Learning Discriminative Visual Codebook for Human Action Recognition

Qing Lei
Cognitive Science Department, Xiamen University, Xiamen, China
Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen, China
College of Computer Science and Technology, HuaQiao University, Xiamen, China
leiqing@hqu.edu.cn

Shao-zi Li and Hong-bo Zhang
Cognitive Science Department, Xiamen University, Xiamen, China
Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen, China
szlig@xmu.edu.cn, hong8757@163.com

Abstract—This paper explores how to improve BOW model for human action recognition in real environment. Traditional codebook learning uses single appearance based local features, thus spatial and temporal correlations of local features are ignored. This leads to a considerable amount of mismatch between sample vectors and noisy visual words resulted from background clutters. To improve the performance of BOW modeling in real settings, we propose a novel action modeling approach. First, two-level feature selection is applied in the pre-process phase of codebook learning to remove noisy features, thus descriptive and discriminative features are obtained. Then spatial-temporal pyramid matching (STPM) is employed in the feature coding process, in which we model human actions considering not only the appearance similarity between local features but also the spatial relationship of features in space and time. We validate our approach on several benchmark datasets and experimental results show that our approach significantly outperforms K-means clustering on more challenge datasets such as KTH, UCF sports and Youtube datasets.

Index Terms—BOW, human action recognition, codebook learning

I. INTRODUCTION

Video-based human action recognition has been a major research topic in computer vision. The goal of this research is two-fold: deciding what actions are in a video (classification) and where the actions are in the video (localization). Automatically and robustly recognizing human actions in the real-world environment has a wide application in a variety of fields including human-computer interaction, intelligent surveillance, video retrieval and identity authentication. However, the accurate recognition of actions is a highly challenging task because it is influenced by various aspects such as inter-class variation, background clutters, low resolution, occlusion, variation of views and illumination etc.[1]-[3]. Most of the current approaches are either attempting to compute effective features from raw video frames [4]-[8] or trying to learn a powerful codebook for action representation [9]-[14].

In recent years, bag-of-words (BOW) has been extremely popular in computer vision. Traditional BOW model [4]-[8] has been a dominated choice for human action recognition which employs K-means to obtain action-specific codebooks and finds the nearest-neighbor visual word to quantize feature. Video frames is represented by the statistic histogram of a set of "visual words", where unsupervised k-means algorithm is applied to learn a codebook from all feature vectors, and local features are projected to the nearest visual word of learned codebook based on distance measurement. Finally the distribution of centers is computed to obtain the final video representation. However, visual words obtained from k-means clustering are seldom descriptive and effective, especially when they are learned from local patches of images or videos. Significant approximation errors are generated when it is applied in real environment. The reason why BOW modeling is ineffective in realistic settings might be largely due to three shortcomings: 1) each local feature is assigned to the visual word that is closest to it in terms of Euclidean distance which will creates a considerable amount of approximation errors when noisy visual words are generated from background clutters. 2) K-means clustering commonly cluster or quantize the local patches based on computing the similarity of local patches in appearance-based feature space which is unreasonable since it largely neglects the spatial and
temporal contexts of the local features. 3) The clustering process is unsupervised. Earlier works [15]-[16] have shown that K-means process will asymmetrically divide feature space which move clusters to denser regions because of its "mean-shift"-like update rules.

Hence, how to learn an effective and discriminative codebook is a popular research topic in human action recognition. Many reported works are trying to improve the descriptive and discriminative ability of visual words. As we know, unlike document retrieval, the spatial and temporal correlations between local features are significantly useful and important for image or video classification. To overcome the above-mentioned shortcomings of BOW modeling for human action recognition, we present a novel discriminative codebook learning method for robust action modeling. A two-level feature selection method is proposed which considers both inner-class and inter-class differences respectively to obtain descriptive and discriminative visual words. Knowing that spatial and temporal information between local features can be useful for feature classification, we employ spatial-temporal pyramid matching strategy to construct a set of action-specific codebooks that preserve spatial relationship between visual words in three-dimensional spatial and temporal space.

The rest of this paper is organized as follows. Section 2 introduce and summarizes the related traditional work on visual codebook generation. In section 3, we describe our feature selection method and spatial-temporal pyramid matching based feature coding process. The experimental results are presented in Section 4. Finally, we provide concluding remarks and future research in Section 5.

