NII-ISM, Japan at TRECVID 2007: High Level Feature Extraction

Duy-Dinh Le[1], Shin'ichi Satoh[1], and Tomoko Matsui[2]

[1] National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan 101-8430
[2] The Institute of Statistical Mathematics,
4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan 106-8569
ledduy@nii.ac.jp, satoh@nii.ac.jp, tmatsui@ism.ac.jp

Abstract. This paper reports our experiments on the concept detection task of TRECVID 2007. In these experiments, we have addressed two approaches which are selecting and fusing features and kernel-based learning method. As for the former one, we investigate the following issues: *(i) which features are more appropriate for the concept detection task?*, *(ii) whether the fusion of features can help to improve the final detection performance?* and *(iii) how does the correlation between training and testing sets affect the final performance?*. As for the latter one, a combination of global alignment (GA) kernel and penalized logistic regression machine (PLRM) is studied. The experimental results on TRECVID 2007 have shown that the former approach that fuses simple features such as color moments, local binary patterns and edge orientation histogram can achieve high performance. Furthermore, the correlation between the training and testing also plays an important role in generalization of concept detectors.

1 Feature-based Approach

1.1 Framework Overview

In our framework as shown in Figure 1, features are extracted from the input keyframe image. In the training stage, we use these features to train SVM classifiers with RBF kernel. These SVM classifiers are then used to compute raw output scores for the test keyframe image in the testing stage. These output scores can be further combined by a certain fusion method for computing the final output score. In order to return K shots most relevant for one concept query, all normalized final output scores of shots are sorted in descending order and top K shots are returned. In the case of a shot consisting of several subshots, only the maximum score among subshots' scores is used for that shot.

1.2 Feature Extraction

We used three types of features including grid color moments, edge direction histogram (which are described in the baseline system [1]) and the local binary patterns.

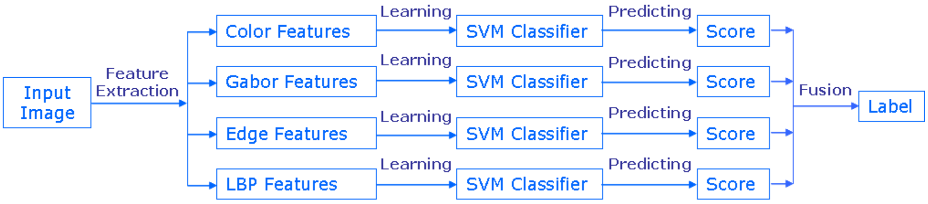

Fig. 1. The evaluation framework.

The extracted features are normalized to zero mean and unit standard deviation and then stored for training and testing. Specifically, the normalized vector $x^{norm} = (x_1^{norm}, x_2^{norm}, ..., x_N^{norm})$ of an input raw vector $x^{raw} = (x_1^{raw}, x_2^{raw}, ..., x_N^{raw})$ is defined as follows:

$$x_i^{norm} = \frac{(x_i^{raw} - \mu)}{\sigma}$$

where x_i^{norm} and x_i^{raw} is the i-th element of the feature vectors x^{norm} and x^{raw} respectively, N is the number of dimensions.

μ is the mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i^{raw}$$

and σ is the standard deviation

$$\mu = \sqrt{(\frac{1}{N}\sum_{i=1}^{N} x_i^{raw} - \mu)^2}$$

1.3 Fusion Method

For each feature, we trained three classifiers for three development sets of TRECVID 2005, TRECVID 2006 and TRECVID 2007 mentioned above. The raw score of each classifier which is the output of SVM classifier is converted to a normalized score by using the function defined in the baseline system [1] that is

$$S_{norm} = \frac{1}{1 + \exp(-S_{raw})}$$

where S_{norm} and S_{raw} are the normalized score and the raw score respectively.

There are two types of fusion. The first one is used to fuse normalized scores trained on different datasets using one feature and the second one is used to fuse scores of different features. The fusion output score is computed as follows:

$$S_{fusion} = \frac{\sum_{i=1}^{N} \alpha_i S_i}{\sum_{i=1}^{N} \alpha_i}$$

where S_{fusion} is the fusion score, S_i is the score to be fused, α_i is the weight of score S_i.

