TRECVID 2007 High-Level Feature Extraction By MCG-ICT-CAS*

Sheng Tang, Yong-Dong Zhang, Jin-Tao Li, Ming Li
Na Cai, Xu Zhang, Kun Tao, Li Tan, Shao-Xi Xu, Yuan-Yuan Ran
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{ts,zhyd,jtli,mli,ncai, zhangxu,ktao, tanli, xushaoxi, ranyuanyuan}@ict.ac.cn

ABSTRACT

We participated in the high-level feature extraction task in TRECVID 2007. This paper describes the details of our system for the task. For feature extraction, we propose an EMD-based bag-of-feature method to exploit visual/spatial information, and utilize WordNet to expand semantic meanings of text to boost up the generalization of detectors. We also explore audio features and extract the motion cues in compressed domain for detecting concepts highly associated with audio/motion. We use Ordered Weighted Average (OWA) fusion method to combine the SVM-based multi-modal concept detection results. Experiment results show that our methods are effective.

1. Overview

In high level feature extraction task, we divide development set randomly into 3 subsets based on frames in the proportion of 5:3:2. 50% is the training set for training classifiers, and 30% is the fusion set for training fusion method, and the rest 20% is the validation set for evaluating our method.

We focus on a set of novel features for SVM-based concept annotation: (1) Bag-of-features: We propose an EMD-based bag-of-feature method to exploit visual/spatial information; (2) Textual features: We utilize WordNet to expand semantic meanings of text to boost up the generalization of detectors; (3) Audio features: We utilize audio features based on short-time frames such as short-time energy, zcr, sub-band energy, sub-band energy ratio and MFCC to detect concepts highly associated with audio information; (4) Motion features: For detecting events or scene highly associated with motion, we extract the motion cues by exploiting a fast motion feature extraction method. This method is based on motion segmentation according to different camera motion type in compressed domain. Since the evidence from different sources tends to be complementary, fusion of them is of key importance. We tried several heuristic fusion methods. Experiments show that OWA outperforms the other fusion methods such as Max, Min, Average, Product, Weighted Average and Adaboost.

2. Basic Visual and Audio Features

Basic Visual Features

Our system extracts six basic visual features [1] for each key frame of the video shots. The basic visual features are: (1) Color Histogram (CH) ; (2) Color Correlogram (CC) ; (3) Color Moments (CM) ; (4) Co-occurrence Texture (CT); (5) Wavelet Texture Grid (WTG); (6) Edge Histogram (EH).

Basic Audio Features

We divide all the audio shot clips into audio short-time frames. All the audio features are extracted based on audio frames. For each audio frame, we calculate the following audio features: (1) Short-Time Energy; (2) Shot-Time Average Zero-Crossing Rate; (3) Sub-Brand Short-Time Energy; (4) Sub-Brand Short-Time Energy Ratio; (5) MFCC.

*This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).

3. SIFT-based Features

We use two SIFT-based approaches: basic visual keyword and EMD-guided bag-of-feature (improved visual keyword). Visual keywords is a traditional method that has been adopted in several papers and proved to be effective for a lot of application, such as object recognition, image categorization, copy detection and so on. Furthermore, we propose to introduce spatial information to bag of features by using earth mover's distance (EMD).

Basic Visual Keyword

We build a visual vocabulary of SIFT points detected from keyframes based on [1]. We choose approximately 1,000 keyframes from TRECVID-2007 development set, which contain only positive samples over all the 36 high-level concepts in LSCOM-lite lexicon. However, these 1,000 keyframes have already included all kinds of visual information of negative samples because of the intra class diversity in TRECVID-2007 development set. With K-means, about 650,000 SIFT points are quantized into 828 clusters, and each cluster represents a visual keyword. As depicted in [2], we also use tf-idf that is classical in text retrieval to form the visual keywords vector for each keyframe.

EMD-guided Bag-of-Feature

Earth Mover's Distance (EMD) provides an effective way for us to measure how much effort would be required to transform one bag into the other when comparing bags of local features. However, the complexity of EMD is so expensive that we adopt embedded EMD which provides a way to map the weighted point sets from the metric space into the normed space with low distortion.

Let B be a bag-of-feature and $f(B)$ be the embedded EMD vector for B. G_l is the embedded grids, each with side lengths 2^l, $l = -1, ..., \log(D)$, where D is the diameter of the feature space. We can get a sparse vector:

$$f(B) = \left[\frac{1}{2} G_{-1}(B), ..., 2^l G_l(B), ..., D G_{\log(D)}(B) \right].$$

It is proved [3] that in the embedded space, the normed distance between $f(B_1)$ and $f(B_2)$ is an estimation of the exact EMD distance for bags-of-feature B_1 and B_2, where the EMD embedding has an upper bound of distortion of $O(\log(D))$.

