The France Telecom Orange Labs (Beijing) Video Semantic Indexing Systems – TRECVID 2010 Notebook Paper

Kun Tao[1], Yuan Dong[1, 2], Jiqing Liu[2], Shan Gao[2], Jiwei Zhang[2], Tianxiang Zhou[2], Guorui Xiao[2], Hongliang Bai[1], Xiaofu Chang[1], Chengyu Dong[1]

[1]France Telecom Orange Labs (Beijing), Beijing, 100190, P.R.China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, P.R.China

ABSTRACT

In this paper, we described the latest video semantic indexing systems developed at France Telecom Orange Labs (Beijing). In our previous systems for TRECVID 2009, the features of color, edge, texture and SIFT were used. This year, some new features based on local descriptors were added for performance improvement. Three Full runs (130 concepts) based on later fusion and one Light run (10 concepts) based on early fusion were submitted, among which we compared the results of unsupervised and supervised late fusion. The effect of cross-domain fusion was also investigated. The run of F_A_FTRDBJ-HLF-2_2 achieved our best MAP of 0.075, which was based on a two-step linear weighted fusion of 19 features. In particular, we used a group of unified weights for all concepts. Such a strategy showed good generalization ability on diverse internet video data.

1. INTRODUCTION

This year, we submitted 4 runs for the video semantic indexing (SIN) task [1]. We followed the basic system structures which were used in our TRECVID 2009 evaluation [2], but some new elements were added so as to face the challenge of using a new internet video corpus:

- Most of our experiments were taken on a 30 concept corpus. A group of unified weights were also trained on that corpus and were used as the fusion parameters of all 130 concepts.
- We integrated several kinds of new features into our systems, including soft-assigned SIFT histograms and the histogram of oriented edges (HOG) features.
- We investigated the effect of cross-domain fusion but didn't see significant improvement in the final evaluation.

As the start of a new round, a new corpus of internet videos (IACC) [1] was used in semantic indexing task. The diversity of video types brought more difficulties to the detection of semantic contents. The change of concept number was also a great challenge to computational capability. We didn't try to make exhaustive experiments on all 130 concepts, but selected a 30 concept subset for our evaluative experiments. The subset includes the 10 concepts of "Light-10" corpus and the other 20 concepts selected from "Full-130" corpus. We call our subset "FT-30" corpus, and the subset selected by NIST for the evaluation of "Full" submission is named as "TREC-30". FT-30 includes some typical concepts of scenes, objects, programs and events:

Airplane_Flying, Anchorperson, Animal+, Beach, Bicycles, Boat_Ship*, Bus*, Cats, Chair, Charts, Cityscape*, Classroom*, Construction_Vehicles, Crowd, Dark-skinned_People+, Demonstration_Or_Protest*, Female_Person, Flowers+, Hand*, House_Of_Worship, Instrumental_Musician, Laboratory, Nighttime*, Roadway_Junction, Running+, Shopping_Mall, Singing*, Sitting_Down+, Sports, Telephones*.*

10 concepts with "*" came from Light-10 corpus, and 5 other concepts with "+" were finally selected into TREC-30. Our internal evaluation and parameter selection were all based on FT-30. Then selected parameters were used for Full-130 testing.

4 categories of low-level features were used in our system: color, edge, texture and local descriptors. Based on the features used in 2009, two kinds of new features were added: soft-assigned SIFT histogram and the histogram of HOG descriptors. Probabilistic latent semantic analysis (PLSA) [3], instead of LSA, was used in the dimension reduction of 3-level pyramid histogram features. The details will be discussed in Section 2.

For the first 3 Full runs, SVM was used to detect concepts by each low-level feature. Then the classification scores were combined by a cascaded structure. We also submitted a Light run using multiple kernel SVM for comparison.

This year, the researches on cross-domain learning and usage of ontology relations were encouraged. Our third run is a cross-domain run which used the shots and labels of TRECVID 2005-2009. No semantic context information was used because we thought that both the great imbalance of positive sample numbers in dataset and

the uncertainty of co-occurrence relationships of different concepts are big problems. We were not so confident about the results of using ontology relationships based on existing methods.

