XJTU at TRECVID2008 High-Level Feature Extraction

Zhe Wang, Guizhong Liu, Xueming Qian, Zhi Li,
DanPing Guo, Nan Nan, Huaixia Jiang

Signal Processing Multi-Disciplinary Research Center,
School of Electronic and Information Engineering,
Xi'an Jiaotong University, Shanxi, P.R.China
zhe.wang313@gmail.com

Abstract

In this paper, we present our experiments in TRECVID 2008 about High-Level feature extraction task. This is the first year for our participation in TRECVID, our system adopts some popular approaches that other workgroups proposed before. We proposed 2 advanced low-level features NEW Gabor texture descriptor and the Compact-SIFT Codeword histogram. Our system applied well-known LIBSVM to train the SVM classifier for the basic classifier. In fusion step, some methods were employed such as the Voting, SVM-base, HCRF and Bootstrap Average AdaBoost(BAAB).

Keywords: high-level feature, semantic concept, low-level feature, support vector machines, multiple classifiers model, classifier fusion.

1. INTRODUCTION

In our first participation in TRECVID high-level feature extraction task, our mainly goal is that constructing intact high-level feature extraction system and mastering the pivotal technologies of this research area. Due to lack of experience and time, there are some missteps in our work. For example, since time pressure, we set the limit of iterative times while training the HCRF fusion method that was adopted in A_XJTU_2_2, A_XJTU_3_3 and A_XJTU_6_6. However, the astringency of the HCRF is deficient with less iterative times. So, the 3 runs are unsuccessful. In A_XJTU_1_1 and A_XJTU_5_5, we attempted the BAAB fusion method to combine the results of the multiple classifiers model. But, the sample data that was distributed to train the BAAB fusion method was too less. Therefore, the performances of the 2 runs are also dissatisfactory. The overview of our high-level feature extraction system is as shown in Figure 1 and the all 6 runs are described simply as follow:

- A_XJTU_1_1: Basic low-level feature, New Gabor, Compact-SIFT Codeword and RSBag, BAAB fusion method.
- A_XJTU_2_2: Combined all of the runs results.
- A_XJTU_3_3: Basic low-level feature, New Gabor, Compact-SIFT Codeword and the HCRF fusion method.

- A_XJTU_4_4: Basic low-level feature, New Gabor, Compact-SIFT Codeword and voting fusion method.
- A_XJTU_5_5: Basic low-level feature, New Gabor, Compact-SIFT Codeword and BAAB fusion method.
- A_XJTU_6_6: Basic low-level feature and the HCRF fusion method.

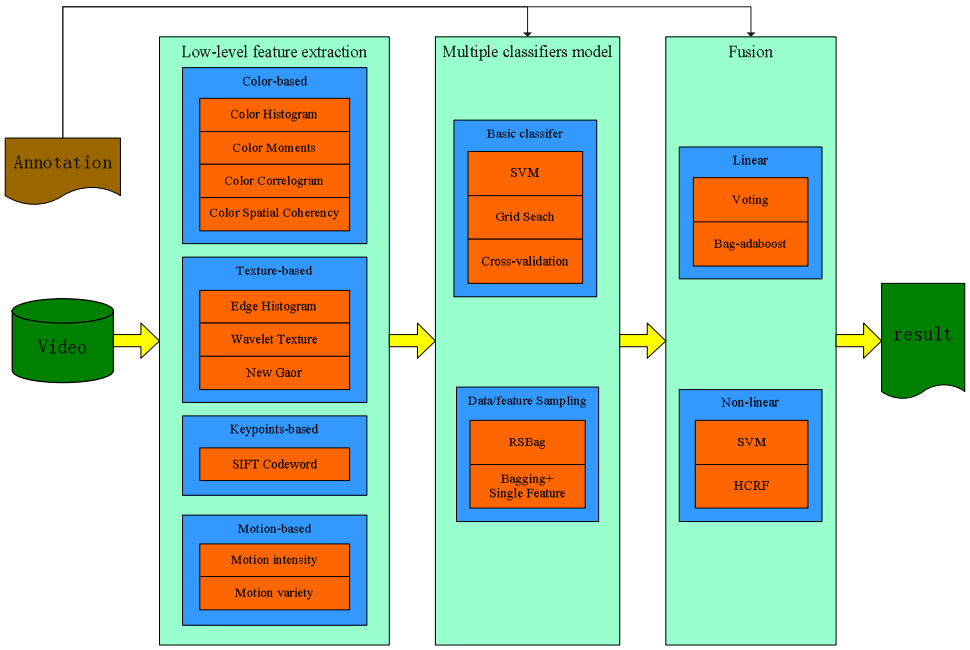

Figure 1: Overview of XJTU high-level feature extraction system

2. BISIC LOW-LEVEL FEATURE

Referring to the previously research of our SIGPRO lab and the TRECVID workshop, we chose 7 basic low-level features based on color、texture、keypoint and motion in the high-level extraction task.