II. RELATED WORKS

A. Feature Detection and Description

Recent proposed feature detection approaches for human action recognition can be divided into two categories: dense sampling [4] and interest point detection [5]-[7]. Dense sampling extracts video blocks at regular intervals of positions throughout or videos at all locations. In addition, multi-scale sampling in space and time is also considered. It has been shown in [4] that features extracted from dense sampling can produce highly accurate results in simple datasets. However, number of features generated by dense sampling is rapidly increased with the incremental of training samples. Moreover, noisy features severely declined the classification performance.

Spatio-temporal interest point (STIP) detection method is proposed and became popular in recent researches. STIP methods are based on the observation that events were frequently occurred in positions with abrupt changes both in time and space. How to detect accurate interest points is of vital importance. Accordingly, typical response function is presented and computed at every location in a video where the extreme points correspond to the keypoints.

Gabor and Gaussian mixed filtering detection algorithm is proposed by Dollar [5] which calculate convolutions separately in spatial domain based on Gaussian filter and in time based on Gabor filtering. The response function has the form:

\[ R = (1^t * g * h_0)^2 + (1^t * g * h_1)^2 \]

where \( g(x, y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \) is the 2D Gaussian smoothing kernel applied only along the spatial dimensions. 

\[ h_0(t; \tau, \omega) = -\cos(2\pi\omega)e^{-\tau^2/2} \]

and \( h_1(t; \tau, \omega) = -\sin(2\pi\omega)e^{-\tau^2/2} \) are a quadrature pair of 1D Gabor filters applied temporally. The two parameters \( \sigma \) and \( \tau \) correspond to the spatial and temporal scale of the detector.

Laptev and Lindeberg [6] propose Harris3D detector as a space-time extension of the Harris corner detector.

They use \( \mu = g(\omega_1, \omega_2) \times (L_x, L_y, L_z) \) \( (L_x, L_y, L_z) \) to compute the convolution of a 3×3 spatial-temporal second-moment matrix composed of first order spatial and temporal derivatives with a Gaussian smooth kernel for each point in video. Where, \( \sigma \) and \( \tau \) correspond respectively to the spatial and temporal scale of a Gaussian smoothing function \( g \) and \( L_x, L_y, L_z \) represent respectively the gradient of video point in x, y, and t direction. The final locations of space-time interest points are given by local maxima of \( H = \det(\mu) - k\text{trace}^2(\mu) \).

Dense sampling and STIP methods have same problem when applied in real environment that noisy features are generated by background clutters or other variations such as camera motion, low resolution and illustration variation etc. This will significantly decline the accuracy of learned classifier. In this paper, we use Harris3D detector and HOG/HOF descriptor proposed by Laptev [7] as the basis feature detection and description, and propose a two-level feature selection approach to choose descriptive and discriminative features. The detail will be introduced in Section 3.1.

B. BOW Modeling and Spatial Pyramid Matching Strategy

STIP based feature detection generated significant outlier features when it applied in realistic scenes. It generates approximation errors and spreads to visual codebook learning. Traditional BOW modeling use K-means clustering and the nearest-neighbor vector quantization to obtain action representation. It discards information about the spatial layout of local features, thus a significant number of mismatches are generated due to noisy visual words learned from background clutters. Therefore it significantly declines the performance on classification.

To modeling the spatial layout of the local features, spatial pyramid matching strategy (SPM) [17] is proposed. Spatial pyramid matching method partitions the image or video into increasingly finer spatial local patches. Typically, \( 2^l \times 2^l \) patches, \( l=0, 1, 2 \) are used. Then histograms of local features for each patch are computed and concatenated to form the final representation of image or video. The resulted “spatial pyramid” is a computationally efficient extension to the unprincipled BOW modeling, and has shown very promising
performance on many image classification tasks. A typical flowchart of the SPM approach based on BOW is illustrated in Figure 1.

In this paper, a discriminative visual codebook learning method for human action recognition is proposed. In purpose of choose descriptive and discriminative features, we propose an evaluating method to measure the discriminate ability of local features which concerns on the difference between visual words of the same action-specific codebook, also the difference between different learned action-specific codebooks. To improve the classification performance affected by mismatches of local features and visual words, we employ STPM method which is an extension of spatial pyramid matching that use a spatial-temporal division on video and statistic the occurrence of low-level features over pre-defined spatial-temporal bins to obtain spatial information compensated action modeling.