1.4 Experiments

For our experiments, the development set was collected from the datasets of TRECVID 2005, TRECVID 2006 and TRECVID 2007. Specifically, they are the development set of TRECVID 2005 that consists of 137 video programs with 74,523 keyframes described in [1], the development set of TRECVID 2007 that consists of 110 video programs with 21,532 keyframes generated by CLIPS-IMAG and annotated by MCG-ICT-CAS and the test set of TRECVID 2006 in which for each concept, positive samples are relevant shots and negative samples are irrelevant ones. The testing set of TRECVID 2007 consists of 22,084 keyframes extracted from 109 video programs by CLIPS-IMAG.

In order to handle the problem of imbalanced training sets (99% is negative), we select randomly maximum 10,000 samples for each positive and negative set. We use LibSVM [2] to train SVM classifiers with RBF kernel. The optimal (C, g) parameters are found by conducting a grid search with 5-fold cross validation on a subset 1,500 samples stratified selected from the original dataset. Figure 2 shows an example of this searching process.

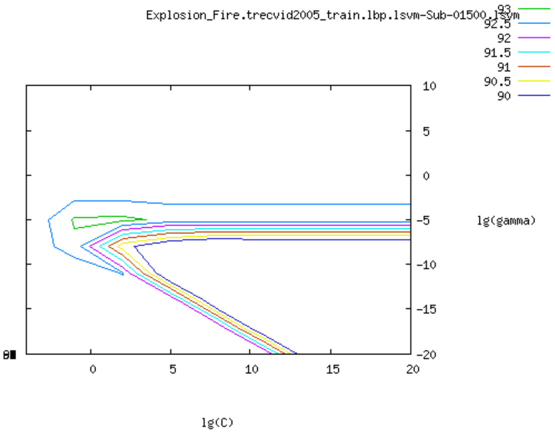

Fig. 2. The output of the grid search process for finding optimal parameters of a SVM classifier. The best performance of 5-fold cross validation is 93.13% corresponding to $(logC, logG) = (2, -5)$

Performance of individual features As shown in Figure 3, GCM feature performs the best while EOH feature performs the worst. GCM feature. GCM feature works very well on concepts such as Maps, Sports and Meeting while LBP feature works well on concepts such as Charts, Boat-Ship, Mountain and People-Marching. EOH feature outperforms the other features the concept Truck. These facts are reasonable since for example, as shown in Figure 4, Sports shots that are found from several submissions are mainly shots with football field.

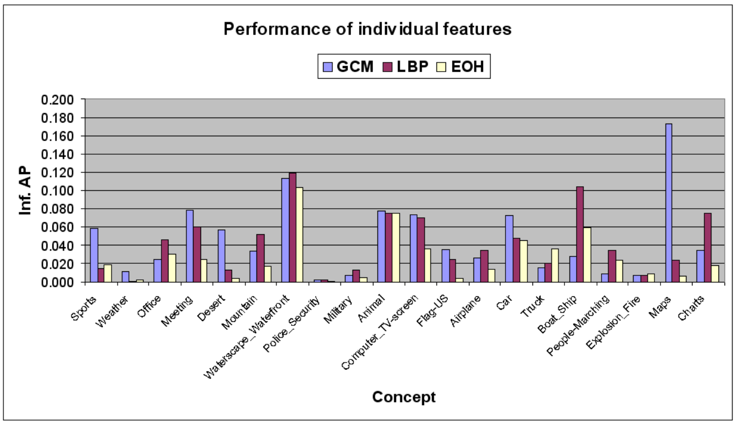

Fig. 3. Performance of individual features. Mean of Inf. AP of GCM, LBP and EOH are 0.047, 0.043 and 0.027 respectively.

Performance of fusing two features Figure 5 shows the performance of fusing any two features. For this experiment, fusion of GCM and LBP is the best while fusion of GCM and EOH and fusion of LBP and EOH have comparable performance.

Performance of fusing all features Figure 6 shows a comparison of fusing features. Obviously, the more features to fuse, the better performance.

Performance of using different training sets Figure 7 shows a comparison of using different training sets. The best performance is obtained when using all the training sets. Furthermore, the performance of using the training set of this year TRECVID 2007 is better that of using the training sets of TRECVID 2005 and TRECVID 2006.

1.5 Discussion

From the experimental result, we have learned several things as follows:

- Color feature is one of the most important features for the concept detection task.
- Local binary pattern feature outperforms edge orientation histogram feature.
- Fusion of many features can help to boost the final performance since features complement each other.
- The correlation of the training set and testing might affect the performance. Furthermore, the more traing data is used, the better performance.

