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

Manuscript received July 1, 2013; revised Sept. 1, 2013; accepted Oct. 1, 2013. Project supported by the 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.JC200903180630A, ZYB200907110169A), the Fundamental Research Funds for the Central Universities of Huaqiao University(No.11QZR04). Corresponding author: szlig@xmu.edu.cn (Shao-zi LI).

doi:10.4304/jcp.8.12.3093-3102

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-dimensioned 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 images 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: $R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$, where $g(x, y; \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$ is the *2D* Gaussian smoothing kernel applied only along the spatial dimensions. $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega) e^{-t^2/\tau^2}$ are a quadrature pair of *1D* Gabor filters applied temporally. The two parameters σ and τ 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 $\mu = g(.; \sigma_i^2, \tau_i^2) * ((L_x, L_y, L_t)(L_x, L_y, L_t)^T)$ 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, σ and τ correspond respectively to the spatial and temporal scale of a Gaussian smoothing function g, and L_x, L_y, L_t 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 $H = \det(\mu) - k\,trace^3(\mu)$, $H>0$.

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.

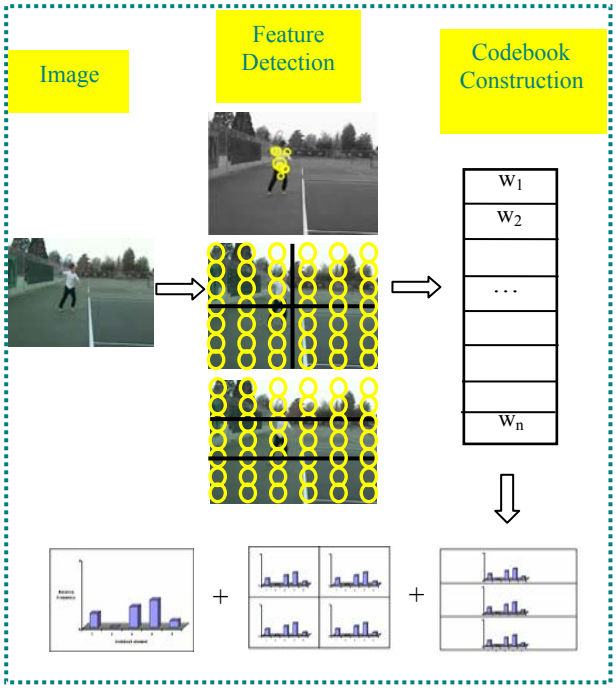

Figure 1. Traditional BOW modeling and spatial pyramid matching based action representation

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 summarizes 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.

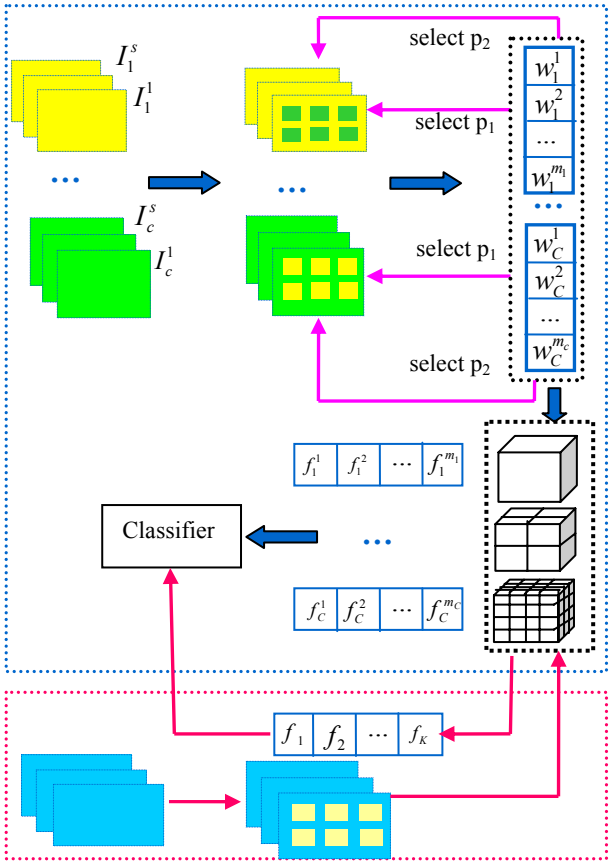

Figure 2. Outlines the workflow of the proposed approach.

