

Research Article

Extract the Relational Information of Static Features and Motion Features for Human Activities Recognition in Videos

Li Yao^{1,2}

¹Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, Jiangsu Province, China

²State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu Province, China

Correspondence should be addressed to Li Yao; yao.li@seu.edu.cn

Received 29 March 2016; Revised 14 July 2016; Accepted 7 August 2016

Academic Editor: Hong Man

Copyright © 2016 Li Yao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Both static features and motion features have shown promising performance in human activities recognition task. However, the information included in these features is insufficient for complex human activities. In this paper, we propose extracting relational information of static features and motion features for human activities recognition. The videos are represented by a classical Bag-of-Word (BoW) model which is useful in many works. To get a compact and discriminative codebook with small dimension, we employ the divisive algorithm based on KL-divergence to reconstruct the codebook. After that, to further capture strong relational information, we construct a bipartite graph to model the relationship between words of different feature set. Then we use a k -way partition to create a new codebook in which similar words are getting together. With this new codebook, videos can be represented by a new BoW vector with strong relational information. Moreover, we propose a method to compute new clusters from the divisive algorithm's projective function. We test our work on the several datasets and obtain very promising results.

1. Introduction

Recognizing human activities in video automatically is a promising technology in computer vision. There are a lot of application scenarios for it, such as content-based video retrieval, intelligent video surveillance, human-computer interaction, and e-health. Although lots of researchers have paid attention to this problem, it remains challenging to recognize human activities in the videos because of the great variance caused by illumination change, camera motion, and background cluster and so on.

To recognize the human activities pattern from massive videos, researchers extract discriminate features for further process. Below, we briefly review some works based on static local features and motion features.

Regarding local static features, Lowe [1] proposed a descriptor which is invariant to translations, rotations, and scaling transformation. This descriptor detects interest points from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a

local neighborhood around each interest point. In this paper, we use a dense version of SIFT descriptor which has been proven to be useful for tasks such as object categorization, texture classification, image alignment, and biometrics [2]. On the other hand, to make use of color information in image, color-based sift descriptor has been proposed [3].

As for local motion features, to capture temporal information in videos, Chen and Hauptmann [4] proposed a Mo-SIFT descriptor that detects interest points and encodes their local appearance and explicitly models local motion. Wang et al. [5] proposed an approach to describe trajectories densely. Laptev et al. [6] proposed a STIP descriptor that computed each interest point's descriptors of the associated space-time patch.

However, local static features and local motion features only contain partial information of human activities in video. Moreover, we believe that local motion and static features are complementary for action recognition in unrestricted videos [7]. Researchers have paid attention to fusion multimodality for getting complementary information [8]. As a similar work to this paper, Liu et al. [7] first extract local static feature and

local motion feature from videos. After that, they use static information to prune motion features and use Page-Rank to prune local static features. Moreover, they employ the divisive algorithm based on KL-divergence for code word clustering. And features are mixed up during the classification phrase.

A low dimension BoW may lose important relation information, while a high dimension may lead to the curse of dimensionality. So, reconstructing original codebook to a smaller dimension with less lost information is necessary. To solve this problem, Fulkerson et al. [9] used Informational Bottleneck to obtain meaningful feature clusters. And Pereira [10, 11] used distributional cluster words/features. Each word cluster can then be treated as a single feature and thus dimensionality can be drastically reduced.

In our work, we use a bimodel to capture hybrid features before the classification phrase. We do it with two reasons: (1) capturing relational information in a direct way and (2) reducing the BoW dimensionality apparently.

As our first contribution, we employed the divisive algorithm based on KL-divergence. This algorithm uses divergence to generate an information loss criterion and is implemented iteratively like k -means. This makes it possible to capture compact and discriminative codebook with smaller dimension effectively and efficiently.

As our second contribution, we initiate employment of bimodel to get hybrid feature representation. We construct a bipartite graph to model the relationship between codebooks of different feature set. Then, we use a k -way partition to get a new codebook. With this new codebook, videos can be represented by a new BoW vector with strong relational information. A similar work [12] employs bimodel to get joint audiovisual codebook. The bimodel needs the clusters. We propose a method to compute new clusters from the divisive algorithm's output, while Ye et al. [12] generate the new clusters directly.

2. Methods

Figure 1 is the flowchart of the proposed system. In this paper, we propose using several technologies for better exploiting relational information of static feature and motion feature for human activities task.

