to differentiate between two closely related classes.

We achieve an overall performance improvement without adding any additional terms in the optimization function. The availability of parallel processing hardware will allow us to build appearance and motion specific dictionaries simultaneously. Therefore the computational requirement is the same as building a class-specific dictionary with combined motion and appearance features.

4. CONCLUSION

In this paper, we have presented an efficient way of constructing a sparse dictionary to represent activities for discriminative action classification. Experimental results on two popular datasets with varying complexity demonstrate that building separate appearance and motion specific dictionaries for each class significantly improves the classification performance compared to a shared dictionary and class-specific dictionary. In addition our representation adds more discriminative power to the video representation and can be extended to different video based applications.

In future we plan to explore our sparse feature representation method with different generative and discriminatory classification schemes.

5. REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[2] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Underst.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.

[3] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[4] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[5] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[6] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3501–3508.

[7] Kai Guo, Prakash Ishwar, and Janusz Konrad, "Action recognition using sparse representation on covariance manifolds of optical flow," in *Advanced Video and*

Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on. IEEE, 2010, pp. 188–195.

[8] Gabriel Peyré, "Sparse modeling of textures," *Journal of Mathematical Imaging and Vision*, vol. 34, no. 1, pp. 17–31, 2009.

[9] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[10] Qiang Zhang and Baoxin Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[11] Zhuolin Jiang, Zhe Lin, and Larry S Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1697–1704.

[12] Yan Zhu, Xu Zhao, Yun Fu, and Yuncai Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Computer Vision–ACCV 2010*, pp. 660–671. Springer, 2011.

[13] Tanaya Guha and Rabab K Ward, "Learning sparse representations for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1576–1588, 2012.

[14] Changhong Liu, Yang Yang, and Yong Chen, "Human action recognition using sparse representation," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*. IEEE, 2009, vol. 4, pp. 184–188.

[15] Anan Liu and Dong Han, "Spatiotemporal sparsity induced similarity measure for human action recognition.," *JDCTA*, vol. 4, no. 8, pp. 143–149, 2010.

[16] Zan Gao, An-An Liu, Hua Zhang, Guang ping Xu, and Yan bing Xue, "Human action recognition based on sparse representation induced by l1/l2 regulations," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, Nov 2012, pp. 1868–1871.

[17] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[18] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[19] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[20] Allen Y. Yang, John Wright, Yi Ma, and S. Shankar Sastry, "Feature selection in face recognition: A sparse representation perspective," Tech. Rep., 2007.