- Color Histogram: same as [1], a 166-dimensional histogram was applied to represent global color in HSV color space.
- Color Moments: a picture was divided to 5×5 grid and the first 3 moments were extracted from the 3 channels in Lab color space over every grid region. Than, the 225- dimensional feature vector was constructed.
- Color Correlogram: a 166- dimensional auto-correlogram histogram was extracted in HSV color space. The distance parameter is 8.
- Color Spatial Coherency: combining the dominant color and the spatial coherency, we proposed a 31-dimensional Color Spatial Coherency descriptor. The first 10 bins describe 10 dominant colors. The second 10 bins describe respectively the proportion between the each dominant color in the picture. The third 10 bins describe the per-dominant-color spatial coherencies same as MPEG-7 [2].The last one bin is the spatial coherency of the entire picture represented in MPEG-7. All of the bins were normalized to (0, 1).
- Edge Histogram: a 64-dimensional global edge histogram with 8 edge direction bins and 8 edge magnitude bins based on a Sobel filter.

– Wavelet Texture: a picture was divided to 3×3 grid and we performed 2D wavelet transform to each grid using the Haar wavelet with the 4 decomposition levels. Then, the 216-dimensional vector was composed of the means and the variances of 12 high frequency sub-bands in each grid region.
– Motion Information: the mean and the variance of motion intensity of one shot.

3. NEW GABOR AND COMPACT-SIFT CODEWORD

We attempted to propose 2 advanced low-level features NEW Gabor texture descriptor and the Compact-SIFT Codeword histogram.

3.1 NEW GABOR

The Gabor function transforms with 8 orientations and 9 scales are utilized to filter the image. The mean and variance of the Gabor filtered image are regarded as Gabor Texture Descriptor. The sum of the means with different scales and orientations of every filtered positive sample keyframes is calculated as the Gabor energy. We can get the Gabor energy values of all the positive samples for each concept. The energy histogram is generated for each concept. The max value column of the histogram is considered as the elementary energy image.

The different scale energy of the elementary energy image and other keyframe are compared to choose the matching 5 scales to get the scale invariant Gabor Texture Descriptor. For the rotation invariant quality, Discrete Fourier Transform is employed on the scale invariant Gabor Texture Descriptors of all the positive samples. Now, the scale invariant and rotation invariant Gabor texture descriptors of the keyframes for each concept are acquired.

This method makes the energy of all the frames aggregate to one energy rank for one concept. So the features have the scale invariant property. Based on it, Discrete Fourier Transform is used for rotation invariant.

3.2. COMPACT-SIFT CODEWORD

In the past year for high-level feature extraction task, SIFT [3] character was popularly adopted. Common method is constructing SIFT codeword histogram [4][5]. At first, a large number of SIFT characters were extracted from lots of video keyframes. Then, by the certain clustering approach such as k-means, these SIFT characters were clustered in k centers. Commonly, we considered the k centers as codebook, and this codebook can describe video semantic information as visual words dictionary. At last, the SIFT codeword histogram of one frame could be constructed by computing the number of every SIFT codeword in the frame.

SIFT codeword histogram can represent video visual semantic information well; however, for getting better performance, the histogram needs a high dimension vector as a rule. However, the high dimension vector will bring some disadvantages. Regarding effectiveness, the high dimension vector will make the overfitting problem while training the classifier model. Another hand, it is time consuming while training and testing the classifier model.

To overcome these disadvantages, this paper proposed an effective approach that could reduce the high dimension vector to the lower dimension vector without performance degradation. First, we constructed the visual codebook with 10000 visual codewords and created the 10000-d SIFT codeword histogram for every frame. The histogram was normalized as follow: if the value of the bin in the histogram is not 0, then it is set to be 1, otherwise, 0. Then, for the certain concept, the importance of the every bin in the histogram would be computed by calculating the ratio

between prior probability of positive and negative samples. The particular steps are following:

1) Gain the prior probability $P(Y \mid e_i)$ and $P(\overline{Y} \mid e_i)$. Where, *i* denote the bin in the SIFT codeword histogram. $Y/\overline{Y}$ represents that the certain concept is present/ absent. e_i describes that the *ith* bin is 1.
2) Computed the likelihood radio O_i using the following formulary:

$$O_i = \frac{P(Y \mid e_i)}{P(\overline{Y} \mid e_i) + \varepsilon}$$

Where, ε is the infinitesimal for avoiding that the denominator equal 0.

