Intelligent Multimedia Group of Tsinghua University at TRECVID 2006

Jie Cao, Yanxiang Lan, Jianmin Li, Qiang Li, Xirong Li, Fuzong Lin, Xiaobing Liu, Linjie Luo, Wanli Peng,
Dong Wang, Huiyi Wang, Zhikun Wang, Zhen Xiang, Jinhui Yuan, Bo Zhang, Jun Zhang, Leigang Zhang, Xiao Zhang, Wujie Zheng
State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, P. R. China

Abstract

Our shot boundary detection system of this year is basically the same as that of last year. However, we have made three minor improvements on the system, including the detection of FOIs, flashlight and short gradual transitions. On the data set of last year, the new system achieves better performance than the old one. However, on the data set of 2006, the new system has not performed better as expected. We find that this is mainly due to the inconsistent annotation criteria, a) the inaccurate definition of FOIs, OTHs etc, b) the blurry distinction between CUTs and short gradual transitions, c) the inconsistent annotation of video in video.

In high level feature extraction task, rich and hierarchical / multiple granular visual representations are adopted. A bundle of diversified SVM classifiers are trained sequentially for each feature. These classifiers are then combined with a weight and select fusion algorithm. Also, the RankBoost and the StackSVM fusion algorithms are implemented, and different approaches for representing concept context are evaluated in quest of the performance gain. Our submitted runs (runid: A/B_hua) are ranked highest in MAP of all HFE participants except run B_hua_2 which is ranked 8th. At the same time, a top result for 7 out of 20 concepts is obtained. The results indicate that our weight and select fusion algorithm works surprisingly well, better than all variations of the RankBoost and the StackSVM fusion algorithm.

HFE Runs	MAP	Description
B_hua_1	0.189	RankBoost which takes baseline shot score plus Mediamill 404 dim features as weak rankings
B_hua_2	0.175	RankBoost which takes 110 dim baseline keyframe score as weak rankings, 50 top rankings selected
B_hua_3	0.199	Roundrobin which combines all five other runs on the shot rank basis[1]
B_hua_4	0.179	RankBoost which takes 110 dim baseline keyframe score as weak rankings, 200 top rankings selected
B_hua_5	0.185	Automatic or manual rule for setting concept context with 39 dim concept vector for each shot
A_hua_6	0.192	Baseline, select and weight top 50 SVM classifiers out of 110 trained on 22 features respectively

We submitted results for both automatic search and interactive search. In both systems, we use text model, image model and concept model. And in interactive mode, result expansion is adopted to find the relevant shots which are the neighbor of those found in timeline or low level feature space. It is found that in interactive search, the average results given by novices are comparable to results of experts. Therefore, it seems that the comprehension of the users to the topics is more important than the familiarity to the search system.

HFE Runs	MAP	Description
F_B_1_HU01_1	0.0365	uses only Text (ASR) Model
F_B_2_HU02_2	0.0526	uses Text (ASR) Model, Image Model and Concept Model. The models are combined linearly with query-class-dependent weights trained on data of TRECVID2005.
F_B_2_HU03_3	0.0528	uses Text (ASR) and Concept Model. The models are models are combined linearly with equal weights
I_B_2_HU04_4	0.158	based on text search, story browsing and text+concept+LLF (low level feature) feedback, done by experts

[1] After a bug-fix, the roundrobin run turns out to be the best run among both our submitted runs and all submitted runs.

I_B_2_HU05_5	0.167	based on 4, adding components of "concept query input" and "similar image browsing", done by experts
I_B_2_HU06_6	0.136	based on text search, story browsing and text+concept+LLF (low level feature) feedback, done by novices

It is the 3rd time for IMG (Intelligent Multimedia Group, Department of Computer Science and Technology, Tsinghua University) to take part in TRECVID. We participated all the four tasks this year. Our approaches for shot boundary detection, high level feature extraction, search and BBC rushes are presented in Section 1, 2, 3 and 4 respectively.

1. Shot Boundary Detection

1.1 System overview

The shot boundary detection system in 2006 is basically the same as that in 2005, which is described and evaluated in [Yuan07]. As shown in Figure 1, the shot boundary detection is conducted by hierarchical classification architecture. First of all, an FOI detector is employed to recognize the FOIs. Secondly, feature vectors for CUTs, constructed based on the graph partition model, are used to train a SVMs model or classified as CUTs and non-CUTs with the trained model. After all the FOIs and CUTs being detected, multi-resolution feature vectors are constructed to detect GTs. With the hierarchical classification procedure, all kinds of shot boundaries can be detected. In the system of 2006, we have improved the FOI detector and added several post-processing modules after the CUTs and GTs detector, e.g. flashlight detection and short gradual transitions detection. On the evaluation data of 2005, the new system indeed performs better than the system of 2005. However, on the evaluation data of 2006, the improvement is not so significant (perhaps worse). We will present the implementation details of the improvements and analyze why the system does not performed as expected.