III. OUR APPROACH

A. Overview of our approach

Traditional STIP and BoW based human action classification methods are easily influenced by background clutters and camera motions when applied to action recognition in realistic scenes. Most of the detected features (almost 50% of KTH, 70% of UCF and Youtube in our research) are noise and unhelpful for classification. It further declined the discriminate ability of learned action codebooks. Different from traditional BOW modeling, we learn action class-specific codebook for each class. This learning method can calculate the discriminate ability of detected features to construct effective codebooks. It effectively improves the performance of BoW modeling in realistic scenes. On the other hand, this codebook learning method is incremental and provides another advantage that each class is modeled independently of others and hence the painful repetition of the training process when a new class of data is added to the system is no longer necessary.

We introduce the framework of our approach as illustrated in Figure 2. In training phase, we divide training videos into C categories according to action classes. First of all, we use Harris3D detector proposed by Laptev [6] to extract spatial-temporal interest points (STIPs) from videos. Second, 162-dimensioned histograms of gradient and optical flow (HNF) are computed for each STIP to obtain the descriptive feature vector. Third, traditional K-means clustering algorithm is applied in feature vectors belongs to the same action category to acquire K centers; therefore a set of action-specific codebooks is obtained by clustering on feature vectors for different action categories. We use the resulted action-specific codebooks as our preliminary codebooks.

Then two-level selection process (select $p_1$ and select $p_2$) is proposed to obtain discriminative features and remove outlier features. To acquire descriptive and discriminate local features, select $p_1$ is presented to measure the discriminative ability that distinguish different visual words of the same codebook, and select $p_2$ is designed to measure the discriminative ability that distinguish different visual words of different codebooks. Afterwards, vector quantization based on spatial-temporal pyramid matching is applied to encode the descriptors based on the learned codebook. Note that spatial consistency of local neighbor region is an important property of visual entities. So the feature coding process is not only supervised by appearance similarity between local features, but also considering spatial relationship between local features. Finally, the distribution of visual words is summarized and inputted into classifier.

For a test video, we follow the same procedures as training to detect STIPs and compute the feature descriptors of STIPs, then coding the feature descriptors as a set of visual words for each action-specific codebook, after this process, feature codes are passed into a trained classifier for recognition.

Figure 2 outlines the workflow of the proposed approach. More specifically, the upper module of blue dashed line is training process, and the bottom module of red dashed line is test process. In training phase, the blue arrows from left to right in turn represent feature detection, codebook learning, feature coding, and classification, respectively. The pink arrows represent two-level feature selection. In test phase, test sample is processed according to the same procedures, and is classified into pre-defined categories by classifier.

B. Feature Selection

Our feature selection approach is based on two assumptions. First, if a local feature has a strong discrimination, then the confusion of its projection on its belonged codebook is very slight. Namely, obvious
difference exists between the distance of this feature apart from its corresponding visual word and the distance of it far from other visual words. Second, if a local feature has more distinctiveness, then the confusion of its projection on different codebooks is also very slight. Namely, there is a low probability that this feature is projected onto some visual word of its non-specific action category when encoded in all codebooks.

For the first assumption, we propose select $p_1$ algorithm to remove noisy features that arise ambiguity between its corresponding visual word and other visual words of the same codebook in feature coding.

For the second assumption, to deal with outlier features that cause ambiguity between different action-specific codebook, we propose select $p_2$ algorithm. It encodes feature descriptors in a global scope, namely to find the nearest visual word in all codebooks. If the projection resulted codebook is different from local feature truly belongs to, this local feature should be removed from feature sets. The implementation detail is described in Algorithm 2.

Algorithm 1: select $p_1$

Input:
Number of action categories: $C$
Extracted STIP feature set: $P = \{p'_1, p'_2, \ldots, p'_s\}$

Codebook for $i$-th action category: $W_i = \{w^{1}_{i}, w^{2}_{i}, \ldots, w^{n}_{i}\}$

Parameter: $\lambda$

Process:
For $i = 1$ to $C$ (for every action category)  
For $j = 1$ to $s_i$ (for every STIP feature of $i$th action category)  
Calculating the nearest two centers $w^{p^1}_i, w^{p^2}_i$ for $p'_j$ in $W_i$ based on Euclidean distance

\[
\text{if } \frac{||p'_j - w^{p^1}_i||}{||p'_j - w^{p^2}_i||} < \lambda \quad (1)
\]

then $P_i \leftarrow P_i - \{p'_j\}$

Return $(P_i)$.  