Algorithm 1: select p_1

Input:

Number of action categories: C

Extracted STIP feature set: $P_i = \{p_i^1, p_i^2, ..., p_i^{s_i}\}$

Codebook for *i-th* action category: $W_i = \{w_i^1, w_i^2, ..., w_i^{m_i}\}$

Parameter: λ

Process:

For $i = 1$ to C *(for every action category)*

For $j = 1$ to s_i *(for every STIP feature of ith action category)*

Calculating the nearest two centers w_i^p, w_i^q for p_i^j in W_i based on Euclidean distance

$$\text{if } fabs(\frac{\| p_i^j - w_i^p \|_2}{\| p_i^j - w_i^q \|_2}) < \lambda \qquad (1)$$

$$\text{then } P_i \leftarrow P_i - \{p_i^j\}$$

Return (P_i).

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 2: select p_2

Input:

Number of action categories: C

Number of extracted STIP features: M

Extracted STIP feature set:

$$P = \{p_1^1, p_1^2, ..., p_1^{s_1}, p_2^1, ..., p_2^{s_2}, ..., p_C^1, ..., p_C^{s_C}\}$$

Codebook for all action categories:

$$W = \{w_1^1, w_1^2, ..., w_1^{m_1}, w_2^1, ..., w_2^{m_2}, ..., w_C^1, ..., w_C^{m_C}\}$$

Process:

For k=1 to M (*every STIP feature* p_i^j)

For q = 1 to C *(for every action category)*

Calculating the nearest visual word w_l^t for p_i^j in W based on Euclidean distance

$$\text{if } l \neq i \text{ then } P \leftarrow P - \{p_i^j\}$$

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 $\varsigma = \{g_i \mid i = 0, ..., D^l - 1\}$ 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 $h_i^l = \{h_{ij}^l \mid j = 0,..., D^l\}$ is computed by all the pixels in the 3D-grid. Concatenate features at same level to construct the level-feature $h^l = \{h_i^l \mid i = 0,..., D^l\}$, and normalize the level-feature. Finally, weighted by $w_i = \frac{1}{2^{L-l}}(l = 1,..., L-1, w_0 = w_1)$, 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 $H_i = \{h_i^0\{x_0, y_0, t_0\}, h_i^1(x_1, y_1, t_1),..., h_i^{L-1}(x_{L-1}, y_{L-1}, t_{L-1})\}$ in spatial-temporal pyramid matching, where h_i^j represents the weighted local histogram represented of grid in *jth* level of V_i, and x_j, y_j, t_j ranges from 0 to 2^j. Therefore, a dimensioned spatial-temporal pyramid based final descriptor is obtained for each video.

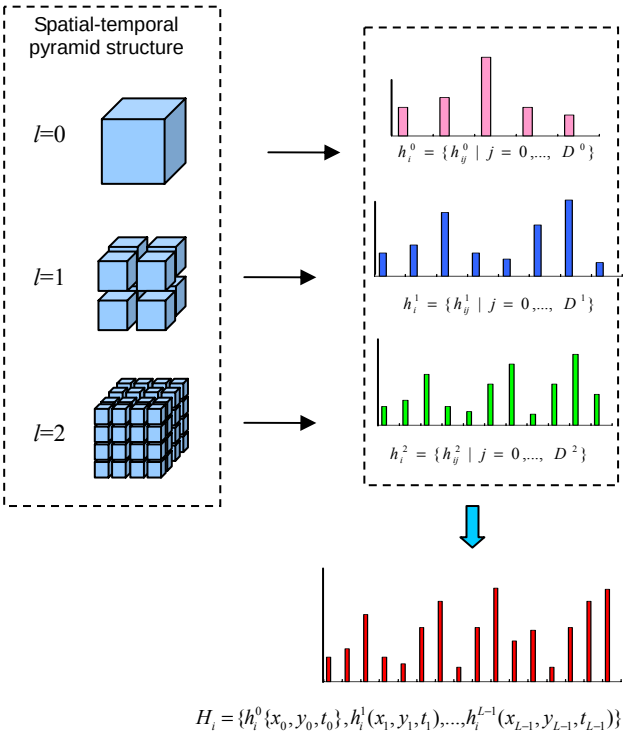

Figure 3. Spatial-temporal Pyramid Matching Structure Feature

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, nonlinear kernel STPM that uses spatio-temporal pyramid histograms and Chi-square kernel is applied in our experiments for recognition.

In nonlinear SVM based classification, all training samples are represented by the distribution of all visual words of all action categories. And for a given test video, feature coding is implemented on all learned codebooks and obtain the corresponding distribution of all visual words as its final representation. Finally, C-to-1 SVM classifier is used for recognition. Codes of training and test samples are passed into a trained SVM classifier with χ^2 kernel as shown in equation 2 for classification.

$$K(H_i, H_j) = 1 - \sum_{m=1}^{K} \frac{(h_{im} - h_{jm})^2}{\frac{1}{2}(h_{im} + h_{jm})} \qquad (2)$$

Where $H_i = \{h_i^1, h_i^2,..., h_i^K\}$ and $H_j = \{h_j^1, h_j^2,..., h_j^K\}$ are descriptors of two videos, K is the dimensions of codebook.