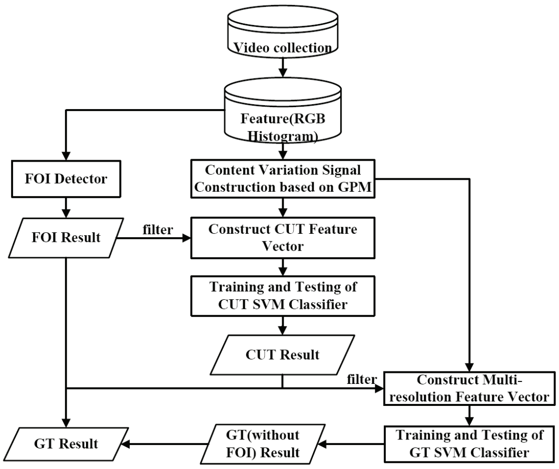

Video collection
Feature(RGB Histogram)
FOI Detector
Content Variation Signal Construction based on GPM
FOI Result
filter
Construct CUT Feature Vector
Training and Testing of CUT SVM Classifier
CUT Result
filter
Construct Multi-resolution Feature Vector
GT Result
GT(without FOI) Result
Training and Testing of GT SVM Classifier

Figure 1. The flowchart of the shot boundary detection system 2005

1.2 FOIs detector

During the FOI, two adjacent shots are spatially and temporally well separated by some monochrome frames [Lienhart01], whereas monochrome frames seldom appear elsewhere. Lienhart proposed to firstly detect all monochrome frames as the candidates of FOIs, and then precisely locate the positions of FOIs [Lienhart99]. The similar idea is improved and extended by Truong [Truong00]. The FOIs detection module of our 2005 system is also based on the similar idea but with different implementation [Yuan05]. As we have stated, the approaches based on this idea consist of two stages: monochrome frames detection and FOIs location. For the purpose of monochrome frame detection, the mean and the standard deviation of pixel intensities are commonly adopted to represent the visual content. Furthermore, the FOIs are located by tracking the characteristics of the mean and standard deviation curve. The previous work has presented different implementations and reported the effectiveness of the idea [Lienhart01, Truong00, Yuan05]. In [Yuan05], we used the mean of the intensities of each frame for monochrome frame detection. However, we find that both the mean and standard deviation of intensities are better alternatives, since for monochrome frame, it not only shows extreme darkness or brightness, but also reveals the uniform spatial distribution, which can be well characterized by the standard deviation of the intensities. Hence, we have replaced the monochrome frame detection module in the FOIs detector. The idea is that only the frames exhibiting both unitary color and uniform spatial distribution are classified as monochrome frames. In the system of 2006, the idea is a rule-based implementation.

1.3 Flashlight and video-in-video detector

Based on the analysis for the false alarms of our 2005 system, we believe that a post-processing module after the CUTs and GTs detectors is necessary. On the one hand, though the adopted graph partition model (GPM) is somewhat robust to various abrupt illuminance changes such as flashlight [Yuan07], some of flashlight scenes are still misclassified as shot boundaries. On the other hand, many shot transitions within video-in-video scenes are classified as shot boundaries by our system though they are not annotated as shot boundaries in the ground truth. Therefore, we adopted the edge change ratio (ECR) [Lienhart99, Zabih95, Kim02, Heng03] method to remove the false alarms, such as flashlights and transitions within video-in-video scenes.

The edge change ratio (ECR) method was originally proposed by Zabih to perform shot boundary detection [Zabih95]. In ECR method, Canny edge detector is firstly employed to calculate the edge map of each frame, i.e., the structural representation of the visual content And then the edge change ratio between adjacent frames is calculated to decide whether shot transitions occur. Lienhart [Lienhart99] points out that ECR usually do not out-perform the simple color histogram methods, but are computationally much more expensive. Despite this depressing conclusion, the edge feature finds their applications in removing the false alarms caused by abrupt illumination change. For example, the ECR value of flashlight scenes will be large enough since the structure of the frame basically remains unchanged under the abrupt illuminance changes, while the ECR value of real CUT boundaries will be small because of the mismatch of the structures. Kim [Kim02] and Heng [Heng03] independently designed flashlight detectors based on the edge feature, in which edge extraction was required only for the candidates of shot boundaries and thus the computational cost was decreased. Besides flashlight scenes, the ECR approach can also remove the false alarms caused by shot transitions within video-in-video scenes since for shot transitions within video-in-video, only small portion of the content varies while the majority remains unchanged.

1.4 About the short gradual transitions

As shown in Figure 1, with the hierarchical classification architecture, CUTs detector is before the GTs detector. However, our CUTs detector is based on statistical machine learning approaches. It has difficulties in distinguishing the CUTs and short gradual transitions (e.g., lasting less than 10 frames). Our 2005 system often classifies the short gradual transitions as CUTs. Although the short gradual transitions no more than 5 frames can be successfully matched with CUTs in the ground truth, many detected short gradual transitions more than 5 frames are not considered right by the evaluation program. We note that NIST has presented an explanation for this case in the guide [nist02]. Therefore, we add the special treatment for the training of short gradual transitions to the system of 2005. After the CUTs and GTs detectors, we use an additional short GTs detector to further select the short GTs from the CUTs results. The training method of short GTs is the same as the training approaches of CUTs and GTs.

1.5 The evaluation results

The shot boundary system of 2005 has achieved the top performance in the evaluation of TRECVID 2005. However, it seems that much room for improvement exists. This year, we made three minor improvements on the system based on its common types of misclassifications. We evaluated the contribution of each improved module and the overall performance of the improved system (i.e., the 2006 system) on the data set of 2005. The experiments show that each improvement does work. The following statistics are the rough results:

a. The improvement of FOIs detector increases the overall F-measure from 0.8987 to 0.9009, which means about 10 misclassifications are corrected.

b. The improvement of Flash detector and video-in-video detector increases the overall F-measure from 0.9009 to 0.9088, which means about 36 misclassifications are corrected.

c. The improvement of special treatment of short GTs increases the overall F-measure from 0.9088 to 0.92, which means about 50 misclassifications are corrected.