In our approach, we use multi-resolution grid representation as illustrated in Figure 1 to implement embedded EMD vector, and in each little grid, the original bag-of-feature are adopted to describe the feature of this grid. The main advantages of the spatial pyramid representation are: first of all, because of combination of multiple resolutions, it is robust to failures at individual levels; furthermore, there has been proved that the promotion has become very subtle when L is increased from 2 to 3; Additionally, experimental results show that the enlarged vocabulary doesn't bring about dramatic boost in performance when the size of vocabulary reaches some level.

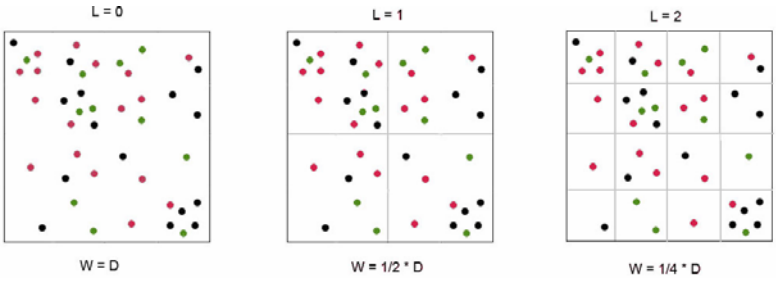

Figure 1: EMD-guided Bag-of-Feature, with M = 3, L = 2.

Grounding on all these vantage points, for controlling the temporal and spatial complexity, we ultimately employ $M = 180$ as the size of the visual vocabulary and $L = 2$, i.e. $\frac{1}{4}D$, $\frac{1}{2}D$, D, as the highest level of multi-resolution representation.

4. Textual features

Text information extracted from videos serves as the most explicit feature to bridge up the semantic gap. So it is necessary and important to exploit text information of videos and incorporate it into concept detectors. Most participants in previous exploited various ways to incorporate text information into detectors and all gained some improvement of baseline. However, they all set up detectors only based on text information directly extracted from videos, which may constrain the generalization ability of detectors in case the training data has little or no correlation with testing data. To boost up the generalization of detectors, we propose a novel approach which utilizes WordNet to add more semantic meanings to detectors as show in Figure 2.

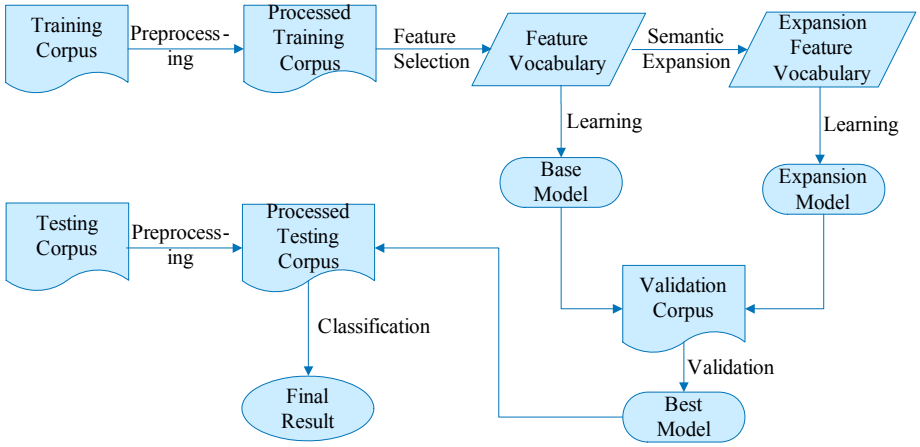

Figure 2: Framework of our text semantic expansion method

5. Motion features

The motion information, including global camera motion and moving objects in video scenes, is very valuable for detecting semantic events such as Walking, Violence or People-Marching. For application on large-scale video corpus, a compressed domain method is used to limit the time consumption. As most motion analysis methods try to estimate the affine model or projective model of camera motion, our approach tries to parses the camera motion in a more direct way. In [4], an effective method is proposed to extract certain kinds of camera motion parameters. Our approach is similar in form except that motion vectors are used as input in stead of optical flow.

Before estimating the global camera motion, some preprocessing steps should be taken, in which the important step is that potential noise MBs (Macro Blocks) with abnormal motion vectors should be eliminated Using DC information components extracted by the method of [5]. Both the marked noise MBs and those intra-coded MBs will be treated distinguishingly at following steps. After preprocessing, different types of camera motion are judged in turn. Some different rules are used to judge the modes of Still, Pan/Tilt, Zoom, Rotation and Irregular. For Pan/Tile, Zoom and Roation frames, corresponding parameters are counted.