The brief summarizations and final MAPs of our 4 submitted runs are listed below:

- F_A_FTRDBJ-HLF-1_1: classifier-level-combination of 19 low-level feature SVMs with equal weights. MAP = 0.070.
- F_A_FTRDBJ-HLF-2_2: linear weighted combination of 19 feature SVMs through logistic regression. MAP = 0.075.
- F_C_FTRDBJ-HLF-3_3: cross-domain fusion between the results of F_A_FTRDBJ-HLF-2_2 and the results of 05-09 TRECVID models. MAP = 0.070.
- L_A_FTRDBJ-HLF-4_4: kernel-level-combination of 14 low-level features with equal weighted multiple kernel learning. MAP = 0.063.

2. THE LOW-LEVEL FEATURES

A total of 19 low-level visual features were used in our semantic indexing systems. They basically belong to 4 categories: color features, edge features, texture features and local descriptor features. The first 3 categories include 7 types of features which have been proved effective in last year: Color Auto-Correlograms (CAC), Color Coherence Vector (CCV), Grid Color Moments (GCM), Edge Coherence Vector (ECV), Edge Direction Histogram (EDH), Gabor feature (Gabor) and Local Binary Patterns (LBP) .They are regarded as one group named CEGL. The local descriptors include SIFT and HOG, which can be transformed into different kinds of histograms by means of Bag-of-Visual-Words. More details will be described below.

2.1 SIFT and PHOW

SIFT descriptors can be extracted by three means: The first is based on traditional DoG interest point detection [4]. The second method also uses the DoG key-points but doesn't calculate the orientation of key-points. All descriptors are extracted at the orientation of $\theta = 0$. The third method samples key-points on a grid with spacing of 6 pixels, and the orientations will not be calculated too. Features based on above three methods are named as "SIFT", "SIFT-NO-ORIENTATION" and "DENSE-SIFT" respectively.

Then the Bag-of-Visual-Words method can be used. Codebooks are trained respectively for the SIFT descriptors extracted by above 3 methods. Each codebook has 512 visual words. If we calculate the histogram of visual words for whole image, it can be named as Histogram of Visual Words (HOW). If a pyramid representation is used, it will become Pyramid HOW (PHOW) [5]. 2-level pyramids used 1x1 and 2x2 regions, which is named as "2L-PHOW".

TABLE I
Four Groups of Features

Group Name	Feature Name	Dim.
CEGL	Color Auto-Correlograms (CAC)	256
	Color Coherence Vector (CCV)	360
	Grid Color Moments (GCM)	108
	Edge Coherence Vector (ECV)	320
	Edge Direction Histogram (EDH)	365
	Gabor feature (Gabor)	240
	Local Binary Patterns (LBP)	256
S6	SIFT.HOW	512
	SIFT.2L-PHOW	2560
	SIFT. 3L-PHOW-PLSA	512
	DENSE-SIFT.HOW	512
	DENSE-SIFT.2L-PHOW	2560
	DENSE-SIFT. 3L-PHOW-PLSA	512
SS3	SIFT.HOW-SOFT	512
	SIFT-NO-ORIENTATION. HOW-SOFT	512
	DENSE-SIFT. HOW-SOFT	512
H3	HOG.HOW	512
	HOG.2L-PHOW	2560
	HOG. 3L-PHOW-PLSA	512

3-level pyramid uses 1x1, 2x2 and 4x4 regions. Last year, we used latent semantic analysis (LSA) [6] to get the low-dimension representation of 3L-PHOW. This time it was replaced by the method of probability latent semantic analysis (PLSA) [3]. PLSA is based on a mixture decomposition derived from a latent class model. This results in a principled approach which has solid foundation in statistics. In our experiments, all 3L-PHOW features were projected to the 512D 3L-PHOW-PLSA representation.

2.2 Soft Assignment

We also used a soft assignment technology to build up the histograms. Although the codebook is the same as hard assignment, each descriptor can contribute to the top 3 nearest codebook centers in soft assignment. The weights of contribution are decided by the distances and the ranks of neighbors.

$$Weight_{ni} = \frac{1/(ni * Dist_{ni})}{\sum_{i=1}^{3}(1/(i * Dist_i))} \qquad ni = 1,2,3 \tag{1}$$

The $Dist_{ni}$ is the distance between the descriptor and the *ni-th* nearest center. The sum of weights is 1.0 and the

three weights can be added to the corresponding bins of histogram.