The total improvement of F-measure is 0.0213, which means about 96 misclassifications are corrected. Of course, based on an about 90%

hit rate, the improvement of performance is not very significant. We employed the improved system to the data set of 2006. However, the improvement is not as large as that of 2005 data. We examined the detection results and analyzed the causes of misclassification. We found that many errors are due to the inconsistent annotation criteria of the ground truth:

a. The inaccurate definition of FOIs, OTHs etc. In the non-2006 data, the boundary with instant transitions before and after the monochrome frames is considered as two CUTs and all the other transitions with monochrome frames in between are considered FOIs (or OTH). However, all the transitions with monochrome frames in between are considered FOIs.

b. The blurry distinction between CUTs and short gradual transitions. Our system is based on machine learning methods. It can not distinguish CUTs and very short gradual transitions successfully. Therefore, the system labels most of the short gradual transitions as CUTs. Fortunately for the before evaluation, in the non-2006 data, most of the short gradual transitions last less than 6 frames. According to the evaluation rule[2], even a short gradual transition less than 6 frames is labeled as a CUT, it can also matched with the ground truth. Therefore, we have not paid special treatment to the classification of CUTs and short GTs. Unfortunately, in the 2006 data, many short gradual transitions last longer than 6 frames but less than 10 frames. Our system has found such boundaries but labeled them as CUTs. In the result, the performance is not as good as expected.

c. The inconsistent annotation of video-in-video scenes. We observed that in TRECVID 2005 many transitions within video-in-video scenes are not considered. Therefore, we have added the flash and video-in-video detector to remove such false alarms. However, in TRECVID 2006, many transitions within video-in-video are considered. We wonder whether it is determined by the area of the inside video, but it seems not.

2. High Level Feature Extraction

2.1 Overview of concept detection system

Video indexing and retrieval is still in its childhood compared with the matured text indexing technology which has already been successful applied to semantic text retrieval. The underlying gap is the lack of concrete basic indexing unit in video. Current research trend in TRECVID shows strong favor of the generic visual indexing [Amir03, Fan04, Snoek06, Snoek06a] as a cornerstone for video retrieval. Concept detection, in its most general form, serves exactly for this purpose. However, generic visual indexing for multimedia archives is far from satisfying the user need due to its low accuracy and lack of robustness. On the other hand, a vocabulary of about 1000 visual concepts for describing rich semantics is proposed recently [LSCOM], which calls for more powerful tools to reliably detecting much more concepts at the same time. Apart from the machine learning techniques, concept detection pipeline, context factors, authoring metaphor previously adopted in visual indexing, we decide to borrow ideas from the cognitive neuroscience since the human vision system is a real miracle considering its generalization ability, computational efficiency and elegancy for recognizing innumerable objects with ease. It is reported [Thorpe96] that object detection in complex scenes can be achieved in less than 150 ms, which is on the order of the latency of the signal transmission from the retina to inferotemporal cortex (IT), the highest area in the ventral visual stream thought to have a key role in object recognition. What is more, all this can be done in a feed-forward manner without top-down attention involved as observed in the pop-out experiments [Treisman80].

After fifty years' exploration, there are a few mostly accepted properties of the ventral stream architecture for visual processing[3] [Riesenhuber02]. Among them are a basic feed-forward processing of information, a hierarchical build-up of invariance to position and scale and to viewpoint and more complex transformations, together with an increasing complexity of the optimal stimuli for the neurons. Very recently, several attempts [Serra05, Mutch06] have been made to simulate the basic human visual processing principles proposed in [Riesenhuber99] and the results on several benchmarking difficult datasets such as Caltech Object 101 are encouraging. In these attempts, a common 'HMAX' fusion model with successively alternating simple (S) and complex (C) layers (named after the V1 simple and

[2] *Gradual transitions can only match gradual transitions and cuts only cuts, except in the case of very short gradual transitions (5 frames or less), which, whether in the reference or in a submission, will be treated as cuts*. from the guideline of TRECVID 2006: http://www-nlpir.nist.gov/projects/t2002v/sbmeasures.html

[3] Here we overlook the difference among visual object/scene/event recognition since we assume they undergo similar path through the ventral pathway.

complex cells discovered by Hubel and Wiesel) is adopted. In a nutshell, the key new aspect of these approaches is that a task-specific hierarchical and rich representation is provided for each object category. The hierarchical architecture progressively builds about 6000 rich and redundant features being invariance to position and scale and preserving the selectivity to specific stimulus [Serre06]. As a result, a state-of-the-art detection performance is achieved for 101 object categories. We are thus motivated to propose the principle of rich representation for rich semantics, and enable our concept detection system by adding richer representations. However we did not attempt to build a whole new hierarchical neuro-scientifically sound system due to the limited time and resources.

From past experience from TRECVID, we know that no best single feature fits for all concepts, and no best single feature fits for every concept either. In contrast, an ensemble which selectively combines many good features is more robust and more flexible for general visual concept detection. Instead of redesign a hierarchical feature extraction pipeline as done in [Serra05, Mutch06], we lay our hierarchical visual representations (with text and motion based representations altogether), with the feature extractors and a bundle of diversified classifiers for each feature respectively, in parallel, which form the ensemble architecture as shown in Figure 2. The feature extractors (as shown in Figure 2 with different kinds of arrow lines) functions similar to the simple cells, and each bundle of the diversified classifiers built for each feature (as shown in Figure 2 with the same kind of arrow line as the corresponding feature) resembles the complex cells. The bundled classifiers are generated sequentially within the RankBoost algorithm. In the following stage, we weight each classifier output respectively and then get a maximum value within a fixed number of classifiers by selecting the top n high scored ones where n is a parameter. We also implement the RankBoost and the StackSVM fusion algorithms for comparison. In addition a concept context level is cascaded to utilize the inter-concept relationship for fusion. The overall concept detection system architecture is illustrated in Figure 2.

