IEICE TRANS. INF. & SYST., VOL.E97–D, NO.3 MARCH 2014

LETTER

Topic-Based Knowledge Transfer Algorithm for Cross-View Action Recognition

Changhong CHEN[†a)], **Shunqing YANG**[†b)], *Nonmembers*, *and* **Zongliang GAN**[†c)], *Member*

SUMMARY Cross-view action recognition is a challenging research field for human motion analysis. Appearance-based features are not credible if the viewpoint changes. In this paper, a new framework is proposed for cross-view action recognition by topic based knowledge transfer. First, Spatio-temporal descriptors are extracted from the action videos and each video is modeled by a bag of visual words (BoVW) based on the codebook constructed by the k-means cluster algorithm. Second, Latent Dirichlet Allocation (LDA) is employed to assign topics for the BoVW representation. The topic distribution of visual words (ToVW) is normalized and taken to be the feature vector. Third, in order to bridge different views, we transform ToVW into bilingual ToVW by constructing bilingual dictionaries, which guarantee that the same action has the same representation from different views. We demonstrate the effectiveness of the proposed algorithm on the IXMAS multi-view dataset.

key words: *cross-view human action recognition, topic distribution of visual words, latent dirichlet allocation, bilingual dictionary*

1. Introduction

Human action recognition aims to recognize the actions of one or more agents from a series of observations. Local representations [1], [2] have been widely used in human action recognition. However, most local representations are based on appearance. Appearance-based features perform well in recognizing actions with limited view variations, but tend to be powerless against large view variations. The major reason lies in the obvious changes in the appearance of actions under different viewpoints. It is difficult even for people to recognize them when the viewpoints change greatly, such as the horizontal view and top view. As a result, most local representations become less discriminative for cross-view action recognition.

To deal with this problem, the techniques at present can be divided into two kinds. The first kind seeks the geometric properties, which remain stable under view changes. A typical algorithm is proposed by Junejo et al. [3]. A simple and interesting action representation called self-similarity descriptors is presented in this paper. It is found to be highly stable under view changes. This method is effective in most cases. However, the experimental results show the approach performs poorly when the top view as source (training) or target (testing) view. The second kind is based on transfer learning, which aims to construct view-stable and discriminative features. Given a pair of views, the features are learned from them and a bridge between them is constructed. For a new action class observed in one view, transfer learning enables recognition in the second view through the bridge. The most encouraging method is proposed by Zheng et al. [4], who learn the two dictionaries of source and target views simultaneously to ensure the same action have the same representation. Although these transfer learning algorithms achieve good performance, they are harder to transfer actions for view combinations that involves the top view.

Transfer learning provides a good choice for cross-view recognition. However, the feature representation is also an important factor determining the performance of the transfer learning algorithm. Inspired by the successful application of topic models in unsupervised recognition [1] and scene categorization [5], we propose a new algorithm for cross view action recognition in this study. The major contributions lie in: (1) A topic-based knowledge transfer algorithm combined with topic distribution of visual words (ToVW) representation and knowledge transfer is proposed, which bridges the semantic gap across view-dependent vocabularies while preserving discrimination among action categories. (2) The proposed algorithm is effective for view combinations that involve the top view.

2. Feature Extraction

Sparse spatio-temporal features are firstly extracted from the action video. Among various interest point detection methods, the one proposed by Dollar et al. [6] is the most widely used. However, it uses local information within a small region and tends to generate spurious detection in background areas. In this paper, we adopt the algorithm proposed by [7], which overcomes the shortcomings of the Dollar detector.

2D Gabor filter of the selected orientation is used on the frame difference. The eight Gabor filters are applied separately and eight different responses are computed at each frame. The interest points are detected when the total response is larger than the predetermined threshold value, the maximum number of extracted interest points in each frame is 20. Most interest points are located on the moving parts. For different views, their locations and their distributions are different. The aim of our work is to make up the influence of the difference. Then fixed size spatio-temporal cuboid is

Manuscript received September 12, 2013.
Manuscript revised November 7, 2013.