So, O_i can represent the of the *ith* bin in the SIFT codeword histogram. The larger O_i, and the better discrimination of the *ith* bin.

3) The lower dimension SIFT codeword histogram called as Compact-SIFT codeword was constructed by reserving and connecting the top *N* bins with the larger O_i. In this paper, the *N* is 100.

Then, the 100 dimensions Compact-SIFT codeword histogram was used to training the classifier model that could gain the higher efficiency and effectiveness.

4. MULTIPLE CLASSIFIERS MODEL

4.1 BASIC CLASSIFIER

In the previous TRECVID high-level extraction task evaluations, SVM was proved as the top performance classifier. So, we chose the LIBSVM [6] as our system basic classifier and the RBF kernel was adopted. As well-known, performance of SVM classifier is remarkably influenced by choice of SVM learning parameters. In this paper, a grid search strategy [7] was adopted and the optimal parameters were selected by 5-fold cross-validation on development data.

4.2 MULTIPLE CLASSIFIERS MODEL

In the recent years, multiple classifiers model was favored in video index and retrieval research. It's a powerful method for increasing the performance of classification [8]. Usually, the multiple classifiers model was constructed by sampling in data space and feature space. At the same time, data and feature sampling can helpfully overcome the problems of classifier overfitting and information redundancy in learning space. In this paper, we used 2 approaches to build the multiple classifiers model:

1) Data bootstrapping and single feature were adopted. The data sampling times is 5.
2) Combining bagging and RSM, RSBag method [1] was adopted to build the multiple classifiers model. Different from [1], the data sampling times is 5 and the feature sampling times is 3.

5. FUSION

In our system, the voting [9], SVM-base, HCRF [10] and the Bootstrap Average

AdaBoost(BAAB) —the four kinds of fusion methods were applied to combine the results of multiple classifiers model. In the voting method, we used the Simple Majority rule to judgment the final decision class that gain the majority score through voting the multiple classifiers output. The SVM-base is the trainable fusion method. First, the score result from multiple classifiers are normalized and concatenated to a score vector. Than, we used these score vectors to train the SVM classifier. The final decision was gained through the classifier.

In this paper, our attention mainly focused on the BAAB method that is a linear weighted fusion method. The weights of the multiple classifiers were assigned by the AdaBoost approach [11]. For one concept, the detailed process for estimating the classifiers weights as follow:

1) The N subsets of the data that divided from the development data for classifiers fusion were set up by the N times bootstrap [12] samplings.
2) Rank the multiple classifiers according to the mean of classifier AP(average precision) that are validated by testing the multiple classifiers through each subset.
3) Estimate the weight of every classifier for each subset adopting the AdaBoost ensemble learning algorithm.
4) Gain the final linear fusion weight of every classifier averaging the weight of every classifier for each subset.

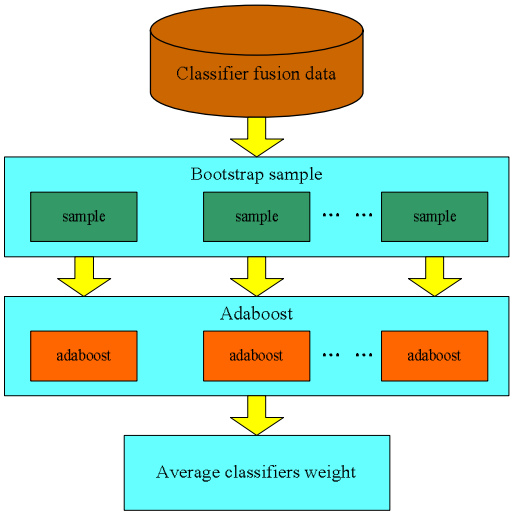

Figure 2: The BAAB classifiers fusion method

6. EXPERIMENTS AND RESULT

We submitted a total of 6 runs, but cause of some mistakes, the A_XJTU_2_2, A_XJTU_3_3 and A_XJTU_6_6 were failed, the performances of the A_XJTU_1_1, A_XJTU_4_4 and A_XJTU_5_5 were not satisfactory, the MAPs were 0.022, 0.040 and 0.034 respectively.

We show the AP of the A_XJTU_4_4 and A_XJTU_5_5 runs in following figure.