IV. Performance Evaluation

In the experiment section of this paper, we present the performance of our approach on three human action recognition datasets: KTH [23], UCF Sports [24] and Youtube [25]. For all datasets, we use Laptev's 3DHarris spatial-temporal interest point detection [6] and HNF feature description [7] as our baseline feature extraction method.

In Laptev's method, HNF feature is obtained by concatenating HOG feature and HOF feature. HOG feature is computed in the neighborhood of detected points. The size of each volume is related to the detection scales. Each volume is subdivided into a 3×3×2 grid of patches. Orientations of the gradient are divided in 4 bins, and it generates a 72-dimensioned feature vector to describe each interesting point. HOF feature is also computed, while orientations of motion flow are divided into 5 bins, so it generates a 90-dimensional feature vector for each point. HNF feature concatenates two vectors and forms the final 162-dimensional feature vector. Then traditional K-means quantization is employed to obtain the initial visual action-specific codebooks, where K=100 is used for each action category. We choose λ=0.8 in feature selection process p_1. Finally, a χ^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.

Figure 4: Sample frames from the KTH dataset

We split the datasets into training set and test set according to different subjects. For each experiment we choose four videos of n_1 subjects for training and actions of remaining $n_2 - n_1$ (n_2 is the total number of subjects) subjects for test. Then the training set contains $4n_1$ videos and the test set contains $4(n_2 - n_1)$ 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 list 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

	BOW	BOW+FS	BOW+FS+STPM
Boxing	98%	98%	100%
Handclapping	95%	98%	94%
Handwaving	94%	95%	99%
Jogging	78%	74%	91%
Running	86%	84%	81%
walking	93%	97%	100%
Average	90.8%	91.14%	94.64%

TABLE 2:
AVERAGE ACCURACY COMPARISON ON KTH

	Schuldt et. al [19]	Wang et.al [21]	Le et.al[22]	Gall et.al[26]	Our approach
Average	71.7%	92.1%	93.8%	93.5%	94.64%

	box	han	han	jogg	run	wal
boxing	0.98	0.02	0	0	0	0
handclapp	0.04	0.96	0	0	0	0
handwavir	0.01	0.05	0.94	0	0	0
jogging	0	0	0	0.78	0.2	0.02
running	0	0	0	0.13	0.86	0.01
walking	0	0	0	0.06	0.01	0.93

(a) Confusion matrix of BOW on KTH

	box	han	han	jogg	run	wal
boxing	1	0	0	0	0	0
handclapp	0.04	0.96	0	0	0	0
handwavir	0	0.01	0.99	0	0	0
jogging	0	0	0	0.91	0.09	0
running	0	0	0	0.19	0.81	0
walking	0	0	0	0	0	1

(b) Confusion matrix of our approach on KTH

Figure 5. The confusion matrixes of BOW (a) and our approach (b) on KTH

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, riding, 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.

Figure 6: Sample frames from the UCF Sports dataset

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 significant 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

	BOW	BOW+FS	BOW+FS+STPM
diving	93.00%	93%	100.00%
golf	78.00%	78.00%	78.00%
hswing	92.00%	92.00%	92.00%
kicking	70.00%	80.00%	80.00%
lifting	100.00%	83.33%	100%
riding	66.70%	75.00%	75.00%
running	69.20%	69.20%	69.20%
skating	75.00%	83.80%	83.80%
swing-bench	80.00%	80.00%	95.00%
walking	85.00%	90.90%	90.90%
Average	80.89%	82.52%	86.39%

TABLE 4
AVERAGE ACCURACY COMPARISON ON UCF SPORTS

	Wang et.al [19]	Guha et.al [10]	Le et.al[20]	Our approach
Average	72.2%	83.8%	86.8%	86.39%

	diving	golf	hswing	kicking	lifting	riding	running	skating	swing-ber	walk
diving	0.93	0	0	0	0.07	0	0	0	0	0
golf	0	0.78	0	0	0	0.11	0	0	0	0.11
hswing	0.08	0	0.92	0	0	0	0	0	0	0
kicking	0	0	0	0.7	0	0.05	0.05	0	0	0.1
lifting	0	0	0	0	1	0	0	0	0	0
riding	0	0	0	0.08	0	0.67	0.25	0	0	0
running	0	0	0	0.15	0	0.07	0.692	0	0	0.07
skating	0	0	0	0.08	0	0	0	0.75	0	0.17
swing-be	0	0	0	0.05	0	0	0.05	0.1	0.8	0
walk	0	0.05	0	0	0	0.05	0.05	0	0	0.85