2.1. Features Extraction and BoW Model

2.1.1. Local Static Features. We detect local interest points by a Harris-Laplace detector densely and use a SIFT descriptor to encode these points. SIFT is invariant to rotation, scale, and light change. Moreover, the dense SIFT has been proven to be useful for tasks such as object categorization, texture classification, image alignment, and biometrics [2].

2.1.2. Local Motion Features. In our work, we employ the Dense Trajectory descriptor [5] as original motion feature. Dense Trajectory firstly samples points in different spatial scale densely. After that, tracking is performed in the corresponding spatial scale. Finally, descriptors are computed along the trajectory. In this paper, we simply use the default parameters for feature extraction.

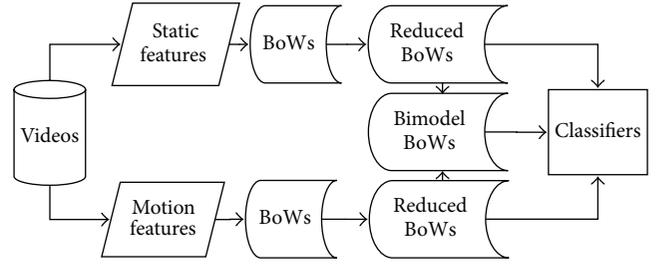


FIGURE 1: Flow chart of proposed method.

2.1.3. BoW Model. BoW model has been widely used in many works and has been shown to be efficient in many tasks. This model clusters all features to several clusters and uses these clusters to discrete features from a video. Although BoW is effective and efficient, it may lose information with a low dimension or lead to curse of dimensionality with a high dimension. So, we will detail our codebook reconstruction technology in the next section for this problem.

2.2. Codebook Reconstruction. Although BoW model is efficient in many computer visual tasks, it has two obvious drawbacks. First, it is inevitable to lose information with low dimension; second, it may lead to the curse of dimensionality with high dimension. So, in this paper, we use a two-phrase procedure to get the BoW representation. Firstly, we use k -means to get a large codebook. Then, a divisive algorithm based on KL-divergence is employed to reconstruct the initial codebook to a small and discriminative codebook. To make use of bimodel, we propose a method to compute new clusters from projective function of original words.

2.2.1. Compute Original Words' Projective Function. Suppose $p_1(x)$ and $p_2(x)$ are probability distributions taken from random variable X . The Kullback-Leibler (KL) divergence between $p_1(x)$ and $p_2(x)$ is defined as

$$\text{KL}(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}. \quad (1)$$

On the other hand, the Jensen-Shannon (JS) divergence is defined as

$$\text{JS}_\pi(p_1, p_2) = \sum_{i=1,2} \pi_i \text{KL} \left(p_i, \sum_{j=1,2} \pi_j p_j \right), \quad (2)$$

where $\pi_1 + \pi_2 = 1$.

Let $C = (C_1, \dots, C_m)$ represent activity classes, and $X = (x_1, \dots, x_n)$ represent the original codebook. Then the information about C captured by X can be measured by mutual information $I(C; X)$. Suppose $Y = (y_1, \dots, y_k)$ is the new codebook we get; then we can measure the quality of the new codebook by the loss of MI, which is defined as

$$\begin{aligned} Q(Y) &= I(C; X) - I(C; Y) \\ &= \sum_{i=1}^k \pi(y_i) \text{JS}(\{p(C | x_t) : x_t \in y_i\}), \end{aligned} \quad (3)$$

where $\pi(y_i) = \sum_{x_t \in y_i} \pi_t$, $\pi_t = p(x_t)$. And, after some derivation, we can rewrite Q as follows:

$$Q(Y) = \sum_{i=1}^k \pi(y_i) \sum_{x_t \in y_i} \pi_t * \text{KL}(p(C | x_t), p(C | y_i)). \quad (4)$$

In this paper, we can use an iterative procedure like k -means algorithm to obtain the optimal new vocabulary using five major steps as follows:

- (1) Perform initialization: for every original word x_t , assign it to y_j ($1 \leq j \leq |C|$) with $j = \max(p(C_j | x_t))$. After that, we get $|C|$ initial word clusters. And then, each cluster is split to several groups which result in k initial clusters, say $Y = (y_1, \dots, y_k)$.