Algorithm 2: select $p_2$

Input:
Number of action categories: $C$
Number of extracted STIP features: $M$
Extracted STIP feature set: $P = \{p'_1, p'_2, \ldots, p'_M\}$

Codebook for all action categories: $W = \{w^{1}_1, w^{2}_1, \ldots, w^{n}_1, \ldots, w^{1}_C, w^{2}_C, \ldots, w^{n}_C\}$

Process:
For $k=1$ to $M$ (every STIP feature $p'_j$)  
For $q = 1$ to $C$ (for every action category)  
Calculating the nearest visual word $w^q_j$ for $p'_j$ in $W$ based on Euclidean distance

\[
\text{if } j \neq q \text{ then } P \leftarrow P - \{p'_j\}
\]

Return $(P)$.  

C. Spatial-Temporal Pyramid Matching Based Feature Coding

We use traditional K-means algorithm to cluster extracted features firstly. Note that unlike document representation, spatial information of local features is an important property of visual entities, specifically spatial consistency of local regions essentially exists in various kinds of visual objects. It provides that a useful clue should be taken into account for visual object representation. In our approach, feature coding process is supervised not only by the appearance similarity between local feature and visual words but also by the spatial layout of local features and visual words.

Spatial-Temporal Pyramid Matching is a 3D extension of spatial pyramid matching. It has been successfully used in many visual recognition tasks, such as sports video classification [18]. Spatial-temporal pyramid feature is built by constructing an $L$-level pyramid which partitions a video into 3D grids in a joint spatial-temporal space. Figure 3 shows an example of a spatial temporal pyramid with $L = 3$.

For each level $l$ the 2-dimensional spatial location and 1-dimensional time dimension are divided into $2^l$ cells. For 3D-grid $\zeta = \{z_i, |i = 0, \ldots, D' - 1\}$ at level $l$ in the pyramid, HOG and HOF features which respectively corresponding to histogram of oriented gradient and histogram of optical flow are used to describe human actions. The direction angle is quantified to $k$ bins and
the grid feature \( h_i^l = \{ h_i^l \mid j = 0, ..., D^l \} \) is computed by all the pixels in the 3D-grid. Concatenate features at same level to construct the level-feature \( h_i = \{ h_i^l \mid i = 0, ..., D^l \} \), and normalize the level-feature.

Finally, weighted by \( w_i = \frac{1}{2^l} \)\( (l = 1, ..., L-1, w_0 = w_L) \), level-features of all multiple scale grids are concatenated into the final feature description of a video. Namely, for a given video \( V_i \), it is represented by \( H_i = \{ h_i^l \mid x_i, y_i, t_i \} \), \( h_i^l(x_i, y_i, t_i), ..., h_i^{l-1}(x_{l-1}, y_{l-1}, t_{l-1}) \) in spatial-temporal pyramid matching, where \( h_i^l \) represents the weighted local histogram represented of grid in \( j \)-th level of \( V_i \), and \( x_i, y_i, t_i \) ranges from 0 to \( 2^j \). Therefore, a dimensioned spatial-temporal pyramid based final descriptor is obtained for each video.

**Figure 3. Spatial-temporal Pyramid Matching Structure Feature**

For test, the same feature detection and description procedures as training samples are performed on the test video, and feature descriptors of test sample are obtained. Since we don’t know the classes of these features belong to, we directly encode these features based on learned codebooks without feature selection process. Finally, a \( \chi^2 \) kernel SVM is used to recognize human actions.

We compared the performance of our approach with baseline method as well as other existing approaches. To make our experiment comparable to earlier work, we apply the same evaluation setting and metric as prior art in each dataset.

**A. KTH Dataset**

KTH dataset [23] contains 600 videos with 6 action categories including: boxing, handclapping, handwaving, jogging, running and walking performed by 25 subjects in four scenes: outdoors, outdoors with scale variation, outdoors with different clothes and indoors (refer to Figure 4). The average length of videos is 4 minutes with resolution 160×120, and frame rate 25fps.
We split the datasets into training set and test set according to different subjects. For each experiment we choose four videos of \( n \) subjects for training and actions of remaining \( n_2 - n \) (\( n_2 \) is the total number of subjects) subjects for test. Then the training set contains 4\( n \) videos and the test set contains \( 4(n_2 - n) \) videos. For evaluation, we train a multi-class SVM and evaluate on the test sets. The final average precision (AP) metric is obtained by taking the average of AP for each subject.

