The France Telecom Orange Labs (Beijing) Video High-level Feature Extraction Systems – TrecVid 2009 Notebook Paper

Yuan Dong[1, 2]*, Xianyu Zhao*[1]*, Zhongxuan Liu*[1]*, Chengyu Dong*[1]*, Jiqing Liu*[2]*, Liang Lu*[2]*, Zhe Wei*[2]*,*
Guorui Xiao[2]*, Shiguo Lian*[1]*, Ronggang Wang*[1]*, Kun Tao*[1]

[1]France Telecom Orange Labs (Beijing), Beijing, 100080, P.R.China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, P.R.China

ABSTRACT

In this paper, we described the video high-level feature extraction systems developed at France Telecom Orange Labs (Beijing). In our systems, four categories of low-level visual features, namely color, edge, texture and SIFT local descriptors, were extracted. Two approaches to fusing the representative capabilities of these visual features were investigated for different runs. Under the setting of late fusion, separate SVM classifiers were constructed for each low-level visual feature and their outputs were combined in a weighted manner. While under the setting of early fusion, a composite kernel was constructed to merge multiple kernels of various visual features and one SVM was trained on such composite kernel. The evaluation results on the 2009 TrecVid high-level feature extraction task were presented, among which the A_FTRD-HLF-5 run achieved an MAP of 0.17 and was above the median for all concepts.

1. INTRODUCTION

We submitted 6 runs for the high-level feature extraction (HLFE) tasks [1]. All of these systems for HLFE were trained on publicly available annotations [2].

The generic structure of these 6 runs includes key frames sampling, low-level feature extraction, classifiers for high-level features.

Firstly, a subset of key frames was sampled from each shot. For each shot, its middle frame was extracted as reference key frame (RKF). Meanwhile, for shots containing sub-shots (based on shot boundary information provided by NIST), one NRKF frame was obtained for each sub-shot.

These key frames were then fed to following low-level feature extraction modules. In our systems, we used 4 categories of low-level features, namely color, edge, texture and SIFT local descriptors, which would be discussed in details in Section 2.

Support vector machines (SVMs) are constructed to detect high-level concepts on top of these low-level visual representations. In Section 3, we present two fusion strategies to combine the representative capabilities of various low-level features, i.e. kernel-level-combination and classifier-level-combination.

Finally, test shots were ranked by the maximum concept detection scores over the key frames within that shot.

A brief summarization of our submitted 6 runs is listed below; and, in Table I, we present their mean average precision (MAP) performances on the 20 concepts of TrecVid 2009.

- A_FTRD-HLF-1: classifier-level-combination of 12 low-level feature SVMs with equal weights.
- A_FTRD-HLF-2: kernel-level-combination of multiple kernels for 12 low-level features with equal weights.
- A_FTRD-HLF-3: classifier-level-combination with logistic regression over 5 low-level feature groups.
- A_FTRD-HLF-4: kernel-level-combination with multiple kernel learning (MKL) for 6 concept categories.
- A_FTRD-HLF-5: classifier-level-combination with logistic regression over 2 low-level feature groups.
- A_FTRD-HLF-6: kernel-level-combination with multiple kernel learning for 20 concepts.

TABLE I
THE PERFORMANCES OF 6 RUNS FOR HIGH-LEVEL FEATURE EXTRACTION

RUNID	MAP
A_FTRD-HLF-1	0.1645
A_FTRD-HLF-2	0.1658
A_FTRD-HLF-3	0.1620
A_FTRD-HLF-4	0.1642
A_FTRD-HLF-5	0.1704
A_FTRD-HLF-6	0.1495

2. LOW-LEVEL VISUAL FEATURES

A total of 12 low-level visual features were used in our HLFE systems. They basically belong to 4 categories:

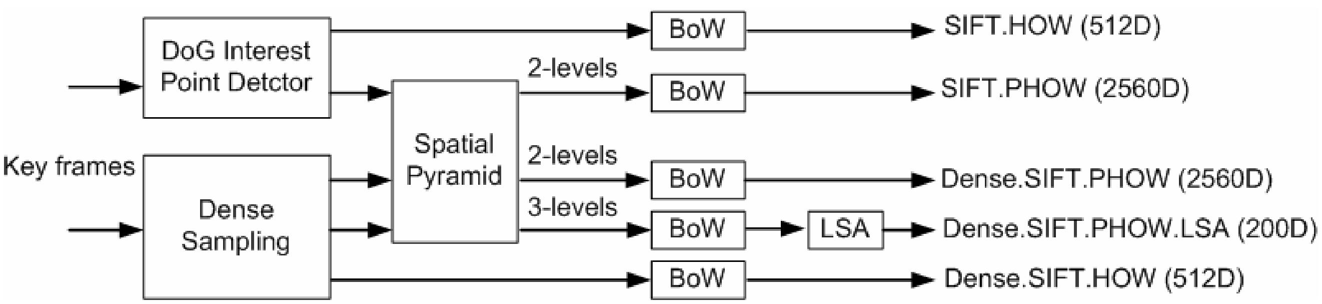

Fig. 1 SIFT local descriptors

TABLE II
COLOR, EDGE AND TEXTURE FEATURES

Feature	Description	Dim.
CAC.Global	Color Auto-Correlograms on whole image	256
CCV.GRI5	Color Coherence Vector with GRI5 spatial layout	360
GCM.GR4x3	Grid Color Moments with GR4x3 spatial layout	108
ECV.GRSH	Edge Coherence Vector with GRSH spatial layout	320
EDH.GRSH	Edge Direction Histogram with GRSH spatial layout	365
Gabor.GRSH	Gabor filter with GRSH spatial layout	240
LBP.Global	Local Binary Patterns on whole image	256

SIFT local descriptors, color features, edge features and texture features.

2.1 SIFT local descriptors

SIFT descriptors [3] were extracted at some DoG interest points as well as at points on a grid with spacing of 6 pixels. These sparse and dense SIFT descriptors were then quantized into visual words using K-means clustering. In our systems, two separate codebooks were used for sparse and dense SIFT descriptors respectively. Each codebook has 512 visual words.

Histograms of these SIFT visual words are computed for whole image as well as for sub-regions to incorporate spatial information. Histograms of visual words for each sub-region were then concatenated into so called Pyramid Histogram of Visual Words (PHOW) [4] [5]. We used a spatial pyramid of 1x1 and 2x2 regions in our PHOW representation.

In our systems, we also built a PHOW representation of 3 levels (having 1x1, 2x2 and 4x4 regions) and projected such representation to a low dimensional latent semantic space with latent semantic analysis (LSA). This was similar to the method used in [6].

A summary of the use of SIFT local descriptors was shown in Fig. 1.

2.2 Color, edge and texture features

There were 3 types of color feature descriptors in our systems, including Color Auto-Correlograms (CAC), Color Coherence Vector (CCV) and Grid Color Moments (GCM). Two types of edge features, Edge Coherence Vector (ECV) and Edge Direction Histogram (EDH), were used in our systems. And for texture features, we used Gabor feature and Local Binary Patterns (LBP). These features were also extracted with proper spatial layout partition to add spatial information to them, e.g. GRI5, GRSH, GR 4x3 [7].

A summary of these color, edge and texture features was presented in Table II.

3. FUSION STRATEGIES

Support vector machines (SVMs) were used for supervised learning of high-level semantic concepts over low-level visual features. To combine the representative capabilities of various low-level features, two fusion strategies were tried in our systems. One was based on classifier-level combination, or so called late fusion. The other was based on kernel-level combination, or early fusion.

3.1 Late fusion: Classifier-level combination

Under this setting, for each low-level feature, an SVM was trained using that representation to detect high-level concepts. Detection scores from these SVMs were then combined in a weighted manner to derive final score. In Table III, we summarized the kernel configurations of SVMs for various low-level features and corresponding MAP performances on TrecVid 2009 dataset.