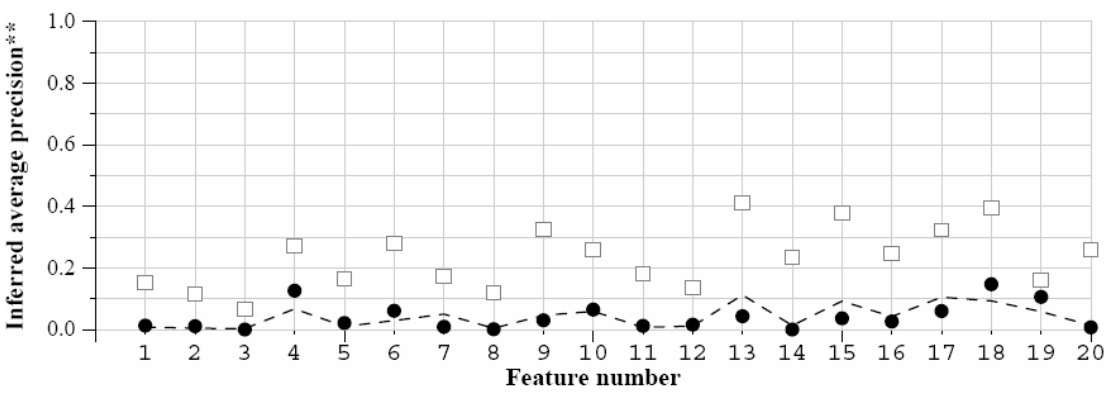

Figure 3: Average Precision Performance of A_XJTU_4 run

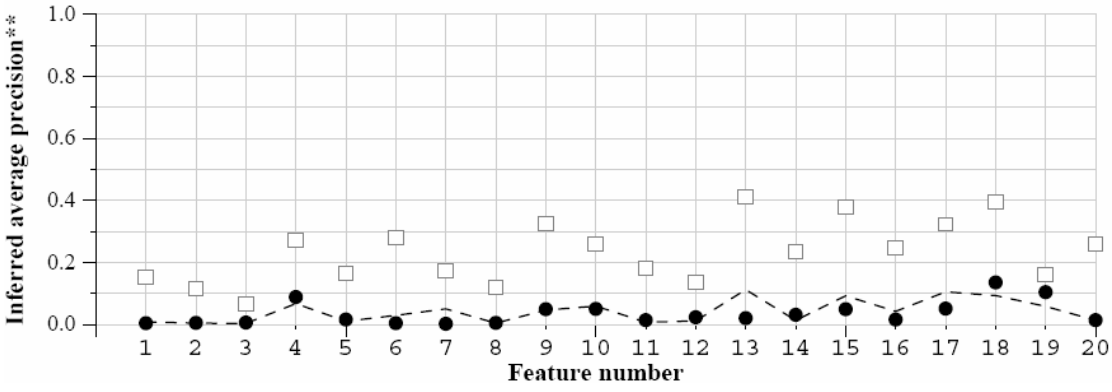

Figure 4: Average Precision Performance of A_XJTU_5 run

References

[1] Murray Campbell,Alexander Haubold.“IBM Rsearch TRECVID-2007 Video Retrieval System.” *in NIST TRECVID Workshop, 2007.*

[2] ISO/IEC/JTC1/SC29/WG11 Doc 15938-3 . Multimedia Content Description Interface-Part3:Visual [S]. 2000-10.

[3] D. Lowe. “Distinctive image features from scale invariant keypoints.” *IJCV, 60(2):91–110,2004.*

[4] Makus Koskela,Mats Sjoberg. “ PicSOM Experiments in TRECVID 2007.” *in NIST TRECVID Workshop, 2007.*

[5] James Philbin, Ondrej Chum, Josef Sivic.”Oxford TRECVid 2008 – Notebook Paper.” *in NIST TRECVID Workshop, 2007.*

[6] C.C. Chang and C.J. Lin, “LIBSVM: a library for support vector machines”, *http://www.csie.ntu.edu.tw/~cjlin/libsvm/ , 2008.*

[7] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. ” A Practical Guide to Support Vector Classification.” *http://www.csie.ntu.edu.tw/~cjlin/papers/guide/, March 12, 2008.*

[8] Rachid Benmokhtar, Benoit Huet. ”Classifier Fusion: Combination Method For Semantic Indexing in Video Content.” *ICANN , part II, pp. 65-74,2006.*

[9] B. Achermann and H. Bunke, “Combination of classifiers on the decision level for face

recognition," *technical repport of Bern University, 1996.*

[10] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol.29, no.10, Oct. 2007, pp.1848-1853.*

[11] Y. Freud and R. Schapire, "Experiments with a new boosting algorithms." *Machine Learning: Proceedings of the 13th International Conference,1996*

[12] L. Breiman, "Bagging Predictors," *Machine Learning, vol. 24, no. 2,pp. 123-140, 1996.*