- (2) For each y_j , compute

$$\begin{aligned} \pi(y_j) &= \sum_{x_t \in y_j} \pi_t, \\ P(C | y_j) &= \sum_{x_t \in y_j} \frac{\pi_t}{\pi(y_j)} p(C | x_t). \end{aligned} \quad (5)$$

- (3) For each original word x_t , assign it to another cluster y_j , where y_j is

$$j^*(x_t) = \underset{i}{\text{argmin}} \text{KL}(p(C | x_t), p(C | y_i)). \quad (6)$$

- (4) Compute the measurement Q :

$$Q(Y) = \sum_{i=1}^k \pi(y_i) \sum_{x_t \in y_i} \pi_t * \text{KL}(p(C | x_t), p(C | y_i)). \quad (7)$$

- (5) If $Q < 10^{-3}$, exit the iteration; otherwise repeat steps (2) to (4).

We now discuss the computational complexity of our algorithm. Step (3) of each iteration requires KL-divergence to be computed for every pair, $p(C_j, x_t)$ and $p(C_j, y_i)$. This is the most computationally demanding task and costs a total of $O(mkl)$ operations. Moreover, it can be proven that the objective function decreases at every iteration. So, the total time complexity is $O(mklt)$, where t is the number of iterations.

2.2.2. Compute New Clusters. Given the projection from original words to new words, we need to compute the clusters of the new words. To be specific, let $Y = (y_1, \dots, y_k)$ be the new clusters; for each y_i in Y , we have

$$y_i = \frac{\sum_{x_t \in y_i} (x_t \times \sum_{j=1}^n d(j, t))}{\sum_{x_t \in y_i} \sum_{j=1}^n d(j, t)}, \quad (8)$$

where n is the number of training videos. And $d(j, t)$ represent the entry t of the j th video.

2.3. Bimodel Based Relational Information Extraction. Given two codebooks extract from local static feature and local motion feature, we need to generate a new codebook which has as more relational information as possible. In this paper, we propose using a bimodel for this problem. Bimodel has been applied to IR [13] and Cross-View Action Recognition [14] successfully. To further capture strong relational information, we construct a bipartite graph to model the relationship between codebooks of different feature set. Finally, we use a k -way partition to get a new codebook. With this new codebook, videos can be represented by a new BoW vector with strong relational information.

2.3.1. Construct a Bipartite Graph. In this section, we detail how to model the relationship between two words from two codebooks. Suppose we have n training videos named $D_{\text{tr}}^{\text{sta}} = \{(h_i^{\text{sta}}, l_i)\}_{i=1}^n$ and $D_{\text{tr}}^{\text{mot}} = \{(h_i^{\text{mot}}, l_i)\}_{i=1}^n$ captured from static feature and motion feature individually. Let $W^{\text{sta}} = \{w_i^{\text{sta}}\}_{i=1}^{m_{\text{sta}}}$ and $W^{\text{mot}} = \{w_i^{\text{mot}}\}_{i=1}^{m_{\text{mot}}}$ be the codebooks of static feature set and motion feature set individually. We can construct a graph $G = (V, E)$, where V and E represent the vertices and edges, respectively. To be specific, as G is a bipartite graph, $V = V^{\text{sta}} \cup V^{\text{mot}}$, where each vertex in V^{sta} corresponds to a static word in W^{sta} and each vertex in V^{mot} corresponds to a motion word in W^{mot} . Moreover, each edge in E only connects the vertices between V^{sta} and V^{mot} . The weight matrix of E can be defined as $E = \begin{pmatrix} 0 & S \\ S^T & 0 \end{pmatrix}$, where S is a $|V^{\text{sta}}| \times |V^{\text{mot}}|$ matrix representing the similarity between any pair of words from two codebooks. In this paper, we use a measurement like TF-IDF to measure the similarity. To be specific, each element s_{kl} of S is defined as follows:

$$s_{kl} = \frac{\sum_{i=1}^n h_i^{\text{sta}}(k) h_i^{\text{mot}}(l)}{\sum_{i=1}^n h_i^{\text{sta}}(k) \sum_{i=1}^n h_i^{\text{mot}}(l)}, \quad (9)$$

where $h_i^{\text{sta}}(k)$ denotes the entry of h_i^{sta} corresponding to static words w_k^{sta} and $h_i^{\text{mot}}(l)$ denotes the entry of h_i^{mot} corresponding to motion words w_l^{mot} .