TABLE III
THE CONFIGURATIONS AND PERFORMANCES OF LOW-LEVEL FEATURE SVMS

Feature	Kernel	MAP
SIFT.HOW	χ^2 exponential kernel	0.0606
SIFT.PHOW	Pyramid χ^2 exponential kernel	0.0683
Dense.SIFT.HOW	χ^2 exponential kernel	0.0996
Dense.SIFT.PHOW	Pyramid χ^2 exponential kernel	0.1244
Dense.SIFT.PHOW. LSA	Euclidean exponential kernel	0.0907
CAC.Global	RBF kernel	0.0318
CCV.GRI5	RBF kernel	0.0403
GCM.GR4x3	RBF kernel	0.0426
ECV.GRSH	RBF kernel	0.0630
EDH.GRSH	RBF kernel	0.0485
Gabor.GRSH	RBF kernel	0.0328
LBP.Global	RBF kernel	0.0515

The exponential chi-square kernel for two image *x* and *y* is defined to be [8]:

$$K(x,y)=\exp\left(-\chi^2\left(H_x,H_y\right)/A\right), \qquad (1)$$

where H_x is the histogram of image *x*; the χ^2 distance is evaluated as

$$\chi^2\left(H_x,H_y\right)=\frac{1}{2}\sum_{i=1}^{N}\frac{\left(H_x(i)-H_y(i)\right)^2}{H_x(i)+H_y(i)}, \qquad (2)$$

And, the scaling parameter A is set to be the mean value of χ^2 distances of histograms over development data.

For the 2-levels spatial pyramid representation used in our submissions, the pyramid χ^2 exponential kernel is [5]:

$$K(x,y)=\sum_{l=0}^{1}\sum_{i=0}^{2^{2l}-1}\exp\left(-\chi^2\left(H_x^{l,i},H_y^{l,i}\right)/A^{l,i}\right), \qquad (3)$$

where $H_x^{l,i}$ stands for the histogram of the *i*-th sub-region at the *l*-th level. In (3), all sub-regions are weighted equally.

Parameters in Euclidean exponential and RBF kernels are tuned by a grid search over development data.

We used SVMTorch [9] as the SVM solver. Similar to [10], in order to handle imbalance in the number of positive and negative samples of high-level concepts, we adjust the weights of the positive and negative samples according to their class priors on training data.

In TABLE III, we see that among the 12 low-level features we used, dense gray sift with spatial pyramid information presents best detection performance.

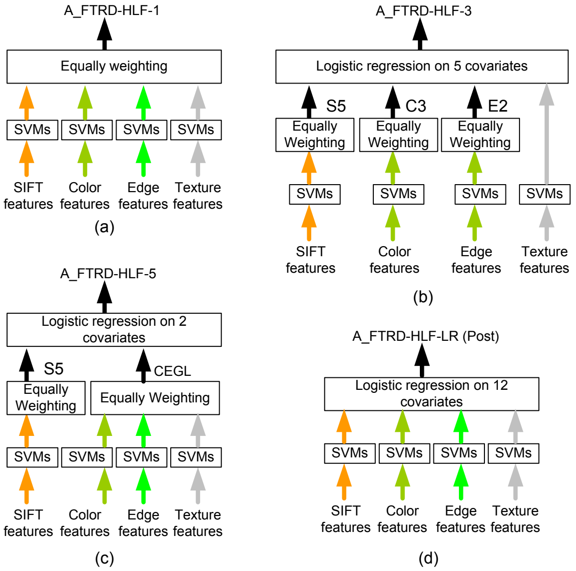

Fig. 2 Different strategies of classifier-level combination

TABLE IV
THE PERFORMANCES OF CLASSIFIER-LEVEL COMBINATIONS

RUNID	MAP
S5	0.1492
C3	0.0596
E2	0.0724
CEGL	0.1225
A_FTRD-HLF-1	0.1645
A_FTRD-HLF-3	0.1620
A_FTRD-HLF-5	0.1704
A_FTRD-HLF-LR (Post)	0.1466

In our 6 submitted runs, 3 out of them fell into classifier-level combination. They differed in the setting of combination weights, as shown in Fig.2.