Though experiments, we found that the soft assignment can improve the result of SIFT obviously. But its effect on DENSE-SIFT is not as good as we expected. The number of DENSE-SIFT descriptors is hundreds of times more than SIFT descriptors, but their vocabularies are at the same size. That might be a reason which reduced the effect of soft assignment. In the future, more experiments will be taken to prove it. Finally in our systems 3 kinds of SIFT histogram features were calculated based on soft assignment and were regarded as one group.

2.3 HOG Descriptors

The HOG descriptor has been proved valuable in the fields such as human detection, so we integrated it into our systems to build up HOW features. Similar to the method in [7, 8], we extracted 124D descriptors. Then a codebook of 512D was built up and corresponding 1-level to 3-level histograms were calculated. The 3-level histograms were also projected to low dimensional space by PLSA.

Above features were organized into different groups. More details of feature groups are showed in Table I.

3. THE FUSION STRATEGIES

Support vector machine (SVM) has been proved to be a very reliable classification method and we chose it to build our classifiers. Last year, three different kinds of kernels were used for different features: Chi-square kernel, Euclidean Exponential kernel and RBF kernel. This year, only Chi-square kernel was used. Of course, some of the features should be normalized in order to satisfy related constrain condition. Then different strategies including late fusion and early fusion were used to get combined results.

In our internal evaluation, 60% shots of the IACC.1.tv10.training [1] dataset were used to train SVM models, 20% for fusion parameter training, and the other 20% were for evaluation. Then selected methods and parameters were applied to the final models.

3.1 Two-Step Late Fusion

AS the 19 features were organized into 4 groups, we used a two-step cascaded fusion strategy to combine the results of different features. First the intra-group fusion was done in each group. Then the results of different groups were combined into the final score by inter-group fusion (Fig. 1).

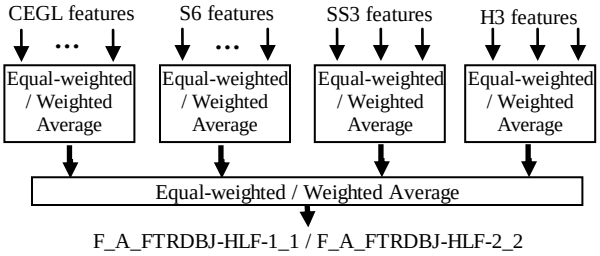

Fig. 1 Two Step Fusion Structure

In the run of F_A_FTRDBJ-HLF-1_1, we used equal weights in both fusion steps. For F_A_FTRDBJ-HLF-2_2, the linear weighted average was used. The weights for both steps were trained by logistic regression using the LIBLINEAR toolkit [9]. Above two runs belong to unsupervised and supervised late fusion respectively.

There are 130 concepts to be tested if someone wishes to submit a Full run, which need a lot of computing time. At the same time, the great number of shots and the usage of new features will also add calculation amount. For the limitation computational capability, we had to use the FT-30 subset to evaluation the effects of different features, classifier parameters and fusion strategies, which resulted in that we couldn't train the fusion weights for every concept out of Full-130 respectively. Luckily we found that using a group of unified weights for all concepts will not cause obvious performance drop. The evaluation results on FT-30 corpus are showed in Table II.

In the F_A_FTRDBJ-HLF-2_2 run, only one group of weights were trained from the FT-30, which were used for all 130 concepts. In fact, our final results on Light-10 concepts are not very good. The MAP of F_A_FTRDBJ-HLF-2_2 on Light-10 is 0.0628, but the MAP on TREC-30 is 0.75. It means that our method worked quite well on the other 20 concepts including 15 inexperienced concepts. Although such a fusion strategy was initially an unwilling choice and seems conservative, it has its advantage in generalization ability.

TABLE II
Two-step Fusion Results on FT-30 Corpus

Fusion Method / Group	Equal weighted (MAP)	Logistic regression with respective weights (MAP)	Logistic regression with unified weights (MAP)
CEGL	0.0634	0.0594	0.0640
H3	0.0513	0.0509	0.0513
SS3	0.0635	0.0626	0.0636
S6	0.0648	0.0650	0.0640
2-Step Fusion	0.0704	0.0699	0.0694

3.2 Cross-Domain Fusion

In the run of F_C_FTRDBJ-HLF-3_3, we used the TRECVID 2005-2009 video corpus and labels to build up our cross-domain models. In our internal evaluation, only the 3 features in group SS3 were used. Both of the data-level and classifier-level cross-domain fusions were investigated on FT-30.