2.3.2. Discover Bimodel Words. After obtaining the bipartite graph between static feature codebook and motion feature codebook, we present the detail of bimodel words discovery.

(1) Graph Bipartitioning. Given a bipartite graph $G = (V, E)$, bipartitioning is to partition V into two subsets, where vertices in the same subset have strong relation and vertices in the different subset have weak relation. Formally, graph bipartitioning aims at minimizing the following objective function:

$$\text{cut}(V_1, V_2) = \sum_{i \in V_1, j \in V_2} s_{ij}. \quad (10)$$

(2) Efficient k -Way Solution. Actually, finding bipartitioning of bigraph can be understood as classifying each point into two classes, for example, +1 and -1. Suppose q_i is the projection value of vertices i ; good bipartitioning minimized $(1/4) \sum_{(i,j) \in E} e_{ij} \times (q_i - q_j)^2$. However, this may lead to a wrong solution that assigns all vertices to +1 or -1. So, in this

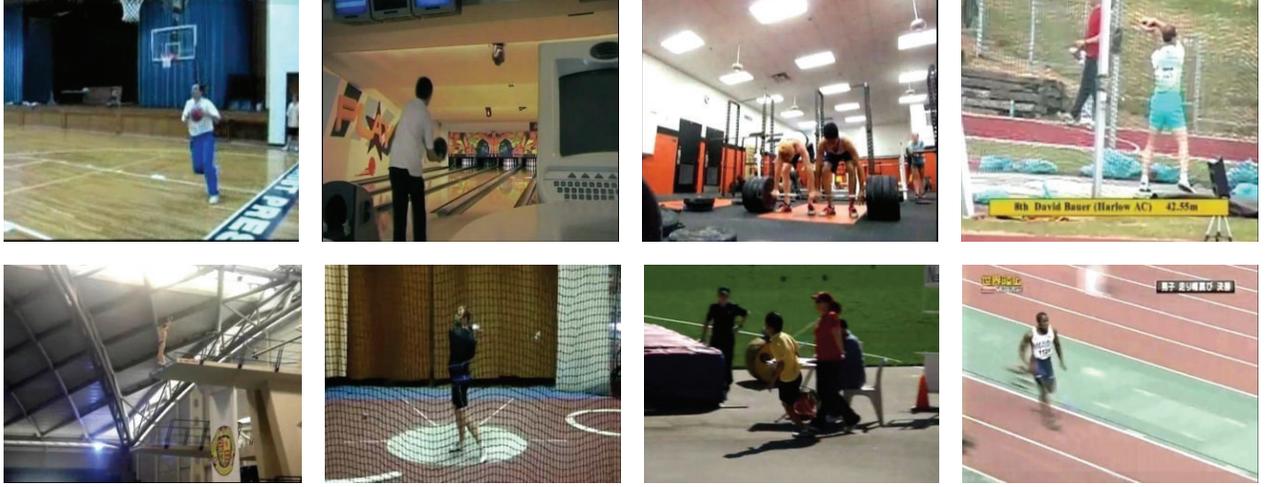


FIGURE 2: Screenshots of Olympic dataset.

paper, we are actually looking for a balanced partition whose objective function looks like the following:

$$\text{Balance Cut}(V_1, V_2) = \frac{\text{cut}(V_1, V_2)}{\sum_{i \in V_1} \sum_j e_{ij}} + \frac{\text{cut}(V_1, V_2)}{\sum_{i \in V_2} \sum_j e_{ij}}. \quad (11)$$

This problem can be solved by spectral clustering, which first constructs a Laplace matrix L as follows:

$$L(i, j) = \begin{cases} -e_{ij} & e_{ij} \in E \\ \sum_k e_{ik} & i = j \\ 0 & \text{else.} \end{cases} \quad (12)$$

After that, bipartitioning of G can be provided by the second smallest eigenvector of the generalized eigenvalue problem $Lz = \lambda Dz$, where $D(i, i) = \sum_j e_{ij}$.