The detailed results such as average precision/accuracy per action class and confusion matrices on KTH dataset are reported in the following. Our method achieves a classification accuracy of 94.64% on the KTH dataset which outperforms all published results in Table 2. A detailed result with comparison to original BOW modeling is given in Table 1. The confusion matrix is provided in Figure 5. As seen from Table 2, our approach achieves superior performance on KTH dataset. Observed on confusion matrix, it can be seen that the average accuracy have been significant increased about 13%, 7%, 5% and 2% respectively for jogging, handwaving, walking and boxing, while a relative decrease on running as 5%. It’s probably because of jogging and running, there’s a strong resemblance between this two action categories. So it can hardly be completely discriminated even by humans as confusion frequently occurs.

### Table 1: KTH: Average Accuracy by Action Class

<table>
<thead>
<tr>
<th>Action</th>
<th>BOW</th>
<th>BOW+FS</th>
<th>BOW+FS+STPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxing</td>
<td>98%</td>
<td>98%</td>
<td>100%</td>
</tr>
<tr>
<td>Handclapping</td>
<td>95%</td>
<td>98%</td>
<td>94%</td>
</tr>
<tr>
<td>Handwaving</td>
<td>94%</td>
<td>95%</td>
<td>99%</td>
</tr>
<tr>
<td>Jogging</td>
<td>78%</td>
<td>74%</td>
<td>91%</td>
</tr>
<tr>
<td>Running</td>
<td>86%</td>
<td>84%</td>
<td>81%</td>
</tr>
<tr>
<td>Walking</td>
<td>93%</td>
<td>97%</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>90.8%</td>
<td>91.14%</td>
<td>94.64%</td>
</tr>
</tbody>
</table>

### Table 2: Average Accuracy Comparison on KTH

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>71.7%</td>
<td>92.1%</td>
<td>93.8%</td>
<td>93.5%</td>
<td>94.64%</td>
</tr>
</tbody>
</table>

### B. UCF Sports Dataset

UCF Sports dataset [24] contains close to 150 action sequences with 10 action categories collected from various sports videos. This dataset exhibits occlusion, cluttered background, motion discontinuity and variations in illumination and scale. The action categories are: diving, golf, hswing, kicking, lifting, running, skating, swing-bench, and walking (refer to Fig.6). The riding action has significant misclassification errors from running and kicking classes. Most of the kick action videos contain walk or run action as prelude by the subject of interest and/or the surrounding people, and the confusion is therefore reasonable.

The recognition results are given in Table 3. The confusion matrix is provided in Figure 7. Table 4 compares the proposed approach with a number of existing ones. Apparently, our approach achieves a classification accuracy of 86.39% on the UCF sports dataset which outperforms some published results (72.2%, 83.8%) and comparable with the state-of-the-art method (86.8%) in this table list. Observed on confusion
matrix, it can be seen that the average accuracy have been significantly increased for most of action classes such as diving (7%), kicking (10%), riding (8.3%), skating (8.8%), swing-bench (15%) and walking (5.9%). It shows that our approach picks up discriminative information and obtains effective human action representation. STPM based action modeling is more robust than BOW modeling for action classification in such challenging environments.
TABLE 3
UCF SPORTS: AVERAGE ACCURACY BY ACTION CLASS