[†]The authors are with Jiangsu Key Laboratory of Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing, China.
a) E-mail: chenchh@njupt.edu.cn
b) E-mail: 1012010536@njupt.edu.cn
c) E-mail: ganzl@njupt.edu.cn

DOI: 10.1587/transinf.E97.D.614

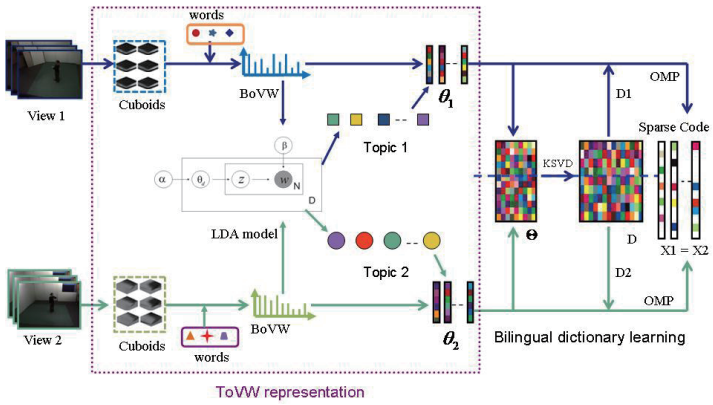

Fig. 1 The process of the topic-based knowledge transfer algorithm.

extracted around the interest points, the size of the cuboid is set to 13*13*19. Gradient-based 100-dimensional descriptor is obtained by PCA for each cuboid.

3. Topic-Based Knowledge Transfer

The process of the topic-based knowledge transfer algorithm is illustrated in Fig. 1. First, Spatio-temporal descriptors are extracted from the action videos and each video is modeled by a bag of visual words (BoVW) based on the codebook constructed by the k-means cluster algorithm. Second, Latent Dirichlet Allocation (LDA) is employed to assign topics for the BoVW representation. The topic distribution of visual words (ToVW) is normalized and taken to be the feature vector. Third, in order to bridge different views, we transform ToVW into bilingual ToVW through transferable dictionary pairs. Bilingual dictionaries are learned for the source and target view, which guarantee that the same action has the same high-level representation from different views. We will introduce the proposed algorithm in detail from ToVW representation formulation and bilingual dictionary construction.

3.1 ToVW Representation

Latent topic models have been successfully applied to scene categorization [5] and human action recognition [1]. In [5], an image is represented by BoVW representation and latent topic models in employed for representing of the distribution of the codewords. The model that best fitting the distribution of the codewords of the test image is founded during recognition. Inspired by the algorithm for scene categorization in [5], we introduce a ToVW representation for video description. In our work, an action video is represented by BoVW representation. Instead of building topic models for recognition directly, LDA model is employed for founding the topic distribution of the codewords, which helps to transform the BoVW representation to the ToVW representation. The process of building the ToVW representation is summarized as follows:

(1) K-means cluster algorithm is employed for the sparse representation of the spatio-temporal descriptors F. The cluster number is chosen as 1000. The cluster centers are saved as the codebook and each descriptor can be represented as a cluster label. Each video is modeled by a BoVW representation.

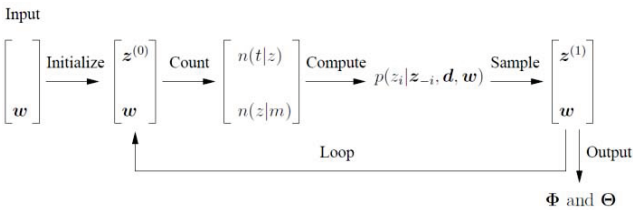

Fig. 2 The procedure of learning LDA by Gibbs sampling [8].

(2) The labels of the descriptor are defined as words and the videos are regarded as documents. The documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The topic of each descriptor is assigned by the LDA. The topic number is chosen as 50. Gibbs sampling is employed for learning LDA as shown in Fig. 2. The topics of the words are initialized randomly. Then the number of terms appearing in each topic $n(t|z)$ and the number of topics appearing in every document $n(t|z)$ are counted. The term t is the element of the codebook and can not be repeated, while word is the element of the document and can be repeated. The computing of $p(z_i|z_{-i}, d, w)$ is to exclude the topic distribution of current word. After obtaining the topic distribution of current word w, a new topic $z^{(1)}$ can be sampled from it. The same method is repeated for updating the topic of next word until the topic distribution of each document θ and the word distribution of each topic ϕ converges. The topic distribution vector θ is normalized to be the ToVW representation. Besides the updated topics, the distribution ϕ is also saved to predict topics for probe videos.

3.2 Bilingual Dictionary Construction

In order to bridge the ToVW representations from two viewpoints, we construct the bilingual dictionary through knowl-

Table 1 Recognition performance of the proposed algorithm. The numbers in the bracket are the average recognition performance of (ours(one), ours(four), [4]).