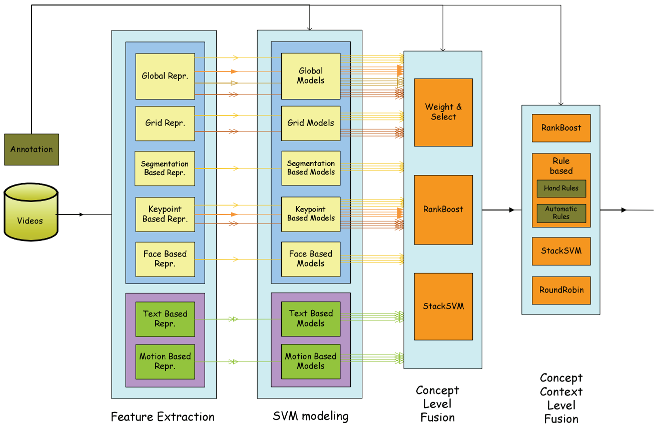


Figure 2. Concept detection system

2.2 Data preparation and annotation

The TRECVID 2005 data is partitioned into three parts, the first is the training set ranging from video #141-277, the second is the fusion set ranging from video #1-100 and the third is the validation set of video #101-140. Our basic fusion model Weight & Select does not require annotation, but the other models and the validation process all require annotation. We selectively annotate ~20 concepts on the video #1-140 by running our baseline fusion and looking down the returned rank list. This strategy is surely sub-optimal but we can not afford the time and human resources for extensive annotation otherwise.

2.3 The representations

We, at our best effort, build five kinds of visual representations. Each representation lies somewhere in the granular hierarchy and each contains several kinds of features which are extracted from this representation. Two more representations based on text and motion analysis are also incorporated. The visual representations are all on the keyframe level and the other two are on the shot level. The total features are about 3 times larger in number than the representations. However, we could not design a total of 39x task specific features for each concept though we try some of them.

The **Global Representation**, which is the coarsest one, contains nine types of global features which are extracted for each keyframe.

These features include two kinds of Color Auto-Correlograms (64 dimensions and 166 dimensions respectively), Co-occurrence Texture (48 dimensions), Color Coherence Vector (72 dimensions), Color Histogram (HSV space, 36 dimensions), Color Moment (LUV space, 9 dimensions), Edge Histogram (72 dimensions) and Wavelet Texture (20 dimensions).

Being of the middle granularity, the **Grid Representation** contains three kinds of layout features extracted from a grid partition of the keyframe. The two types of Color Moment (9 dimensions) and Haar Wavelet Moment (10 dimensions) share a 4x3 grid layout partition. The Edge Histogram (64 dimensions) is extracted from a 5-region layout consisting of four corner regions overlapped by a center region.

The **Segmentation based Representation**, which is also of the middle granularity, is obtained via a standard JSEG segmentation procedure. Each keyframe is down-sampled to 192 pixels in width to speed up the segmentation process. After a keyframe is segmented, the cluttered regions are merged to at most ten regions. Only one type of feature, Color Moment (LUV space, 9 dimensions) namely, is extracted for each merged region since the region is usually homogenous both in texture and color due to the nature of the JSEG method.

The **Keypoint Based Representation**, which is the finest representation we have for now, is generated by the bags-of-keypoint model with both the SIFT [Lowe04] and SURF [Bay06] keypoint descriptors. Obtained by the keypoint detector firstly, the keypoints are then quantized with a sampling based K-mean procedure either on a per concept basis or on a whole training corpus basis. We call these codebooks as concept-dependent one and concept-independent one respectively. Then quantization of the keyframes with these codebooks produces six kinds of histogram features. Two of them are extracted from either a 3x2 or a 4x3 grid layout partition of the keyframe while four others are extracted from the whole keyframe. The 4x3 grid layout uses a codebook with 20 entries while the 3x2 grid uses one with 50 entries in order to balance between description accuracy and robustness. Both of the codebooks are concept-independent. We also generate the other two concept-independent codebooks, one with 10 entries and the other with 100 entries. After extensive experiments on the concept-dependent codebooks, we choose a 10-entry and a 100-entry codebook for each concept. The 10-entry codebooks are combined across all 39 concepts to get a 390-entry concept-independent codebook. The 390-entry codebook is used for feature extraction on its own and combined with the 200-entry concept-independent codebook to get a 590-entry one. The codebook combination mentioned is simple feature concatenation (feature concatenation) instead of entry mix followed by renewed features (center concatenation). All these codebooks are applied to each concept and the resulting histogram feature is sent to the SVM classifier.

The **Face based Representation**, which is a specialized one, is produced with a multi-view face detector [Huang05]. The detected faces, together with a bounding box for the corresponding bodies are taken as a human-oriented segmentation to divide the frame into two parts of human body and background. Then several visual features are extracted to capture the invariance of different kinds of roles, e.g. government leaders or military personals. However, this representation performs not very well and is removed from the feature pool.

The **Text Based Representation** is derived from the unaligned shot text. Simple TF-IDF features are extracted but the results are disappointing compared with the visual ones.

The **Motion Based Representation** is derived from our Low Level Feature extraction algorithm last year [Yuan05] and some motion activity measurement from [Peker01]. For each shot, we connect camera motion feature and motion activity feature to form a 61-dimension feature vector.

Both the motion and text representations are very weak and can not provide comparable result for any concept.