However, as an efficient solution proposed in [13], we can get optimal bipartitioning without computational complex. Suppose we have a matrix L , where $D_1^{\text{sta}}(i, i) = \sum_j e_{ij}$ and $D_2^{\text{mot}}(i, i) = \sum_j e_{ij}$, as follows:

$$\begin{aligned} L(i, j) &= \begin{pmatrix} D_1^{\text{sta}} & -S \\ -S^T & D_2^{\text{mot}} \end{pmatrix} \\ &= \begin{pmatrix} D_1^{\text{sta}} & 0 \\ 0 & D_2^{\text{mot}} \end{pmatrix} + \begin{pmatrix} 0 & -S \\ -S^T & 0 \end{pmatrix}. \end{aligned} \quad (13)$$

Let $S = D_1^{\text{sta}-1/2} S D_2^{\text{mot}-1/2}$; it can be proven that the second eigenvector of L can be expressed in terms of left and right singular vectors (say u_2 and v_2) of S as follows:

$$z_2 = \begin{pmatrix} D_1^{\text{sta}-1/2} u_2 \\ D_2^{\text{mot}-1/2} v_2 \end{pmatrix}. \quad (14)$$

In a general scene, suppose we need to capture k new words containing relational information; the optimal k -way

partitioning solution is provided by the $l = \lceil \log k \rceil$ singular vectors $U = (u_2, \dots, u_{l+1})$ and $V = (v_2, \dots, v_{l+1})$.

To be specific, let $Z = (D_1^{\text{sta}-1/2} U, D_2^{\text{mot}-1/2} V)^T$; we look for k clusters of row space in Z such that the sum of squares $\sum_{i=1}^k \sum_j \text{distance}(i, j)$ is minimized.

Thus, our bimodel based clustering algorithm can be summarized as five basic steps as follows:

Input: n training videos.

- (1) Construct bipartite graph, where each element of S is computed as formula (9).
- (2) Compute matrices D_1^{sta} and D_2^{mot} and $S = D_1^{\text{sta}-1/2} S D_2^{\text{mot}-1/2}$.
- (3) Apply SVD on S ; compute U and V .
- (4) Compute matrix Z , whose size is $(|V^{\text{sta}}| + |V^{\text{dyn}}|) \times l$.
- (5) Run k -means on row vectors of matrix Z to get k clusters.

Output: k clusters.

With k new clusters, each video can be represented as a new BoW vector which contains relational information.

3. Experiment and Analysis

3.1. Experiment on Olympic Dataset. The Olympic dataset (Figure 2) contains videos of athletes practicing different sports [15]. As all the videos are crawled from YouTube, it means that there are little artificial constraints which make the human activities recognition hard. There are 16 sports including high jump, long jump, and basketball. In our experiment, we use the default solution to split the training videos and testing videos. Figure 2 shows some screenshots from this dataset.

We use the package provided by [3] to extract dense SIFT features. For each video, we extract SIFT from the densely sampled grid with default parameters. In our experiment,

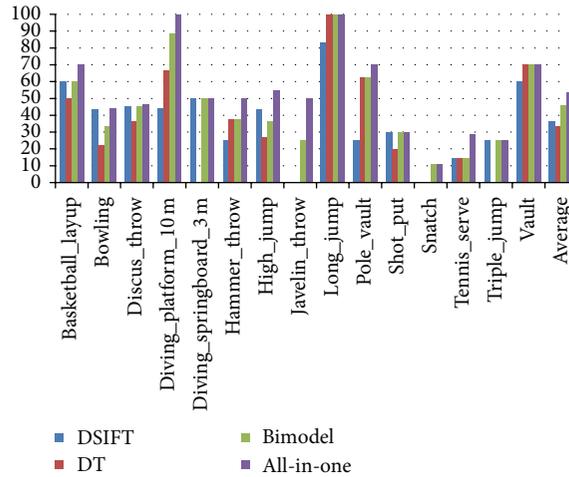


FIGURE 3: Experiment results of our experiment on 16 events of the Olympic dataset.



FIGURE 4: Screenshots of KTH dataset.

we extract nearly 800000 SIFT features. We use the tool provided by [5] to extract Dense Trajectory features with default parameters. Finally, we sample every 100 frames and get about 60000 features for each event. After that, every video is represented by BoW vectors. And we use a grid search and 5-fold cross validation to get optimal parameters for SVM [16] classifiers.

To demonstrate the effectiveness of our method, we implement four experiments for comparison. In the first experiment, we simply used dense SIFT to recognize the testing videos. The second experiment only used Dense Trajectory for recognition. The third experiment extracted the relational information with bimodel. Finally, to demonstrate the influence of our codebook reconstruction method, we combine bimodel and the divisive algorithm we detailed before, which is called all-in-one algorithm.