	Cam0	Cam1	Cam2	Cam3	Cam4
Cam0		(93.7, 99.3, 98.8)	(96.8, 99.6, 99.1)	(98.1, 99.3, 99.4)	(98.2, 99.6, 92.7)
Cam1	(94.9, 99.3, 98.8)		(97.0, 99.6, 99.7)	(98.3, 99.3, 92.7)	(97.7, 99.6, 90.6)
Cam2	(91.7, 99.3, 99.4)	(91.1, 99.3, 96.4)		(98.1, 99.4, 97.3)	(98.0, 99.8, 95.5)
Cam3	(81.8, 99.3, 98.2)	(82.1, 99.3, 97.6)	(89.8, 99.7, 99.7)		(92.5, 99.8, 90.0)
Cam4	(88.7, 99.3, 85.8)	(87.8, 99.3, 81.5)	(94.1, 99.7, 93.3)	(96.7, 99.4, 83.9)	
Aver.	(89.2, 99.3, 95.5)	(97.0, 99.3, 93.6)	(94.6, 99.6, 98.0)	(86.7, 99.3, 93.3)	(91.8, 99.7, 92.4)

edge transfer. We employ the transformable dictionary pair representation [4] for the bilingual dictionary formulation. Suppose we have N video sequences shared in the source and target views. θ_s and θ_t are the ToVW representations of them. Our goal is to find a sparse representation X, which can denotes θ_s and θ_t simultaneously. It can be realized by designing bilingual dictionaries D_s and D_t for the source and target views, which can be realized by:

$$\arg\min_{D_s,D_t,X} \|\theta_s - D_s X\|_2^2 + \|\theta_t - D_t X\|_2^2 \quad s.t. \forall i, \|x_i\|_0 \le s \qquad (1)$$

where the first term is the reconstruction error of the source view and the second term is that of the target view. x_i is the i^{th} representation of X and $\|x_i\|_0 \le s$ is the sparse constraint. The bilingual dictionaries $\{D_s, D_t\}$ can be learned using the K-SVD algorithm [9], as shown in Fig. 1. Sparse representations of θ_s and θ_t are obtained by solving the following optimization problem through the orthogonal matching pursuit (OMP) algorithm [10].

$$(D, X) = \arg\min \|Y - DX\|_2^2 \quad s.t. \forall i, \|x_i\|_0 \le s \qquad (2)$$

As a result, the same actions from two viewpoints can be represented by the same vectors.

4. Experiments and Discussion

4.1 Data Set and Experimental Setup

In this section we evaluate our algorithm with the IXMAS multi-view action data set [11], which contains 12 daily-live actions performed each 3 times by 12 actors taken from 5 different views: four side views and one top view. For an accurate comparison to [4], [12] and [13], we adapt the leave-one-action-class-out strategy for the experiments, which means that each time we only consider one action class for testing in the target view.

We construct the bilingual dictionary for each viewpoint pair with 11 actions of 1 actor and 4 actors, separately. The orphan action class of the other actors is used for testing. 12 set of experiments are carried out for the former circumstance and 495 set of experiments for the later one. The features extracted from the orphan action are clustered according to the distances between them and the trained cluster centers of corresponding view. The labels of the descriptors are regarded as the sparse representation. The topics of the sparse representation are predicted by the trained LDA. The ToVW representation is transferred through the bilingual dictionary and is used for training the SVM classification model. Radial basis function kernel are chosen for the SVM model and its parameters are set as: $C = 512$ and $\gamma = 0.5$. The classification accuracy is reported by averaging over all possible combinations for the selecting orphan actions.

4.2 Experimental Results

In this section, we show the experimental results of our proposed algorithm. The recognition results of our algorithm with one actor and four actors used for constructing the dictionary for all 20 combinations of the training and testing views are illustrated in Table 1. The numbers in the bracket are the average recognition performance of our algorithm with one actor (ours(one)), our algorithm with four actors (ours(four)) and [4]. The first column represents the target views and the first row represents the source views.