For A_FTRD-HLF-1, all low-level feature SVMs were combined with equal weights. For A_FTRD-HLF-3 and A_FTRD-HLF-5, some combination weights were optimized through logistic regression. Instead of carrying out logistic regression over all 12 low-level feature SVMs, they are split into several categories. Within each category, we combined corresponding low-level feature SVMs with equal weights. After this first stage of fusion, we hope the fused detection scores out of each category would be more stable, on which logistic regression was then applied to derive more robust fusion weights. A_FTRD-HLF-3 and A_FTRD-HLF-5 differed in the granularity of the category setting. In Fig. 2, "S5" represents the combination of 5 SVMs related to sift features; similarly, "C3" is for the combination of 3 color feature SVMs and "E2" for combination of 2 edge feature SVMs. "CEGL" stands for the combination of 7 color, edge and texture feature SVMs. In our post evaluation, we also carried out logistic regression over all 12 low-level feature SVMs, i.e. A_FTRD-HLF-LR (Post) as shown in Fig. 2 (d).

In TABLE IV, we listed the performance of these fusion systems. Although the performance of any single systems using color, edge or texture features is inferior to those based on sift local descriptors, their combination, i.e. the "CEGL" system, achieves comparable performance with sift-based systems. The combination of sift local descriptors, color, edge and texture features is shown to further promote performance than using any of them alone. Through carefully training the combination weights by logistic regression, slightly better performance is obtained over equally weighting mode. Though, we also find that it is not robust to carry out logistic regression over all 12 low-level feature SVMs.

3.2 Early fusion: Kernel-level combination

This is a kind of early fusion strategy. A composite kernel is constructed by weighted linear combination of multiple kernels for various low-level features:

$$K\left(x,y\right)=\sum\nolimits_{f}\beta_f K_f\left(L_x^f,L_y^f\right), \tag{4}$$

where L_x^f is one low-level feature whose corresponding kernel is K_f and β_f is the kernel weight. It is expected that such composite kernel could measure input similarities from various aspects (e.g. local gradient information, color, edge, texture, etc.) by integrating similarity measures with respect to various low-level features. An SVM is then trained with this composite kernel to detect high-level semantic concept. This multiple kernel approach has also been studied in [5], [11].

In our 6 submitted runs, 3 submissions belong to this category. The composite kernels are constructed based on linear combinations of the 12 kernels listed in Table III. Different kernel weighting schemes were tried.

For A_FTRD-HLF-2, these 12 low-level feature kernels were combined with equal weights.

For A_FTRD-HLF-6, the kernel combination weights were learned through multiple kernel learning (MKL) for each high-level concept. We adopted the MKL approach proposed in [11]. To prevent MKL from deriving two sparse solutions on the kernel weights, a regularization term $\mathcal{R}=\varphi\sum\nolimits_{f}\beta_f^2$ is added to the MKL objective function [12]:

$$\begin{aligned}&\min_{\beta_f,w_f,b,\xi_i}\frac{1}{2}\sum_{f=1}^{F}\frac{1}{\beta_f}\left\|w_f\right\|_2^2+C\sum_{i=1}^{N}\xi_i+\varphi\sum_{f=1}^{F}\beta_f^2\\&s.t.\ y_i\left(\sum_{f=1}^{F}\left\langle w_f,L_i^f\right\rangle+b\right)\geq 1-\xi_i\ \ \forall i\\&\qquad\xi_i\geq 0\\&\qquad\sum_{f=1}^{F}\beta_f=1,\ \beta_f\geq 0\ \forall f,\end{aligned}\tag{5}$$

where b, ξ_i and C are the standard SVM bias, slack variables and regularization term; w_f is the primal SVM weights associated kernel K_f (with underlying feature transformation function ϕ_f, i.e. $K_f\left(L_x^f,L_y^f\right)=\left\langle\phi_f\left(L_x^f\right),\phi_f\left(L_y^f\right)\right\rangle$):