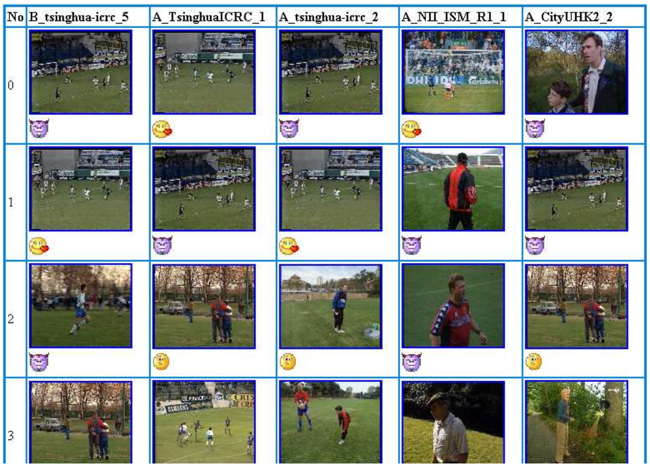

Fig. 4. Top relevant shots for the concept Sports returned by several submissions.

2 Kernel-based Approach

We attempted to use segment-based features in this section. The segments are generated by [3] and each segment has 23 features consisting of the following features:

- areas (in pixel)
- average x
- average y (these two compose center of segment)
- boundary length divided by area
- moment
- average R
- average G
- average B (these three compose average color)
- standard deviation of R
- standard deviation of G
- standard deviation of B (standard deviation of color)
- 12 texture features [4]

The number of segments varies for each key frame and it is difficult to use a conventional method using features with a fixed dimensional vector. Here we treated the segment-based features as a sequence of the segments by sorting them according to the Euclidian distance from the origin. For the segment sequences for each key frame, we applied a combination of global alignment (GA) kernel [5] and penalized logistic regression machine (PLRM) [6, 7]. While standard kernels such as Gaussian and polynomial kernels are vector kernels, the GA kernel is a vector sequence kernel and constructed using similarities based on dynamic time warping (DTW) scores. The GA kernel can effectively handle time series with

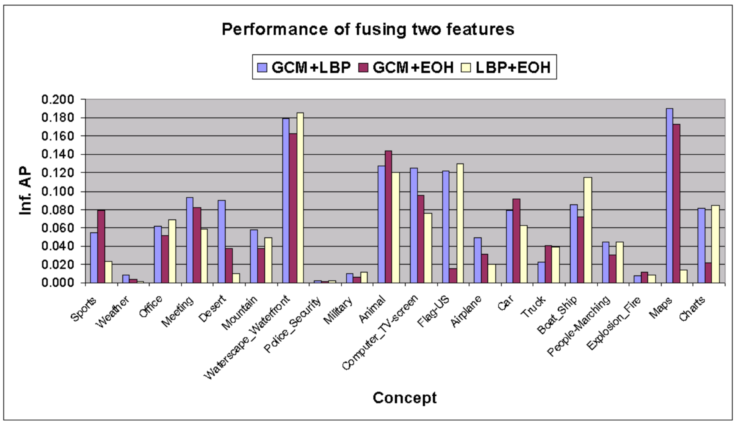

Fig. 5. Performance of fusing two features. Mean of Inf. AP of GCM+LBP, GCM+EOH and LBP+EOH are 0.074, 0.059 and 0.056 respectively.

variable lengths and local dependencies between neighboring states of the time series. It can be considered that we could measure the similarity between the key frames of the segment sequences by utilizing the GA kernel. On the other hand, PLRM is a multi-class classifier and we could estimate one machine for all classes (high-level features) at once. In the concept detection task, each key frame has multiple class labels, so we extended the original PLRM so as to deal with the multiple-labeling problem with fuzzy class representation.

The AP performance was, however, low. One of the main problems is in a sorting method of the segments. We now study new kernels to measure the similarity between the segment-based features.