2.4 The modeling approach

As already proved by the past TRECVID activities, SVM classifier is appreciated for its generalization ability and always produces leading benchmark results. We follow this respectable tradition and smooth the classifier output with the standard Platt's sigmoid regression [Platt00] to achieve a posterior probability estimate. Usually the Radial Basis Function (RBF) kernel is adopted in SVM training. For the Grid Representation and Segmentation based Representation, an EMD kernel [Jing04] is more suitable to account for the set matching nature of the inter-image similarity though it is much more computational intensive. We also used a RBF kernel for the Grid Representation since this kernel works well in the past [Chang05]. For the Keypoint based Representation, the χ^2 kernel is adopted through extensive experimental comparison. However, the parameters of σ, r, C, which stands for the RBF bandwidth, relative significance of positive to negative examples (necessitated by the imbalance in the number of positive vs. negative training examples) and the trade-off between training error and margin, often require a computational intensive grid search procedure [Amir05].

Often the concept detection task suffers from both few positive instances and large imbalanced data sets in which negative instances heavily outnumber the positive instances. These two problems are intrinsic for the concept detection task. We have proposed a modified RankBoost approach, RelayBoost (abbr. RL.Boost) [Yuan05, Wang06], to tackle these two problems. Originally the RankBoost [Freund03] algorithm is designed to produce a single ranking list of the given set of objects in the corpora by combining the weak ranking lists. The algorithm is also provided with weighted constrains about which pairs of objects should be ranked above or below one another. It attempts to find a combined ranking that misorders as few pairs as possible, relative to their weighted pair ordering constrains. RankBoost has an efficient procedure for the bi-partite setting as what we have now. Furthermore, it can combine the weak rankings with principled combination weights which count for the possible correlations among the weak rankings. In TRECVID 2006, RankBoost is used mainly for two purposes: the normal one is to fuse the existing weak ranking lists as shown below while the other is to generate the weak ranking lists and to produce a fused ranking simultaneously as in the current setting. The modified RankBoost algorithm is called RL.Boost in the latter situation since it chooses the features for training in a style similar to the choice of runners in a relay race.

Using SVM as the base classifier, our RL.Boost algorithm actively selects balanced positive and negative examples from the imbalanced data set and trains a classifier with these examples in each round, and combines these classifiers in a boosting-like ensemble. RL.Boost also shares the accumulated training error pattern among multiple features for better performance. Please see [Wang06] for details on this algorithm. This model is capable of integrating different features into one module. It has the advantage of a principled automatic weight assignment for each component classifiers in a linear weighted fusion framework and it neither requires time consuming weight tuning procedure as in [Amir05] nor worries about the over-fitting in the fusion process. Furthermore, the r parameter is not required since the sampled examples are balanced and the computational cost is thus reduced. However, the disadvantage is also obvious. The tightly coupled feature ensemble does not allow for inserting new features. In the after-site experiments last year, we also try another variant which selects one feature in each round by comparing the trained models of all features. In spite of its higher detection performance, the training procedure is very slow. This year, we restrict the range of each bundle of classifiers within the same feature and train 5 diversified SVM classifiers for each feature. This is identical to the RankBoost algorithm so we change the name back from RL.Boost. This parameter of 5 is arbitrarily chosen within our computational limit. We find that the prediction performance does not drop much, which may be an indication of the weak inter-feature correlation. By doing so, we further increase the richness of the representation while limiting the classifiers in one single feature and reducing the training time down N times where N is the number of all features.

After the training, we have gathered a 110 dim model vector as a new feature vector for each keyframe with each concept. This vector is derived from the 22 features used in 7 representations with 5 model score per feature. This model vector serves as the input feature for subsequent concept level and concept context level fusion.

2.5 The concept level fusion

When more rich representation is taken, the fusion problem, the binding problem in the neuroscience area, becomes more important than ever before. A good fusion algorithm should remain as simple as possible while keeping robust to component detection errors. A linear summation form does not meet this requirement since it fails to discriminate between the simultaneous presentation of multiple weak responses and a strong response form a reliable component detector. Again we borrow idea from the neural inhibition rule which states that strong output unit suppresses their less-active neighbors. For each of our diversified models for each concept, a weight has been automatically assigned by the RankBoost algorithm to measure its detection power. We introduced a parameter h, to select the most powerful top h models and inhibit others left. All the selected models are then summed with their scores and the corresponding weight. This simple Weight and Select (WAS) method is able to select the strong models with a non-linear mechanism while removing the noisy ones. h is arbitrarily set to 50. Our experiments shows that this algorithm, despite of its simple form, works surprisingly well.

Three kinds of approaches are adopted this year. The first one is the weight and select algorithm; the second and the third are RankBoost and StackSVM respectively.

As stated in Section 2.4, the RankBoost algorithm is also used to produce a single ranking list given the weak rankings. Here we follow the standard bi-partite RankBoost algorithm introduced in [Freund03] with the model vector.

The Stacked generalization is proposed for minimizing the generalization error rate of one or more generalizers. The StackSVM approach is originally designed for concept context fusion and has been successfully applied before [Iyengar03, Snoek04]. We observe that the concept and context level fusion differ only in the features they used. So we take the StackSVM approach in the concept level fusion also. To be more specific, the 110 dimension model vector is taken for each concept as feature input, along with the annotated labels, is used to train an SVM classifier.

2.6 The concept context level fusion

In the concept context module, we try mainly four approaches. The RankBoost and the StackSVM approaches are identical to the ones used in the concept level fusion module, except the feature input. However, we observed a clear over-fit in the StackSVM algorithm from both fusion levels and thus exclude it from the submitted runs. We hypothesize that the feature variation is suppressed by the smoothed SVM output and the limited number of positive and negative examples labeled in the fusion data set misleadingly exaggerates the ease of the learning task. We also try a Rule based concept context with two variants to capture the mutually exclusive relationship between concepts, for example, outdoor and office. We generate mutually exclusive rules based on our common sense (Hand generated Rules) or derive the rules from occurrence statistics from the training data (Automatic generated Rule), and then remove the shots of the concepts which are exclusive from the target concept. There is a clear over-fit in this rule-based concept context method. One possible reason is the different concept distribution between the two years' video data and the other is the fragile performance if the concepts involved. Apart from these approaches, a fourth Round-Robin fusion algorithm which is often adopted in meta-search applications is incorporated. A round-robin algorithm circularly schedules the results from each run into a fused list. We are interested in the pooling strategy used by NIST and this algorithm is designed to quantitatively evaluate the effect.