<table>
<thead>
<tr>
<th>Action</th>
<th>BOW</th>
<th>BOW+FS</th>
<th>BOW+FS+STPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>diving</td>
<td>93.00%</td>
<td>93%</td>
<td>100.00%</td>
</tr>
<tr>
<td>golf</td>
<td>78.00%</td>
<td>78.00%</td>
<td>78.00%</td>
</tr>
<tr>
<td>hswing</td>
<td>92.00%</td>
<td>92.00%</td>
<td>92.00%</td>
</tr>
<tr>
<td>kicking</td>
<td>70.00%</td>
<td>80.00%</td>
<td>80.00%</td>
</tr>
<tr>
<td>lifting</td>
<td>100.00%</td>
<td>83.33%</td>
<td>100%</td>
</tr>
<tr>
<td>riding</td>
<td>66.70%</td>
<td>75.00%</td>
<td>75.00%</td>
</tr>
<tr>
<td>running</td>
<td>69.20%</td>
<td>69.20%</td>
<td>69.20%</td>
</tr>
<tr>
<td>skating</td>
<td>75.00%</td>
<td>83.80%</td>
<td>83.80%</td>
</tr>
<tr>
<td>swing-bench</td>
<td>80.00%</td>
<td>80.00%</td>
<td>95.00%</td>
</tr>
<tr>
<td>walking</td>
<td>85.00%</td>
<td>90.90%</td>
<td>90.90%</td>
</tr>
<tr>
<td>Average</td>
<td>80.89%</td>
<td>82.52%</td>
<td>86.39%</td>
</tr>
</tbody>
</table>

TABLE 4
AVERAGE ACCURACY COMPARISON ON UCF SPORTS

<table>
<thead>
<tr>
<th>Approach</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al. [19]</td>
<td>72.2%</td>
</tr>
<tr>
<td>Guha et al. [10]</td>
<td>83.8%</td>
</tr>
<tr>
<td>Le et al. [20]</td>
<td>86.8%</td>
</tr>
<tr>
<td>Our approach</td>
<td>86.39%</td>
</tr>
</tbody>
</table>

Table 3 and Table 4 provide the average accuracy of different action classes in the UCF Sports dataset. Table 3 shows the performance of three different methods: BOW, BOW+FS, and BOW+FS+STPM. Table 4 compares our approach with existing methods, showing that our approach outperforms some published results and achieves similar accuracy to the state-of-the-art method.

TABLE 5
YOUTUBE: AVERAGE ACCURACY BY ACTION CLASS

<table>
<thead>
<tr>
<th>Action</th>
<th>BOW</th>
<th>BOW+FS</th>
<th>BOW+FS+STPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>biking</td>
<td>66.67%</td>
<td>70.83%</td>
<td>72.17%</td>
</tr>
<tr>
<td>diving</td>
<td>68%</td>
<td>74%</td>
<td>80%</td>
</tr>
<tr>
<td>golf</td>
<td>68.37%</td>
<td>71.42%</td>
<td>75.51%</td>
</tr>
<tr>
<td>hswing</td>
<td>41.67%</td>
<td>54.17%</td>
<td>60.42%</td>
</tr>
<tr>
<td>jumping</td>
<td>70%</td>
<td>76%</td>
<td>80%</td>
</tr>
<tr>
<td>riding</td>
<td>61.22%</td>
<td>71.43%</td>
<td>78.57%</td>
</tr>
<tr>
<td>spiking</td>
<td>42.42%</td>
<td>57.58%</td>
<td>68.95%</td>
</tr>
<tr>
<td>swing</td>
<td>50%</td>
<td>65%</td>
<td>68%</td>
</tr>
<tr>
<td>walk_dog</td>
<td>45%</td>
<td>60%</td>
<td>63%</td>
</tr>
<tr>
<td>Average</td>
<td>58.14%</td>
<td>67.84%</td>
<td>72.02%</td>
</tr>
</tbody>
</table>

C. Youtube Dataset

Compared to KTH and UCF Sports datasets, the Youtube action dataset [25] is a more challenging dataset with camera motion and jitter, highly cluttered and dynamic backgrounds, compression artifacts, and variable illumination settings. It contains 25 subjects and 11 action categories including basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. Some sample frames about these 11 actions are shown in Figure 8.

Figure 8: Sample frames from the Youtube dataset

We list our recognition results in Table 5. Table 6 compares the proposed approach with a number of existing ones. Our approach achieves a classification accuracy of 72.02% on YouTube dataset which outperforms some published results (65.4%, 70.4%) which is a little lower than the state-of-the-art method (75.8%) in this table list. The confusion matrix is provided in Figure 9. Confusion Observed on confusion matrix, it can be seen that the average accuracy have significant increased about all action categories. When detecting on videos of juggle, shooting and walk_dog, thousands of or more interest points are detected for each video. Most of these points are generated from low resolution or camera motion and useless for classification. Our feature selection method successfully picks out discriminative features and discards outlier features that arises confusion between different action categories. In addition, STPM based codebook constructing method is effective and helpful for learning a more robust action codebook adapted to realistic scenes.