RunID	Method	Inf. AP
NII-ISM-R1	Fusion of baseline features (GCM, LBP, EOH) trained on TV05, TV06 and TV07	0.101
NII-ISM-R2	Fusion of baseline features (GCM, LBP, EOH) trained on TV05	0.061
NII-ISM-R4	Fusion of baseline features (GCM, LBP, EOH) trained on TV07	0.066
NII-ISM-R5	LBP feature trained on TV05, TV06 and TV07	0.043
NII-ISM-R6	Segment based features, GA kernel with PLR machine	0.020

Table 1. The submissions for high level feature extraction task of NII-ISM

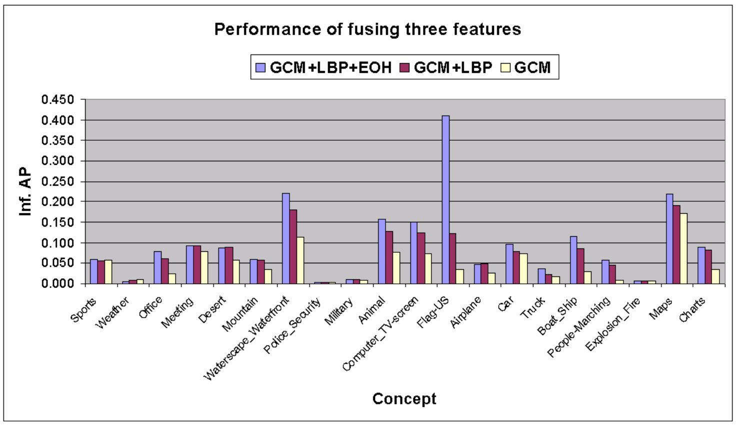

Fig. 6. Performance of fusing features. Mean of Inf. AP of GCM+LBP+EOH, GCM+LBP and GCM are 0.101, 0.074 and 0.047 respectively.

3 Summary

We submitted five runs for TRECVID 2007 high level feature evaluation, as shown in Table 1. The best performance belongs to the system which fuses scores of classifiers trained on different training sets and different features. As shown in Figure 8, our approach achieves high performance while using a small number of features and a simple fusion method.

References

1. Yanagawa, A., Chang, S.F., Kennedy, L., Hsu, W.: Columbia University's baseline detectors for 374 LSCOM semantic visual concepts. Technical report, Columbia University (2007)
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
3. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision **43**(1) (2001) 29–44
4. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(8) (2001) 800–810
5. Cuturi, M., Vert, J., Birkenes, O., Matsui, T.: A kernel for time series based on global alignments. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing. (2007)
6. Tanabe, K.: Penalized logistic regression machines: New methods for statistical prediction 1. In: ISM Cooperative Research Report 143, Estimation and Smoothing Methods in Nonparametric Statistical Models. (2001) 163–194
7. Tanabe, K.: Penalized logistic regression machines: New methods for statistical prediction 2. In: Workshop on Information-Based Induction Science. (2001) 71–76

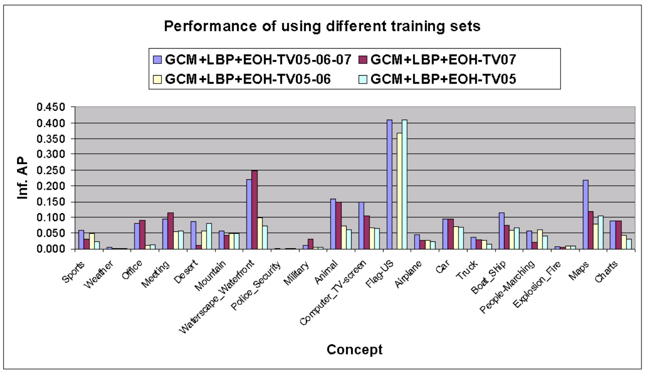

Fig. 7. Performance of using different training sets. Mean of Inf. AP of GCM+LBP+EOH-TV05-06-07, GCM+LBP+EOH-TV07, GCM+LBP+EOH-TV05-06, GCM+LBP+EOH-TV05 are 0.100, 0.066, 0.061 and 0.060 respectively.

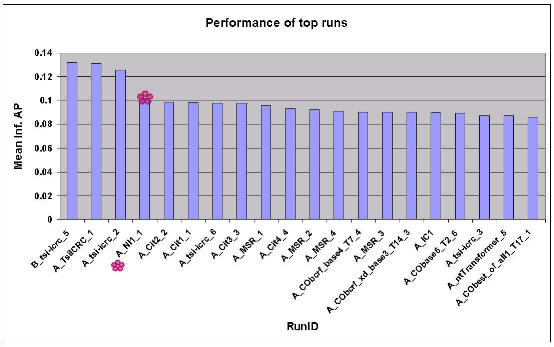

Fig. 8. Performance of top 20 runs. Our best run is ranked the fourth.