SHANGHAI JIAOTONG UNIVERSITY(SJTU-IS) participation in high-level feature extraction at TRECVID 2009

XingHao Jiang, TanFeng Sun

Bin Chen,GuangLei Fu,Bing Feng, RongJie Li

School of Information Security

SHANGHAI JIAOTONG University,SHANGHAI 200240,CHINA

Abstract

In this paper, a description of our participation in high-level feature extraction at TRECVID 2009 will be given. Four runs had been submitted for our effort on the task. For those four runs, they have some common points and are based on one system. The basic elements for our work is different kinds of the descriptive of any frame(picture) retrieved from the videos. We have taken six kinds of image descriptive into consideration this time, which are sift-bow, local binary pattern, color moment, color hist, edge hist and gabor. Then we tried the different combination of those descriptive in the training and classifying of SVM. The result of the task is based on the linear weighted fusion of the separate result of each descriptive. We submitted the following four runs:

RUN 1: sift-bow with vocabulary of 500 words and with soft weighting method and with k-means of different feature data, plus color moment, LBP, color hist, edge hist and gabor, the final result is based on the linear weighted fusion of each result from each descriptive.

RUN 2: remove the sift-bow descriptive from RUN1 and the final result is based on the linear weighted fusion of each result from each descriptive.

RUN 3: sift-bow with vocabulary of 500 words and with soft weighting method and with k-means of all the feature data, plus color moment, LBP, color hist, edge hist and gabor, the final result is based on the linear weighted fusion of each result from

each descriptive.

RUN 4: sift-bow with vocabulary of 200 words and without soft weighting method and with k-means of different feature data, plus color moment, LBP, color hist, edge hist and gabor, the final result is based on the linear weighted fusion of each result from each descriptive.

1 System overview

Input: the input of the system is the frames of the videos. We retrieve those frames from the shot boundary defined by TRECVID and three frames one shot.

Low level descriptive extraction: for those input video frames, five kinds of low level descriptive will be extracted. They are color moment(9dims * 25 block, totally 225 dims), color hist(256 dims), edge hist(80 dims), gabor(48 dims), local binary pattern(256 dims). The details of the low level descriptive extraction will be given in section 2.

High level descriptive extraction: sift-bow descriptive will be extracted from those video frames as well. Sift descriptive has been proven a effective method in local descriptive process in image processing. And we append bow(bag of visual words) processing to the sift in order to keep the consistent form of one vector on frames. In the bow processing , we have taken different methods, which includes different vocabulary size and different dataset for k-means and whether using soft-weighting.the details of the high-level descriptive extraction will be given in section 3.

SVM classifying for each descriptor respectively: with the help of the six kinds of descriptive vector extracted from video frames, we used SVM to train and predict each frame respectively. Six result of one frame will be generated.

linear weighted result fusion: according to the six results of one frame from six kinds of low-level and high-level descriptive, the final result will be given out by the linear weighted fusion. The weight of each descriptive is based on the early testing of a certain dataset from part of the TRECVID video dataset.