2.7 Computational issues

This year we use clusters to train our models due to the prohibitive computation otherwise. Some partial estimation of the running time of training sums up to 600 days for one computer! Fortunately the parallel computing paradigm, which is a natural choice for the uncorrelated concept detection task, ends in less than 10 days. Based on this experience, we strongly recommend that the participants report their time consumption in the HFE task for fair comparison.

2.8 Results

Based on these representations and modeling approaches, we have our 6 runs. A_hua_6 is the only A type run we have which is obtained by running the Weight and Select fusion algorithm to select 50 top classifiers from 110 classifiers based on their training performance measured in Average Precision (AP). Other 5 runs are more or less related to the traditional fusion algorithm. Two runs of B_hua_2 and B_hua_4 are all based on the RankBoost algorithm taking all the 110 dim score for keyframe as the weak rankings with different output combination number. However they show a strong over-fit and return the lowest MAP across the 6 runs. A shot level concept context fusion result of B_hua_1 is produced with the baseline shot scores together with the MediaMill 404 dim concept features as the weak rankings. But no improvement is shown. The Roundrobin run of B_hua_3 takes all other 5 runs as the input and it is the best run we have.

Figure 3 compares our performance with the best and the median performance across all runs for the 10 concepts. Figure 4 evaluates our system on the concepts which is evaluated in both TRECVID 2005 (tv05) and TRECVID 2006 (tv06). Figure 5 gives the MAP of all submitted runs in which our runs are indicated by red bars.

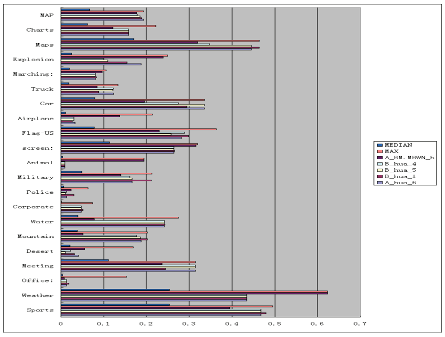


Figure 3. Performance of our system and other systems in tv06

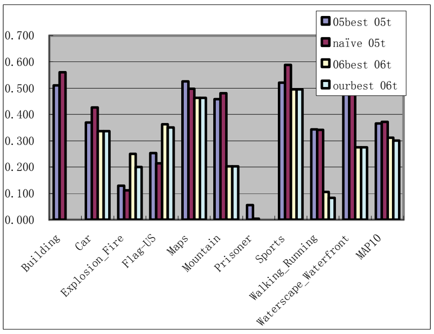


Figure 4. Evaluating our system on both tv05 and tv06

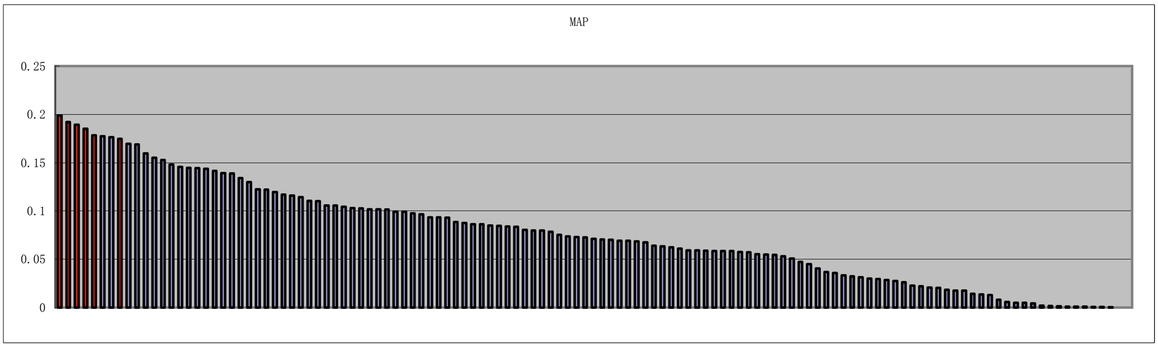


Figure 5. MAP comparison of all submitted runs. Our runs are indicated with red bars.

2.9 Lessons

The following lessons were learnt from across our submitted runs:

a. Rich representation from the parallel layout of hierarchical representations and corresponding features work quite well for detecting semantic concept.

b. The Weight and Select fusion algorithm which is biologically motivated, though still in its primitive form, outperforms the other ones. However, it is possible that the other algorithms may suffer from the improper annotation strategy we take.

c. The RankBoost algorithm builds a diversified bundle of classifiers for each feature and alleviates the burden of the fusion process. In other words, what to fuse is more important than how to fuse.

d. Over-fitting occurs quite often in the fusion process, especially when there is strong mismatch of concept occurrence and broadcasting style between the training and testing data. This has not been mentioned or noticed before, which may due to the relatively short time interval between the training and testing data used in the past.

e. The pooling strategy used does not give the Roundrobin method significantly higher MAP (+3.5%) though it is looked more than other runs by the pooling strategy.

We propose the principle of rich representation for rich semantics, and devise our system architecture for the concept detection task following this principle. Though in its primitive form, the system performs surprisingly well. It benefits from the rich and hierarchical visual representation, a bundle of carefully designed classifiers which are diversified intentionally and a cognitively motivated weight and select fusion algorithm.