$$w_f=\beta_f\sum_{i=1}^{N}\alpha_i y_i\phi_f\left(L_i^f\right) \tag{6}$$

TABLE IV
CONCEPT CATEGORIES

Category	Concept
CST	Cityscape, Street, Traffic-intersection
IPA	Classroom, People-dancing, Person-eating, Person-playing-musical-instrument, Singing
BSW	Boat-ship, Bridge, Harbor
HND	Hand
FHC	Female-human-face-closeup

The parameter, φ, controls the level of sparsity in MKL solution. Larger values of φ would drive MKL towards a more uniform set of weights. In our submissions, φ was set to 10.

As for some concepts there are very limited number of positive training samples, to increase the robustness of learned kernel combination weights we also tried a mode of concept category MKL in our submission of A_FTRD-HLF-4. The idea of grouping concepts into categories was also studied in [13] from a different point of view. In this mode, we group TrecVid 2008/2009 concepts into some categories as shown in TABLE V. For each category, MKL was carried out to learn the kernel weights which are shared across the concepts in that category. Although concepts in one category share kernel combination weights, support vectors and corresponding weights were learned separately for each concept. In each category, positive training samples are the union of positive samples for each concept in that category; and, for negative training samples, intersection of negative samples for each concept is used. For other concepts not in these categories, its weighting scheme is an average of the kernel weights over the 5 categories.

Comparing the detection performances of A_FTRD-HLF-4 and A_FTRD-HLF-6 in TABLE I, we see that such concept category MKL could derive more robust kernel weights than carrying our MKL directly over each concept. However, for our submissions, the MKL approaches did not achieve better performance than kernel combination with equal weights. This might be related to limited positive samples for some concept and large intra-concept variability. All these factors affect the generalization capability of MKL.

4. CONCLUSION

This was our first time participating in TRECVID. Our emphasis is on the evaluation of various low-level features and robust fusion strategies. Experimental results show that both early and late fusion of sift local descriptors, color, edge and texture features could integrate different aspects of knowledge about concepts and achieve better detection performance than any single low-level feature. Our best run ranked the 10th among all 203 runs, and for all concepts, our performances are better than the median performance. When comparing with the best performances on these concepts, we find that there is still large improvement space for our systems.

In future, we plan expand our set of low-level features, e.g. color sift [10], and incorporate motion information to detect concepts involving activities more effectively.

5. REFERENCES

[1] "Guidelines for the TRECVID 2009 Evaluation," http://www-nlpir.nist.gov/projects/tv2009/tv2009.html.

[2] G. Quenot and S. Ayache, "TRECVID 2009 Collaborative annotation," http://mrim.imag.fr/tvca/.

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

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

[5] J. Philbin, M. Marin-Jimenez, S. Srinivasan, and A. Zisserman, "Oxford/IIIT TRECVID 2008 – Notebook paper," http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/oxford.pdf.

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

[7] Yingyu Liang, Xiaobing Liu, Zhikun Wang, et al. "THU-ICRC at TRECVID 2008," http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/thu-icrc.pdf.

[8] J. Zhang, M. Marszalek, S. Lazebink and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, 73 (2): 213-238, 2007.

[9] R. Collobert, S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *JMLR*, 1: 143-160, 2001.

[10] C. G. M. Snoek, K. van de Sande, et al."The MediaMill TRECVID 2008 semantic video search engine," http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/mediamill.pdf .

[11] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proc. ICML*, 2007.

[12] C. Longworth, M. J. F. Gales, "Combining derivative and parametric kernels for speaker verification," *IEEE TASLP*, 2009.

[13] Y. Peng, Z. Yang, J. Yi, L. Cao, H. Li, J. Yao, "Peking University at TRECVID 2008: High Level Feature Extraction," http://www-nlpir.nist.gov/projects/tvpubs/tv8.papers/peking-university.pdf.