Figure 1 is a framework of our system,

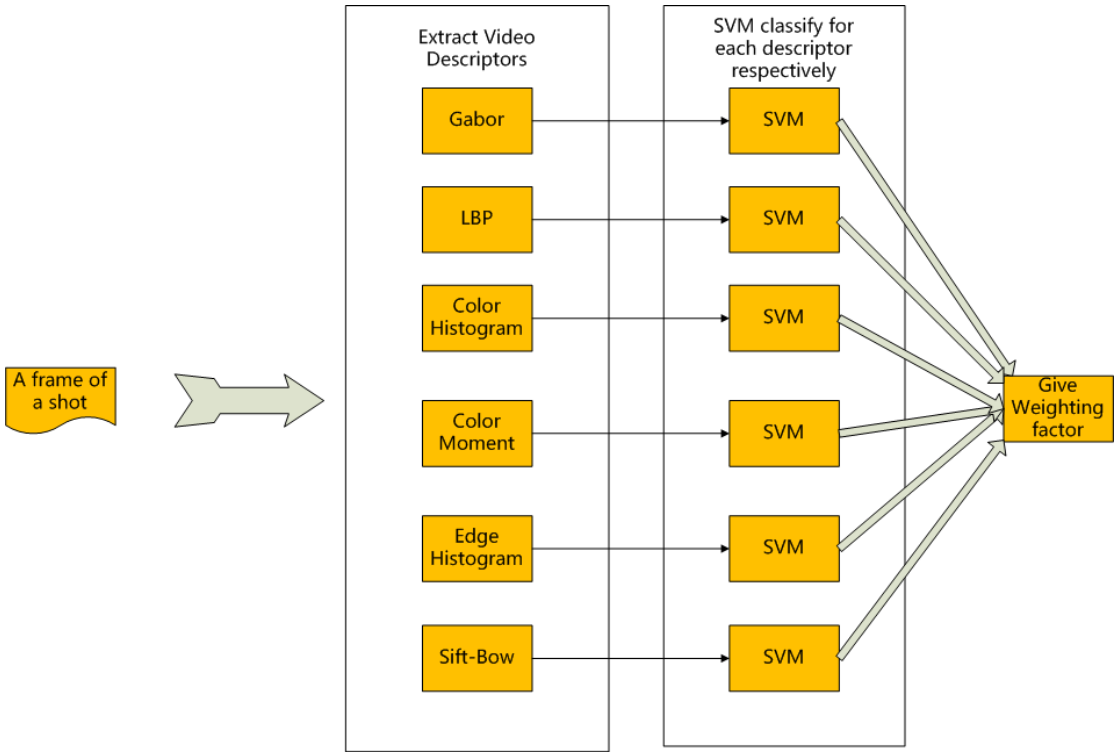

Figure 1 system framework

2 low-level image descriptive extraction

we used five low-level descriptive in our work, which includes local binary pattern, color moment, color hist, edge hist and Gabor.

LBP: The local-binary-pattern (LBP) operator is a theoretically simple yet very effective multi-resolution statistical texture descriptor in terms of the characteristics of the local structure, and has been applied in many areas. We use lbp to produce a histogram whit 256 bins and ranges from 1 to 256 in each bin.

Color Moments: Color moments are measures that can be used differentiate images based on their features of color. Once calculated, these moments provide a measurement for color similarity between images. These values of similarity can then be compared to the values of images indexed in a database for tasks like image retrieval. We have divided a frame into 25 chunks , thus produced a 9*25-dimensional vector.

Color Histogram: color histogram is a representation of the distribution of colors in an image, derived by counting the number of pixels of each of given set of color ranges in a typically two-dimensional (2D) or three-dimensional (3D) color space. It is a flexible construct that can be built from images in various color spaces, whether RGB, chromaticity or any other color space of any dimension. In our runs ,we have adopted the HSV-space and produced a 256-bin histogram per frame.

Edge Histogram: The edge histogram descriptor captures the spatial distribution of edges, somewhat in the same spirit as the CLD. The distribution of edges is a good

texture signature that is useful for image to image matching even when the underlying texture is not homogeneous. In our runs, it produced a 80-bin histogram per frame.

Gabor Wavelets: Gabor wavelets were originally developed to model the receptive fields of simple cells in the visual cortex and in practice they capture a number of salient visual properties including spatial localization, orientation selectivity and spatial frequency selectivity quite well. They have been widely used in face recognition. We have introduced 48 gabor filters thus produced a 48- dimensional vector per frame.

3 sift-bow and knn ,k-means soft-weighting

3.1 BOW scheme

We adopt the Bag-Of-Visual Word scheme to deal with the key-points in the image.

Key-points are salient image patches that contain rich local information of an image. In our experiment we use the sift algorithm to extract the key-points. Key-points are then grouped into a large number of clusters so that those with similar descriptors are assigned into the same cluster. To get the cluster, we use the k-means algorithm. By treating each cluster as a “visual word” that represents the specific local pattern shared by the key-points in that cluster, we have a visual-word vocabulary describing all kinds of local image patterns. With its key-points mapped into visual words, an image can be represented as a “bag of visual words”, or specifically, as a vector containing the (weighted) count of each visual word in that image, which can be used as a feature vector in classification task.

The following figure is the framework of sift-bow module:

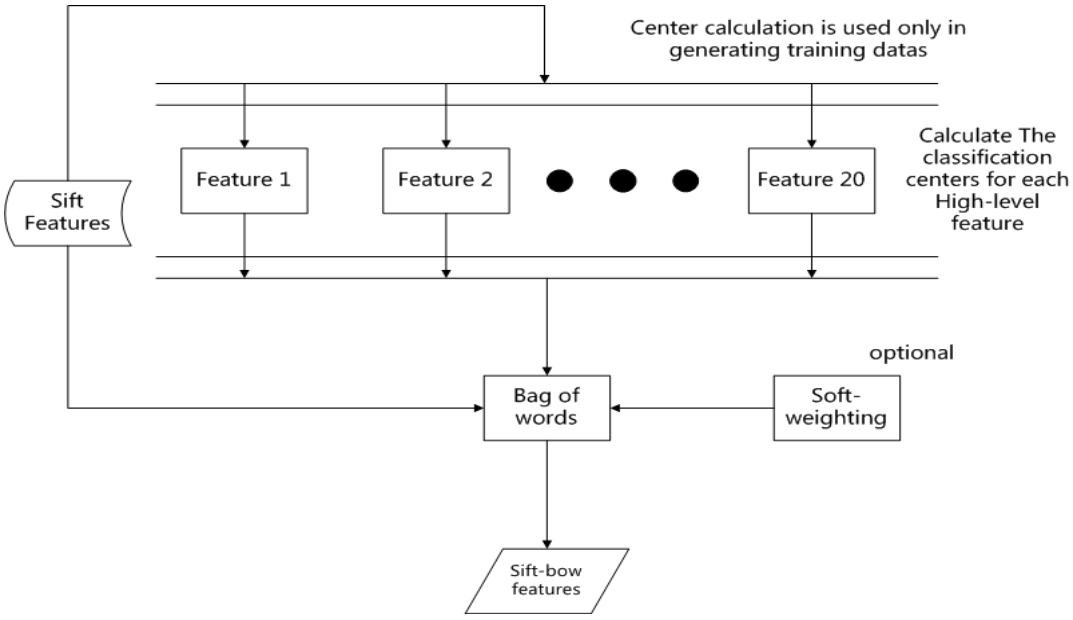

Figure 2 framework for sift-bow module with kmeans for each feature

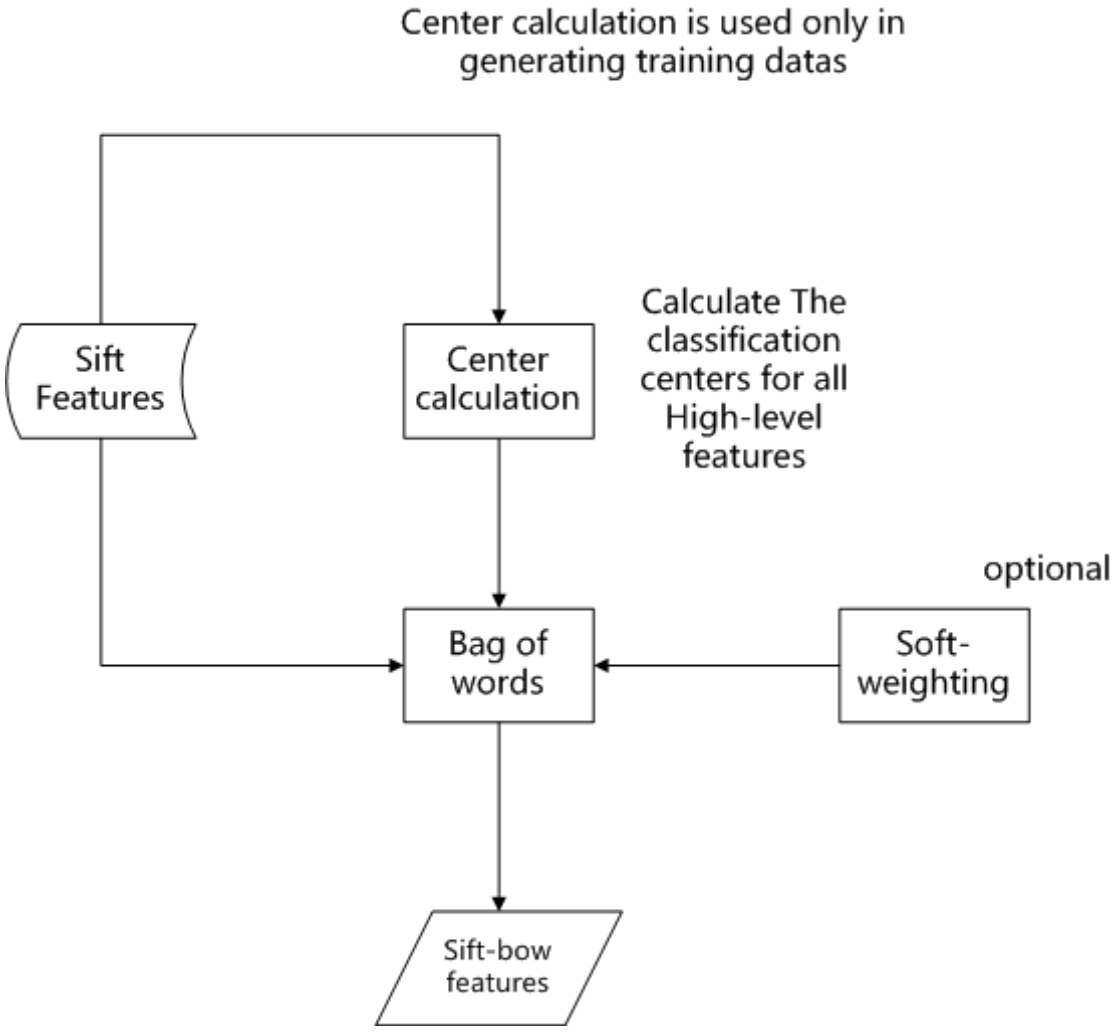

Figure 3 framework for sift-bow module with kmeans for all features

3.2 Soft-weighting scheme

For each key-point in an image, instead of searching only for the nearest visual word, we select the top-N nearest visual words. Suppose we have a visual vocabulary of K visual words, we use a K-dimensional vector T = [t1,t2,…,t_K] with each component t_k representing the weight of a visual word k in an image such that

$$t_k = \sum_{i=1}^{N} \sum_{j=1}^{Mi} \frac{1}{2^{i-1}} similarity(j,k)$$

In this expression, Mi represents the number of key-points whose ith nearest neighbor is visual word k. The measure similarity (j, k) represents the similarity between key-point j and visual word k. Notice that in the above equation the contribution of a key-point is dependent on its similarity to word k weighted by $1/2^{i-1}$, representing the word is its ith nearest neighbor. In our experiment, while we using the weighting scheme, we define the N=4.

4 SVM learning and result fusion and distribution of the weight

We adopt the SVM for classification of our data. In our experiment, we use the

TRECVID2007 data for the training data and the TRECVID2008-2009 data for the testing. For the training data, we fist extract the key-frame from the key-shot presented by the organization. Then we extract the feature by our algorithm, totally there are 6 features. We adopt the back-fusion scheme to deal with the SVM result for each feature which means we first get the SVM result of each feature and then fuse them with some weight scheme. By doing some test, we get the rough distribution of each feature for different scene.

5 experiment results and analysis

The table 1 is the results for our participation in the high-level feature extraction at TRECVID 2009. The bold figure in the table means the best score for the same feature among four runs.