(a) Confusion matrix of BOW on UCF Sports

	diving	golf	hswing	kicking	lifting	riding	running	skating	swing-ber	walk
diving	1	0	0	0	0	0	0	0	0	0
golf	0	0.78	0		0	0.11	0	0		0.11
hswing	0.08	0	0.92	0	0	0	0	0	0	0
kicking	0	0	0	0.8	0		0	0	0	0.1
lifting	0	0	0	0	1	0	0	0	0	0
riding	0	0	0	0.083	0	0.75	0.167	0	0	0
running	0	0	0	0.15	0	0.07	0.692	0	0	0.07
skating	0	0.08	0	0	0	0	0	0.8333	0	0.083
swing-be	0	0	0	0	0	0	0	0.05	0.95	0
walk	0	0.05	0	0	0	0	0	0.05	0	0.9

(b) Confusion matrix of our approach on UCF Sports

Figure 7. The confusion matrixes of BOW (a) and our approach (b) on UCF Sports

C. Youtube Dataset

Compared to KTH and UCF Sports datasets, 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
YOUTUBE: AVERAGE ACCURACY BY ACTION CLASS

	BOW	BOW+FS	BOW+FS+STPM
biking	66.67%	70.83%	72.17%
diving	68%	74%	80%
golf	68.37%	71.42%	75.51%
juggle	41.67%	54.17%	60.42%
jumping	70%	76%	80%
riding	61.22%	71.43%	78.57%
shooting	42.42%	57.58%	60.61%
spiking	63.16%	75.79%	78.95%
swing	63%	70%	75%
tennis	50%	65%	68%
walk_dog	45%	60%	63%
Average	58.14%	67.84%	72.02%

TABLE 6
AVERAGE ACCURACY COMPARISON ON YOUTUBE DATASET

	Liu et.al [25]	Liu et.al [27]	Le et.al[22]	Our approach
Average	65.4%	70.4%	75.8%	72.02%

	biking	diving	golf	juggle	jumping	riding	shooting	spiking	swing	tennis	walk_dog
bikir	0.67	0.03	0.00	0.02	0.00	0.06	0.00	0.00	0.06	0.00	0.16
divir	0.04	0.68	0.05	0.01	0.00	0.02	0.07	0.07	0.02	0.02	0.02
golf	0.01	0.00	0.68	0.09	0.02	0.01	0.11	0.01	0.03	0.01	0.02
juggl	0.05	0.00	0.11	0.42	0.08	0.05	0.06	0.03	0.07	0.04	0.07
jumpi	0.00	0.00	0.01	0.07	0.70	0.02	0.03	0.00	0.11	0.01	0.05
ridir	0.09	0.01	0.00	0.03	0.06	0.61	0.02	0.00	0.01	0.03	0.13
shoot	0.05	0.04	0.08	0.10	0.03	0.04	0.42	0.04	0.04	0.11	0.04
spiki	0.02	0.06	0.03	0.02	0.00	0.03	0.13	0.63	0.04	0.03	0.00
swing	0.10	0.02	0.01	0.06	0.08	0.00	0.02	0.03	0.63	0.00	0.05
tenni	0.04	0.01	0.08	0.10	0.04	0.03	0.07	0.08	0.00	0.50	0.05
walk_	0.14	0.02	0.01	0.06	0.05	0.11	0.02	0.04	0.06	0.04	0.45

(a) Confusion matrix of BOW on Youtube Dataset

	biking	diving	golf	juggle	jumping	riding	shooting	spiking	swing	tennis	walk_d
bikir	0.73	0.03	0.00	0.02	0.01	0.03	0.00	0.00	0.04	0.00	0.14
divir	0.02	0.80	0.04	0.01	0.00	0.01	0.05	0.05	0.00	0.00	0.02
golf	0.01	0.01	0.76	0.05	0.02	0.01	0.07	0.01	0.03	0.01	0.02
juggl	0.02	0.00	0.10	0.60	0.06	0.04	0.03	0.01	0.04	0.03	0.05
jumpi	0.00	0.00	0.01	0.05	0.80	0.02	0.02	0.00	0.06	0.01	0.03
ridir	0.04	0.01	0.00	0.02	0.03	0.79	0.02	0.00	0.01	0.00	0.08
shoot	0.01	0.02	0.05	0.08	0.04	0.03	0.61	0.03	0.02	0.09	0.02
spiki	0.02	0.03	0.02	0.01	0.00	0.02	0.07	0.79	0.02	0.01	0.00
swing	0.06	0.02	0.01	0.02	0.05	0.00	0.02	0.03	0.75	0.00	0.04
tenni	0.02	0.01	0.04	0.07	0.03	0.02	0.05	0.06	0.00	0.68	0.02
walk_	0.10	0.01	0.01	0.03	0.03	0.07	0.02	0.03	0.04	0.03	0.63

(b) Confusion matrix of our approach on Youtube Dataset

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

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 11QZR04.