After getting the global camera motion type and motion parameters, we can make the camera motion compensation. In every pixel, we estimate the motion vector caused by camera motion by the above estimated parameters. After subtracting the camera motion vector from the MB's motion vector, we can obtain the relative motion of the corresponding area (moving object/foreground) to background. We do not compensate noise MBs and intra-coded MBs. If more than half of their 8 neighbors are normal MBs, they obey the majority of normal neighbors' label. Otherwise they will be labeled as background.

We calculate some invariant moments [6] from the mask image of segmentation, and formed them into a feature vector $\vec{V}$ characterizing the information of moving objects. A sliding window of 2 seconds is used in our approach as the detection unit. The feature sequence of a window is $\{\vec{V}_1, \vec{V}_2, \cdots, \vec{V}_n\}$, whose average is $\hat{\vec{V}}$. The absolute values of neighbor difference: $\{\Delta\vec{V}_1, \Delta\vec{V}_2, \cdots, \Delta\vec{V}_{n-1}\} = \{|\vec{V}_1 - \vec{V}_2|, |\vec{V}_2 - \vec{V}_3|, \cdots, |\vec{V}_{n-1} - \vec{V}_n|\}$ are also counted, whose average is $\Delta\hat{\vec{V}}$. Other four statistics about motion type distribution will be also counted in the window. Above two averaged vectors and four statistics form the final motion features. The flowchart of our approach is shown in Figure 3.

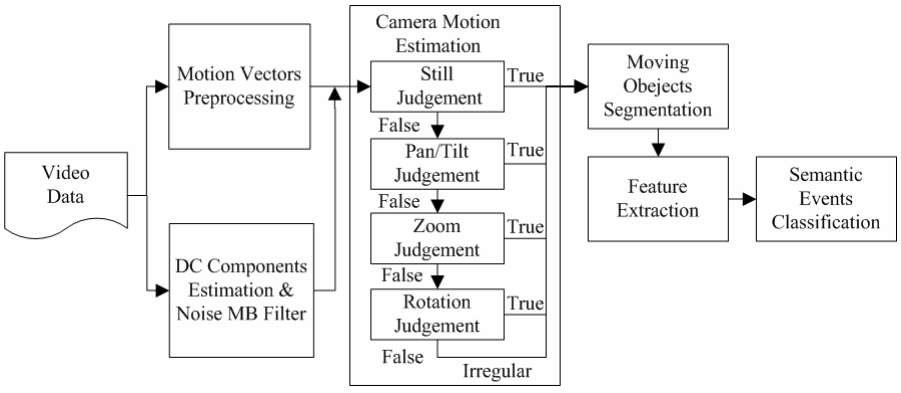

Figure 3: Flowchart of our motion-based approach

6. Fusion

We tried number of fusion method including non-heuristic method and heuristic method. Non-heuristic methods including Max, Min, Average, Product, do not need fusion training. They are simple but not efficient enough. Heuristic methods including Adaboost, Weighted Average and Ordered Weighted Average (OWA) [7] perform better results.

Tested in the develop set, OWA method performs the best for most high level features especially. Compared to Max, Min Average, Product, Weighted Average and Adaboost fusion method, OWA performs the best in 22 concepts of all 36 concepts comparing to 6 other fusion methods. On how many concepts each fusion method performs the best are shown in Figure 4.

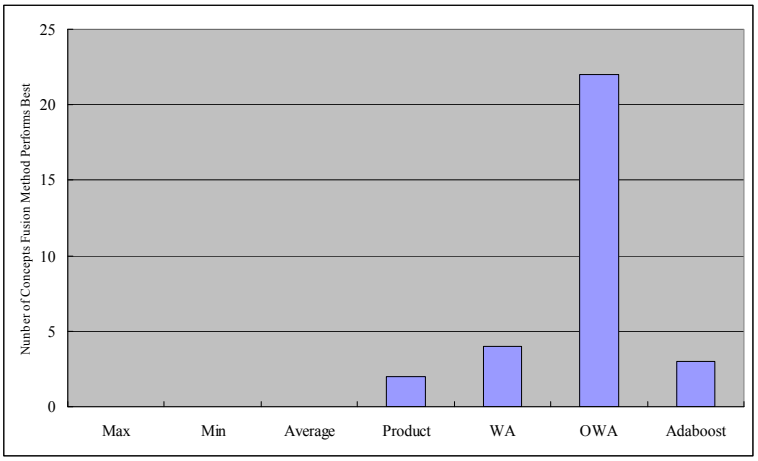

Figure 4: Performance comparison of fusion method

7. Experiments and Result Analysis

We submitted a total of 5 runs. The description and MAP of each run are as shown in the following Table 1. Our best run ranked the 15th among all the submitted 150 runs.