Table 5 and Table 6 compare the proposed approach with existing methods, showing that our approach outperforms some published results and achieves similar accuracy to the state-of-the-art method.
inter-class differences respectively to obtain descriptive method is proposed which considers both inner-class and for robust action modeling. A two-level feature selection present a novel discriminative codebook learning method on human action recognition in real environment. We BOW model for action representation in the application performance in real-world environments.

and achieves competitive performance on UCF Sports that our method outperforms compared methods on KTH KTH, UCF Sports and YouTube datasets. Results show spatial and temporal space. We evaluate our approach on relationship between visual words in three-dimensioned action-specific codebooks that preserve spatial pyramid matching is employed to construct a set of and discriminative visual words. Spatial-temporal

V. CONCLUSION

This paper explores the effectiveness of improved BOW model for action representation in the application on human action recognition in real environment. We present a novel discriminative codebook learning method for robust action modeling. A two-level feature selection method is proposed which considers both inner-class and inter-class differences respectively to obtain descriptive and discriminative visual words. Spatial-temporal pyramid matching is employed to construct a set of action-specific codebooks that preserve spatial relationship between visual words in three-dimensioned spatial and temporal space. We evaluate our approach on KTH, UCF Sports and YouTube datasets. Results show that our method outperforms compared methods on KTH and achieves competitive performance on UCF Sports and YouTube datasets, demonstrating its superior performance in real-world environments.

ACKNOWLEDGMENT

This work was supported by National Nature Science Foundation of China (No.61202143), Doctoral Program Foundation of Institutions of Higher Education of China (No. 20090121110032), Shenzhen Science and Technology Research Foundation (No.JC2009031806 30A, ZYB200907110169A), the Fundamental Research Funds for the Central Universities of Huaqiao University under grant 11IQ3R04.

Table 6

<table>
<thead>
<tr>
<th></th>
<th>Liu et al. [25]</th>
<th>Liu et al. [27]</th>
<th>Le et al. [22]</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>65.4%</td>
<td>70.4%</td>
<td>75.8%</td>
<td>72.02%</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th></th>
<th>biking</th>
<th>diving</th>
<th>golf</th>
<th>jogging</th>
<th>jumping</th>
<th>skipping</th>
<th>tennis</th>
<th>walk</th>
<th>dog</th>
</tr>
</thead>
<tbody>
<tr>
<td>bkie</td>
<td>0.67</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.68</td>
</tr>
<tr>
<td>dirk</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>golf</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>jumpe</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>kne</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>packe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>runn</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>shoote</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>sticke</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>swinge</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>tennis</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>walke</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dog</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.71</td>
</tr>
</tbody>
</table>

(a) Confusion matrix of BOW on Youtube Dataset

(b) Confusion matrix of our approach on Youtube Dataset

Figure 9. The confusion matrices of BOW (a) and our approach (b) on Youtube Dataset.

REFERENCES


Hong-bo Zhang received the B. S. degree in Computer Science and Technology in 2008 from Shenyang Normal University, Shenyang, China. He is currently working toward Ph.D. degree in artificial intelligent in Cognitive Science Department of Xiamen University, China. His research interests include human action analysis, object detection/recognition, and image/video retrieval.

Qing Lei received the B.S. and M.S. degree from Computer Science and Technology College of Huaqiao University, China in 2002 and 2005. She is currently working toward Ph.D. degree in artificial intelligent in Cognitive Science Department of Xiamen University, China. She joined the faculty of Huaqiao University in 2005. Her research interests include human motion analysis and object detection/recognition.

Shao-zi Li received the B.S. degree from the Computer Science Department, Hunan University in 1983, and the M.S. degree from the Institute of System Engineering, Xi’an Jiaotong University in 1988, and the Ph.D. degree from the College of Computer Science, National University of Defense Technology in 2009. He currently serves as the Professor and Chair of Cognitive Science Department of Xiamen University, the Vice Director of Fujian Key Lab of the Brain-like Intelligence System, and the Vice Director and General Secretary concurrently of the Council of Fujian Artificial Intelligence Society. His research interests cover Artificial Intelligence and Its Applications, Moving Objects Detection and Recognition, Machine Learning, Computer Vision, Natural Language Processing and Multimedia Information Retrieval, Network Multimedia and CSCW Technology and others.

© 2013 ACADEMY PUBLISHER