REFERENCES

[1] R. Poppe, Vision-based human motion analysis: An overview, Computer Vision and Image Understanding, vol.108, no. 1-2, pp. 4-18, 2007.

[2] R. Poppe. A survey on vision-based human action recognition. Image and Vision Computing. Vol.28, no.6, pp. 976-990, 2010.

[3] J. K. Aggarwal, M. S. Ryoo. Human Activity Analysis: A Review. ACM Computing Surveys. Vol. 43, 2010.

[4] H.Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of Local Spatio-Temporal Features for Action Recognition. British Machine Vision Conference (BMVC), 2009.

[5] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features", In Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65-72, 2005.

[6] I. Laptev, T. Lindeberg, "Space-time interest points", In Proceedings of the International Conference on Computer Vision (ICCV'03), pp. 432-439, 2003.

[7] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies", In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), pp. 1-8, 2008.

[8] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", In Proceedings of the International Conference on Multimedia (MultiMedia'07), pp. 357–360, 2007.

[9] Angela Yao, Juergen Gall, Luc van Gool. A Hough Transform-based Voting Framework for Action Recognition. Proceedings 23rd IEEE computer society conference on computer vision and pattern recognition - CVPR2010, June 13-18, 2010, San Francisco, California, USA.

[10] Tanaya Guha, Rabab Kreidieh Ward. Learning Sparse Representations for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.34, no. 8, 2012.

[11] J. Sun, X. Wu, Hierarchical Spatio-Temporal Context Modeling for Action Recognition, In Proceedings of the Computer Vision and Pattern Recognition, pp.2004-2011, 2009.

[12] Yi Ouyang, Jianguo Xing, Human Action Recognition algorithm based on Minimum Spanning Tree of CPA Models, Journal of Software, Vol 7, No 7 (2012), 1577-1584, Jul 2012.

[13] Ling Gan, Fu Chen, Human Action Recognition Using APJ3D and Random Forests, Journal of Software, Vol 8, No 9 (2013), 2238-2245, Sep 2013.

[14] Chuanxu Wang, An Algorithm of Unsupervised Posture Clustering and Modeling Based on GMM and EM Estimation, Journal of Software, Vol 6, No 7 (2011), 1201-1208, Jul 2011.

[15] F.Jurie and B.Triggs, Creating Effcient Codebook for Visual Recognition, Proc.10th IEEE Int'l Conf. Computer Vision, 2005.

[16] MacQueen J. Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematical Statistics and Probability, pp 281–297.

[17] Lazebnik S，Schmid C，Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE Conference on Computer Vision and Pattern Recognition (CVPR' 2006). New York, USA, June 17-22, 2006: 2169-2178.

[18] Jaesik Choi, Won J. Jeon, and Sang-Chul Lee. Spatio-Temporal Pyramid Matching for Sports Videos. Multimedia Information Retrieval, page 291-297. ACM,

(2008)
[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In ICPR, 2004.
[20] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D gradients. In BMVC,2008.
[21] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2010.
[22] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. CVPR 2011: 3361-3368.
[23] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach", In Proceedings of the International Conference on Pattern Recognition, pp. 32–36, 2004.
[24] M. Rodriguez, J. Ahmed, and M. Shah, "Action Match a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, June 2008.
[25] J. Liu, J. Luo, M. Shah. Recognizing realistic actions from videos in the wild. In Proceedings of the Computer Vision and Pattern Recognition, 2009. YouTube dataset is available at <http://www.cs.ucf.edu/liujg/YouTube_Action_dataset.html>.
[26] J.Gall and A. Yao and N. Razavi and L. Van Gool and V. Lempitsky. Hough Forests for Object Detection, Tracking, and Action Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 33, No. 11, 2011, pp: 2188-2202.
[27] J.Liu, Y. Yang, I. Saleemi, et al. Learning semantic features for action recognition via diffusion maps. Computer Vision and Image Understanding. Vol.116, No. 3, 2012, pp: 361-377.

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.

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.