3 Search

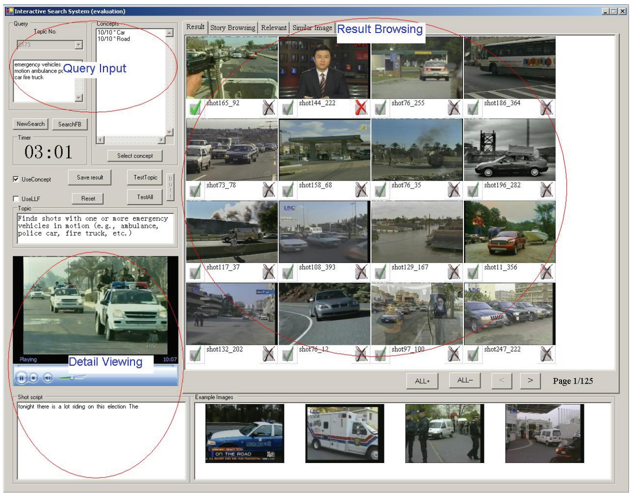
3.1 System overview

We took part in both automatic and interactive retrievals this year. Our system has three basic retrieval models: a text model, an image model and a concept model, all of which support both automatic and interactive retrievals. The text model searches video shots through ASR (automatic speech recognition) script with analysis of query classes. The image model searches video shots through global feature based image classification using SVM. The concept model automatically parses the queries and video shots into concept vectors, and then searches video shots through query-shot similarity measurement using dot product of normalized concept vectors. For interactive retrieval, we also consider the issues of Result Expansion. Time expansion and image expansion is supported with two independent interfaces.

Design of user interface also plays an important role in interactive retrieval. The system is illustrated in Figure 6. The “Query Input” part allows users to enter query text and set concept relevancies. Example images of a query are also shown. The “Result Browsing” section contains not only the browsing interface of search results, but also the browsing interfaces of time and image expansion results named “Story Browsing” and “Similar Image”. In addition, relevant results are collected for browsing. When users browse the results, they can label the shot by clicking the positive flag or negative flag under the key frame. We also enable “Detail Viewing”. Users can watch the video and view the shot scripts with salient words highlighted for details.

3.2 Basic retrieval models

3.2.1 Text model



emergency vehicles motion ambulance po car fire truck

☑ UseConcept

☐ UseLLF

Finds shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)

tonight there is a lot riding on this election The

Figure 6. Search Demo of Tsinghua University in TRECVID 2006

Our text model is based on an OKAPI-TF formula [Paul05, Yan04] as last year. In this year, we developed an automatic query-class-dependent text retrieval system, wherein five query classes are defined empirically. Firstly, the queries are analyzed and classified into five hand-defined classes: “Named Person”, “Object”, “Outdoors Scene”, “Event Scene”, and “General Scene” by query processing. Then the queries are expanded dependent on its category by WordNet [Fellbaum98] and Google Web Search [Google]. The parameters of OKAPI are also trained over the training dataset for each category. The text model also gives a feedback function for interactive retrieval.

The overview of query processing is shown in Figure 7, and we will present the details below.

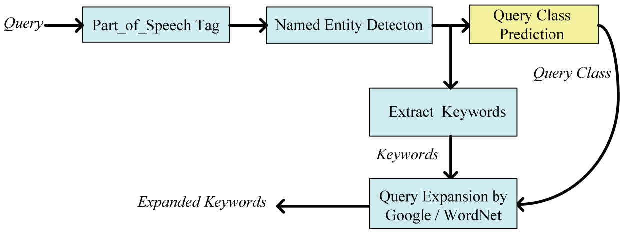


Figure 7. Query preprocess paradigm: class prediction & query expansion

a. Query Class Definition and Prediction

Five classes, i.e. NP, OB, OS, ES, and GS, are defined empirically (see Table 1 for more details). We employ a simple yet effective rule-based algorithm to predict to which class a given query belongs.

Given a new query, we first tag the part of speech properties for the query, and then detect four predefined named entities, i.e., PERSON, OBJECT, OUTDOORS, and EVENT, which indicates NP, OB, OS, and ES, respectively. Queries without named entities are labeled as GS. It is worth noting that we perform a single-label classification for simplicity, although a query might be shared by multiple classes. Moreover, the five classes are predicted with different priorities, that is, if more than one named entities are found in a query, which gives rise to several candidate labels, the one of highest priority will be selected as the final label.

Table 1. Query Class Definition

Class	Description	Priority (Descending)	Example
NP	Named Person. Queries for finding a specific person, possibly with certain actions. Specifically, queries containing person	1	Find shots of Hu Jintao, president of the People's Republic of China

	names are classified as "NP".		
OB	Object. Queries for finding a specific or general object, possibly with certain actions.	2	Find shots of an airplane taking off
OS	Outdoors Scene. Such queries contain keywords which usually characterize outdoor views explicitly or implicitly.	3	Find shots of one or more buildings on fire, with flames and smoke visible.
ES	Event Scene. Queries depicting a scene in which multiple objects, general persons or crowds present with certain actions.	4	Find shots of one or more skiers skiing a slalom course with at least one gate pole visible.
GS	General Scene. Queries cannot be confidently grouped into the above four classes.	5	Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

For PERSON, simple NLP techniques and the capitalization information are combined to find person names. The 39-concept dictionary in the concept detection task is used in OBJECT and OUTDOORS detections. For OUTDOORS, we estimate the conditional probabilities of the specific outdoor concept, given each of the other 38 concepts over the TRECVID 2005 development dataset.