Table 1 Description and MAP of our HLF runs

HLF run	MAP	Description
A_ICT_1	0.090	Best Run
A_ICT_2	0.085	Visual Baseline+SIFT
A_ICT_3	0.071	Visual Baseline+SIFT+Text
A_ICT_4	0.073	Visual Baseline+SIFT+Text+Face+Motion+Audio
A_ICT_5	0.061	Visual Baseline Run (only six basic visual features)

To test our SIFT features, we fuse (by OWA method) the results of visual keywords and EMD-guided bag-of-feature into the baseline result as A_ICT_2. The Average Precision Performances of the A_ICT_2 and the visual baseline A_ICT_5 are shown in Figure 5 and 6 respectively, which shows that the inferred average precision of 20 concepts benefit a lot from SIFT features. Except for 4 concepts – office, meeting, police_security, military – each concept has some boost in a different degree, especially for concepts such as desert, waterscape_waterfront, boat_ship, people-marching, explosion_fire, maps, charts (the promotions are dramatic). The MAP (mean average precision) of this run has also reached 0.085 compared to 0.061 of the baseline run. It indicates our SIFT features are effective.

By comparing the A_ICT_2 with A_ICT_3, we noticed that addition of textual features decreases the MAP value. This is reasonable. One reason is that the ASR and MT about this year's data contain quite some noisy information. Another reason is that the problem of how to fuse textual features with other features still needs further research although our fusion methods already get some satisfying results.

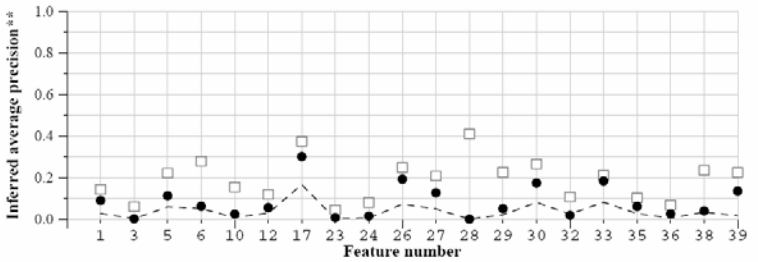

Figure 5: Average Precision Performance of HLF A_ICT_2 (Visual Baseline +SIFT)

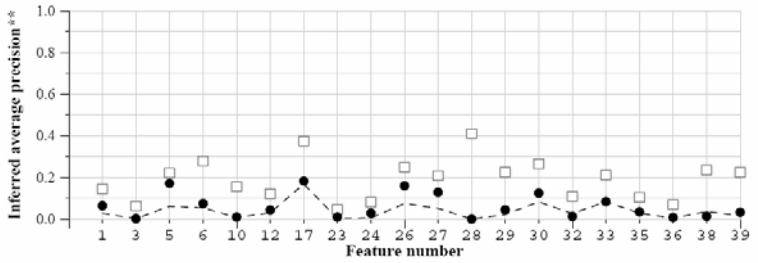

Figure 6: Average Precision Performance of HLF A_ICT_ 5 (Visual Baseline)

By comparing the A_ICT_4 and A_ICT_3, we can find that addition of more features such as motion and audio features increases the MAP slightly.

By comparing the A_ICT_1, A_ICT_4 with A_ICT_2, A_ICT_3 and A_ICT_5, we can find that motion features are good at detecting concepts about events or behavior such as Walking_Running and People-Marching, and can boost the detection precision for other few concepts such as Sports by fusion with other features. For Sports, the two runs (A_ICT_1, A_ICT_4) with motion information are obviously better than the other runs. For People-Marching, although the two runs are not as good as the best of our 5 runs, they are near to the best.

References

[1] Arnon Amir, Janne Argillander, Murray Campbell, etal, “IBM Research TRECVID-2005 Video Retrieval System”, NIST TRECVID-2005 Workshop, Gaithersburg, MD, November 2005.
[2] Josef Sivic and Andrew Zisserman, “Video Google: a text retrieval approach to object matching in videos”; In Proc. ICCV, Oct 2003.
[3] Kristen Grauman, Trevor Darrel, “Efficient image matching with distribution of local invariant features”, In Proc. Computer Vision and Pattern Recognition (CVPR), 2005.
[4] Xingquan Zhu, Xiangyang Xue, Hangzai Luo, and Lide Wu, “A Qualitive Camera Motion Classification Based on Motion Vector”, Journal of Computer Research and Development (in Chinese), Vol. 38, No. 1, Jan. 2001.
[5] B. L. Yeo and B. Liu, “On the extraction of DC sequences from MPEG. compressed video”, in Proc. Int. Conf. Image Processing, vol. II, pp. 260–263, 1995
[6] M. Sonka, V. Hlavac, and R. Boyle, “Image Processing, Analysis, and Machine Vision (2nd Edition)”, pp.259-262, PWS Pub., New York, 1999
[7] Yager R R, Kacprzyk J. “The Ordered Weighted Averaging Operators:Theory and Applications”[M]. Norwell MA:Kluwer,1997.