As Figure 3 shows, the average accuracy of bimodel is obviously higher than the dense SIFT [3] and Dense

Trajectory [5]. And, in most of the cases, the bimodel based accuracy is higher than or the same as the other two features' accuracy. This is in accord with our intuition that relational information contains message from both features which results in better result. Moreover, as the number of testing videos in "javelin_throw" and "snatch" is very small, the single feature based classifiers perform badly. But the bimodel based classifiers can still deal with them. This is due to the fact that bimodel relational information contains more information that single feature does not include. Our experiments results show that the all-in-one algorithm is performing better than other three experiments in almost all cases.

3.2. Experiment on KTH Dataset. The KTH dataset [17] consists of six human action classes. Each action class is performed by 25 people. And every person repeats one action 4 times under different scenarios. Figure 4 shows some screenshots from this dataset.

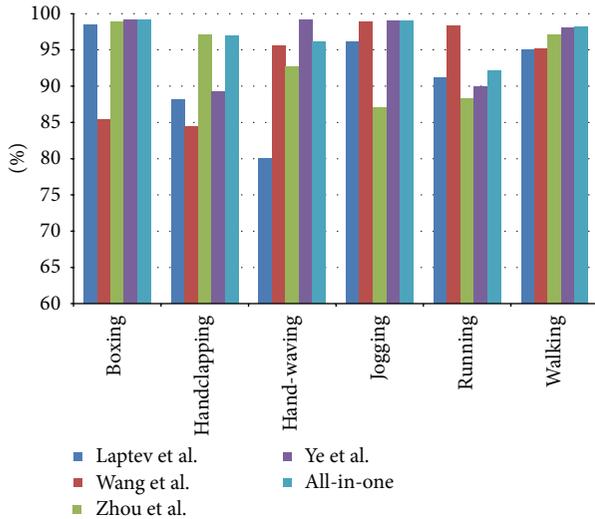


FIGURE 5: Compare proposed method with other BoW based methods on KTH dataset.

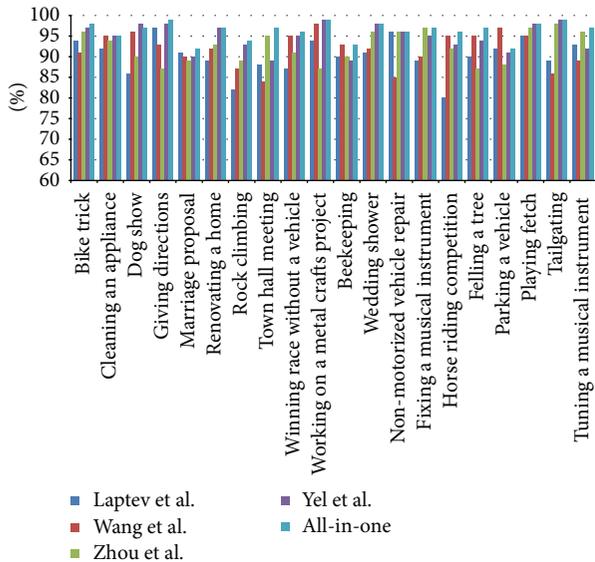


FIGURE 6: Compare proposed method with other BoW based methods on TRECVID MED16 dataset.

As Figure 5 shows, we compare our proposed all-in-one method with other state-of-the-art methods. Among them, Laptev et al. [6] used STIP descriptor. Wang et al. [5] used Dense Trajectory descriptor in multiple scales. And Ye et al. [12] proposed a joint audiovisual bimodal using SIFT and STIP features. Zhou et al. [18] proposed a novel structured codebook construction method to encode rich spatial and temporal contextual information for human action recognition.

It is shown that the proposed all-in-one method is better than other methods for the “boxing,” “hand-waving,” “jogging,” and “walking” actions. Meanwhile, we observe that proposed methods perform relatively worse in “handclapping” class and “running” class. Because the “running” action looks similar to the “jogging” action except the speed and

the “handclapping” action looks similar to the “hand-waving” action, we need more specific information to distinguish them.