Through the comparison of the bold figure, we can see run1 in a obviously best place, it has more features with the best score. And run2, run3 and run4 have got the similar profile. And for the run2 ,it got the worst results, which mean the importance of the place of sift-bow descriptive.(run 2 does not include the sift-bow descriptive).

We found the features about people have got the relatively poor results, which is in our expectation. For the descriptive we used in our work is just image descriptive monotonously. Some descriptive about audio is necessary in analyze the people motion. And further, some specific algorithm aiming at human motion is needed as well.

High-level features	Run 1	Run 2	Run 3	Run 4
Classroom	**0. 049**	0. 032	**0. 049**	0. 031
chair	**0. 004**	0. 002	0. 003	0. 002
infant	0. 000	0. 000	0. 000	0. 000
traffic	**0. 003**	0. 002	0. 001	0. 002
doorway	**0. 017**	0. 009	0. 015	0. 009
airplane_flying	0. 023	0. 006	0. 026	**0. 030**
personal-playing-a-musical-instrument	**0. 017**	0. 008	0. 015	**0. 017**
bus	**0. 007**	0. 001	0. 004	0. 002
person-playing-soccer	**0. 007**	0. 006	0. 006	0. 001
cityscape	**0. 051**	0. 018	0. 031	0. 045
person-riding-a-bicycle	0. 000	0. 000	0. 000	0. 000
telephone	**0. 004**	0. 002	0. 003	0. 002
person-eating	0. 009	0. 002	0. 006	**0. 035**
demonstration-or-protest	**0. 008**	0. 002	**0. 008**	0. 006
hand	**0. 041**	0. 008	0. 035	0. 008
people-dancing	**0. 012**	0. 011	**0. 012**	**0. 012**
nighttime	**0. 038**	0. 028	0. 037	0. 031
boat-ship	0. 063	0. 047	0. 066	**0. 068**
female-human-face-closeup	0. 013	**0. 018**	0. 015	**0. 018**
singing	0. 021	0. 020	**0. 026**	0. 016

Table 1 experiment results in high-level feature extraction at trecvid 2009

6 Reference

1 Evaluating Bag-of-Visual-Words Representations in Scene Classification, Jun Yang, Yu-Gang Jiang, Alexander G.Hauptmann, Chong-Wah Ngo 2007

2 VIDEO CLASSIFICATION BASED ON LOW-LEVEL FEATURE FUSION MODEL, Mickael Girionnet, Denis Pellerin, and Michele Rombaut

3 SVM-based Video Scene Classification and Segmentation, Yingying Zhu, Zhong Ming,2008

4 Local feature view clustering for 3D object recognition,David G.Lowe,December 2001