Table III shows a result of cross-domain experiments on FT-30, the weights for fusion were based on the estimated MAP of each isolated feature. From the Table III, we can see that a simple data-level cross-domain fusion can't bring improvement because there are many apparent differences between samples came from different datasets. In fact, we found that only very few concepts with too little positive samples in IACC.1.tv10.training (such as Airplane_Flying) can get improvement from the additional 05-09 samples.

When the SS3 classifiers trained on 05-09 dataset were applied to test TREC2010 samples, they resulted in very poor MAP. But when their results and the results of classifiers trained on 2010 dataset were combined by means of weighted average, we got a better MAP in the internal evaluation. Thus we combined the results of F_A_FTRDBJ-HLF-2_2 run and the results of SS3 model-2 by weighted averaged, and got the results of F_C_FTRDBJ-HLF-3_3 run.

But the evaluation results made by NIST show that the F_C_FTRDBJ-HLF-3_3 didn't reach expected performance. The MAP of F_C_FTRDBJ-HLF-3_3 is lower than that of F_A_FTRDBJ-HLF-2_2. It seems that the cross-domain learning is still a challenging problem to us.

TABLE III
Cross-Domain Results based on SS3 and FT-30 Corpus

Model ID	Model Description	MAP
Model-1	Models trained on 2010 dataset	0.084
Model-2	Models trained on 05-09 dataset	0.027
Model-3	Models trained on 2010+05-09 dataset (data-level fusion)	0.075
Model-4	Equal weighted fusion of Model-1and Model-2 (classifier-level fusion)	0.063
Model-5	Weighted Average fusion of Model-1and Model-2 (classifier-level fusion)	0.086

3.3 Early fusion: Kernel-level combination

Last year, a run based on kernel-level-combination (multiple kernels with equal weights) won our 2nd best MAP [2]. This year, we tried it again in order to make a comparison with late fusion runs. The run of L_A_FTRDBJ-HLF-4_4 was based on 14 low level features (Excluding the 3 H3 features and 2 3L-PHOW-PLSA features in S6). The training of multi-kernel models is time-consuming, so we only submitted a Light run of 10 concepts.

If only the 10 concepts of Light-10 are investigated, the MAP of L_A_FTRDBJ-HLF-4_4 is better than the MAPs of three late fusion runs (TABLE IV). Such comparison shows that early fusion methods are still very effective without regard for the computation cost.

TABLE IV
The MAP of 4 Submitted Runs on Light-10 Corpus

Run Name	MAP
F_A_FTRDBJ-HLF-1_1	0.057
F_A_FTRDBJ-HLF-2_2	0.063
F_C_FTRDBJ-HLF-3_3	0.057
L_A_FTRDBJ-HLF-4_4	0.064

4. CONCLUSION

This is the second time we participated in TRECVID. Comparing with the HLFE task of last year, the changes of dataset and concept number brought great challenges to every participator. Our attempts on new features and new fusion strategies brought some advantages, which lead to a acceptable result. But the results of cross-domain models are unsatisfactory. In the future, we plan to pay more attention on cross-domain learning and try to find a more efficient combination of features in order to reduce the computational cost.

5. REFERENCES

[1] "Guidelines for the TRECVID 2010 Evaluation," http://www-nlpir.nist.gov/projects/tv2010/tv2010.html.

[2] Y. Dong, etc. "The France Telecom Orange Labs (Beijing) Video High-level Feature Extraction Systems –TrecVid 2009 Notebook Paper," http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2009.

[3] Thomas Hofmann, "Probabilistic Latent Semantic Indexing," Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999.

[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 60 (2): 91-110, 2004.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006.

[6] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE PAMI*, 30 (4), 2008.

[7] "Object Detection using Histograms of Oriented Gradients". http://www.pascal-network.org/challenges/VOC/voc2006/slides/dalal.pdf.

[8] Jianxiong Xiao et al. "SUN Database: Large-scale Scene Recognition from Abbey to Zoo",CVPR 2010.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research 9(2008), 1871-1874.