3.3. *Experiment on TRECVID MED Dataset.* TRECVID MED is a challenging task for the detection of complicated high-level events. We test our proposed method on the prespecified evaluation events in TRECVID MED 2016 development dataset [19], which includes 20 events. This dataset consists of 200,000 videos.

As Figure 6 shows, our proposed method can better take advantage of the useful information among dense SIFT and Dense Trajectory features and get higher accuracy than the other methods for all events except “parking a vehicle” and “dog show.” For “parking a vehicle” event, our method is more concerned with the actions of the human, but human action is very little in the car. For “dog show” event, Ye et al. use audiovisual bimodal and the barking of the dog gave more clues.

4. Conclusions

In this paper, we present using bimodal for extracting the relational information of local static feature and local motion feature. To overcome the weakness of BoW model, we further introduce a divisive algorithm to keep more information among feature discrete. Our experiments have shown that original static and motion features are complementary to their relational information.

Competing Interests

The author declares that there are no competing interests regarding the publication of this article.

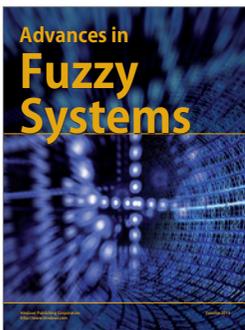
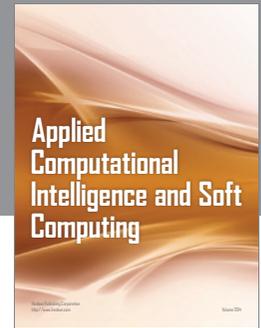
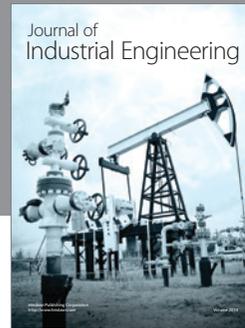
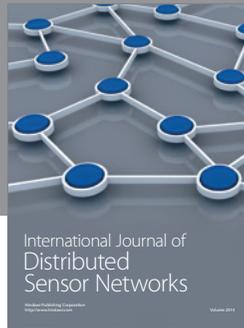
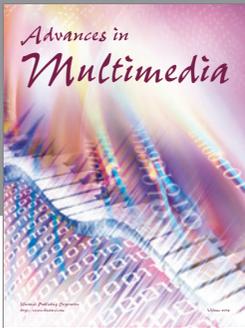
Acknowledgments

This work is supported by National High-Tech R&D Program of China (863 Program) (Grant no. 2015AA015904), China Postdoctoral Science Foundation funded project (2015M571640), Special Grade of China Postdoctoral Science Foundation funded project (2016T90408), CCF-Tencent Open Fund (RAGR20150120), and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

References

- [1] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV '99)*, vol. 2, pp. 1150–1157, IEEE, Kerkyra, Greece, September 1999.
- [2] T. Lindeberg, “Scale invariant feature transform,” *Scholarpedia*, vol. 7, no. 5, article 10491, 2012.
- [3] K. E. A. Van De Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.
- [4] M. Chen and A. Hauptmann, “Mosift: recognizing human actions in surveillance videos,” 2009.

- [5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 3169–3176, IEEE, Providence, RI, USA, June 2011.
- [6] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.
- [7] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1996–2003, Miami, Fla, USA, June 2009.
- [8] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th ACM International Conference on Multimedia (MM '05)*, pp. 399–402, ACM, November 2005.
- [9] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proceedings of the 10th European Conference on Computer Vision: Part I (ECCV '08)*, pp. 179–192, Springer, Berlin, Germany, 2008.
- [10] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 183–190, 1993.
- [11] L. D. Baker and A. K. McCallum, "Distributional clustering of words for text classification," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 96–103, ACM, Melbourne, Australia, August 1998.
- [12] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. T. Lee, and S.-F. Chang, "Joint audio-visual bi-modal codewords for video event detection," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR '12)*, article 39, ACM, Hong Kong, June 2012.
- [13] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.
- [14] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 3209–3216, IEEE, Providence, RI, USA, June 2011.
- [15] <http://vision.stanford.edu/Datasets/OlympicSports/>.
- [16] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.
- [17] <http://www.nada.kth.se/cvap/actions/>.
- [18] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, "Action recognition via structured codebook construction," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 546–555, 2014.
- [19] <http://www.nist.gov/itl/iad/mig/med16.cfm/>.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