It can be seen that the performance of ours(one) is not stable. We check the recognition accuracy of the different action and find that the highest recognition rate of the action "hands up" is 97.0% and the lowest one only 27.3% when the first person are used for constructing the bilingual dictionary. The recognition results of some other actions also have the same circumstance, which is not shown in the paper because of the limitation of pages. We deduce that the major reason for the instability lies in that some action of some people is not standard. In order to validate this conclusion, we do experiments with 4 actors constructing the dictionary and our algorithm yields much better performance in all cases. The average recognition accuracy achieve 99%. When Cam4 is the source or target view, the recognition accuracy is a little lower than other combinations of pairwise views in [4], but the performance is little effected by the viewpoint changes and excellent results are also obtained for Cam4 in our work. Our algorithm tends to be more effective and credible than that of [4], The major reason lies in the ToVW representation is more suitable for action of different views than the BoVW used in [4].

The average recognition performance comparison with [4], [12] and [13] is shown in Fig. 3. Although easy to be affected by the view changes, the performance of our algorithm with one actor (ours(one)) is better than those of [12] and [13]. When training the SVM model with four actors (ours (four)), our algorithm is stable and achieves more than 99% recognition accuracy. That is to say, our algorithm is

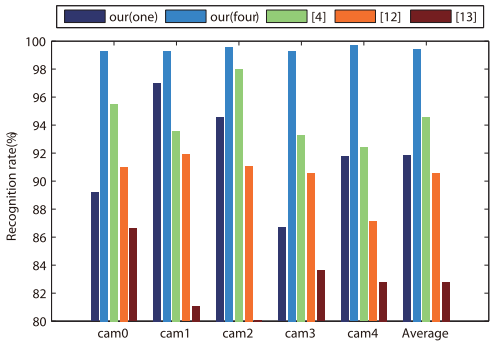

Fig. 3 Average recognition performance.

credible if given enough training samples. Compared with the other three transfer learning algorithms in [4], [12] and [13], our approach yields much better performance in all cases. Especially for the cases where Cam4 is the source or target view, the recognition accuracy is comparable as other combinations of view pairs. It can be concluded that our proposed algorithm is credible for view variation.

Besides the algorithms, the recognition performance is also deeply effected by the choice of low-level features. The same low-level action descriptors of [4], [12] and [13] are employed. The action is represented by a complementary combination of local and global features, which are BoVW representation based on interest point descriptors [6] and shape-flow descriptors [14], respectively. Our low-level features are BoVW representation based on interest point descriptors [7]. [6] uses local information within a small region and tends to generate spurious detection in background areas, which affects the performance of the low-level action descriptors of [4], [12] and [13].

5. Conclusion

This paper propose a new topic-based knowledge transfer framework for cross-view action recognition, which bridges the semantic gap across view-dependent vocabularies while preserving discrimination among action categories. Topic distribution of visual words (ToVW) representation is brought forward to represent the action video and the bilingual dictionary is built for each view pair. The recognition performance on the public IXMAS multi-view dataset illustrates the effective of our algorithm. Although the algorithm achieved the average recognition rate of 99%, we can not simply conclude that our algorithm is practical because the size of the current public database is limited. We will further evaluate our algorithm on a database with more people and more actions.

Acknowledgements

This research was sponsored by the NSF of Jiangsu Province (BK2010523), the NSFs of China (Nos.61172118, 61071166, 61071091 and 61001152), the University Natural Science Research Project of Jiangsu Province (11KJB510012) and the Scientific Research Foundation of Nanjing University of Posts and Telecommunications (NY210073).

References

[1] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," Int. J. Comput. Vis., vol.79, no.3, pp.299–318, 2008.
[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," IEEE Conference on Computer Vision and Pattern Recognition, pp.1–8, 2008.
[3] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.1, pp.172–185, 2011.
[4] J. Zheng, Z. Jiang, J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," British Machine Vision Conference, 2012.
[5] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.524–531, 2005.
[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, 2005.
[7] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," Pattern Recognit., pp.1220–1234, 2012.
[8] Y. Wang, "Distributed gibbs sampling of latent topic models: The gritty details," Tech. Rep., 2008.
[9] J. Zeng, W. Cheung, and J. Liu, "K -svd: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Process., vol.54, no.11, pp.4311–4322, 2006.
[10] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," IEEE Trans. Inf. Theory, vol.53, no.12, pp.4655–4666, 2007.
[11] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," IEEE 11th International Conference on Computer Vision, 2007, ICCV 2007, pp.1–7, 2007.
[12] B. Li, O.I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[13] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[14] T. Du and S. Alexander, "Human activity recognition with metric learning," Computer Vision C ECCV 2008, pp.548–561, 2008.