$$P(outdoor \mid concept\_i) = \frac{P(outdoor, concept\_i)}{P(concept\_i)} \approx \frac{Frequency(outdoor, concept\_i)}{Frequency(concept\_i)}$$

This estimation gives us a quantitative insight of the relevance of words to the outdoor scene. Words or phrases in a new query will be tagged as OUTDOORS, if their probabilities exceed 0.5. The top 5 concepts are natural-disaster, boat_ship, mountain, desert, and snow, which is coherent with our intuition. For OBJECT, each concept is manually annotated as OBJECT or not, since there is no OBJECT concept in the dictionary. There are totally 9 concepts labeled as OBJECT, i.e. Airplane, Boat_Ship, Building, Bus, Car, Computer_TV-screen, Flag-US, Maps, and Trucks. The rule for EVENT is rather straightforward, a word or phrase is marked as EVENT if it is or contains gerunds.

b. Keywords Extraction and Query Expansion

Keywords Extraction. Keywords in a query should be extracted before query expansion. The reasons are two-folds: firstly, some meaningless words such as "find shots of" will generally degenerate the search performance and thus should be abandoned, and secondly, considering that more or less noises will be brought in with query expansion, the query needs to refine and preserve the most representative words. We totally collected a 98-topic repository ranging from TRECVID 2002 to TRECVID 2005 [TRECVID] and utilize the popular TF-IDF scheme in the Information Retrieval community to score and rank the salience of query words in the repository. For a new coming query, words in it will be discarded if their salience values are less than a certain threshold which is set heuristically.

Query Expansion. In query expansion, we look to add additional words to the query which are closely related to the extracted keywords. The expansion is fulfilled by submitting the keywords as seeds into some knowledge resources to discover related and meaningful words, dependent on the query class. We implement two query expansion systems, using Google [Google] and WordNet [Fellbaum], respectively. We use WordNet by extracting all the synonyms for each keyword in the query. We use Google by submitting the keywords to Google and then examine the top M returned snippets to discover the top N most frequent words in the snippet set. Standard stop words are removed and the remaining terms are added to the query. In current system, the snippet count M is empirically set to 300 and N is tuned over the development dataset.

Specifically, the Google system is for the NP queries only and WordNet for the other queries. The intuition is that WordNet will not work for NP queries, since their keywords are person names. While for non-NP queries, words collected from Google are usually irrelevant to the queries possibly due to the inaccuracy of the seed keywords. Based on the above observation, we further employ WordNet to refine the expansion results. Only words with fewer amounts of senses are accepted. Some query expansion examples are listed in Table 2.

Table 2. Examples of keywords extraction and expansion

Original Query	Salient Keywords	Query Expansion (WordNet / Google)

Find shots of water with one or more boats or ships	water boat ship	water douse drench irrigate moisten rain soak sprinkle boat barge ship vessel (by WordNet)
Find shots of Condoleeza Rice	Condoleeza Rice	Rice Condoleezza Secretary State Condoleeza New com Nation Security Bush (by Google)

3.2.2 Image Model

For our image retrieval subsystem, we followed the approach from IBM TRECVID 2005's visual retrieval system [Amir05]. We only used Support Vector Machines (SVM) for the Search Task in visual model. The use of SVM mainly faces two challenges in automatic search task: small number of positive examples and no negative examples. In order to overcome these challenges, we extend positive examples from similar topics of TRECVID 2004 and TRECVID 2005 Search Task and create bags of pseudo-negative samples from unlabelled data in order to train different classifiers. In TRECVID 2006 Search Task, several image and video examples are given for every query. Start time and stop time are included for each of the video examples, from which one keyframe is extracted. Image examples, keyframes of video examples and related images from old similar topics constitutes the whole set of image examples.

Firstly, visual feature are extracted for every image in either the example set mentioned above or TRECVID 2006 test collection. In our system, we finally choose localized edge histograms and color moments which had the top performance for our search experiments. Edge Histogram Layout (EHL) is localized edge histograms with 8 edge direction bins and 8 edge magnitude bins, based on a Sobel filter, extracted from a 5-region layout consisting of four corner regions and a center overlapping region (320-dimensional); Color Moments (CMG) is localized color extracted from a 3x3 grid and represented by the first 3 moments for each grid region in Lab color space as a normalized 81-dimensional vector.

Secondly, after feature extraction step, SVM models should be built in order to classify images corresponding to each topic. It is admitted that performance of SVM classifiers significantly depends on kernel type and parameters of the models. In our system, we use radial basis function, which usually performs better than other kernel functions, and select parameters for each visual feature with TRECVID 2005 dataset. On the other hand, the ratio between positive and negative examples and the number of SVM models for each feature are two specific factors in this system as IBM discussed in [Amir05]. In our system, many experiments about these two factors are taken in order to verify related conclusions in [Natsev05]. We randomly generate 5 different learned bags, each includes about 300 pseudo-negative examples, and fuse the SVM scores of each model using AND logic.

Finally, fusion of results from different feature is also an important step in our subsystem. We tested several fusion methods and finally chose average method due to its simplicity and consistent performance.

For interactive retrieval, we use the Color Moment feature only as the SVM costs too much time to be applicable for online retrieval. And we use the exact feedback results for training without any randomly pseudo-negative examples. To further speedup the SVM, we predict only the most relevant results given by text model and concept model. Experiments in TRECVID2005 show its applicability.

3.2.3 Concept Model

Since concept features have been extracted in High Level Feature Detection Task, shots in the dataset can be described by the relevance of them to the concepts. More exactly, we can map the shots to the vectors in the concept space. Each vector denotes the relevance of the corresponding shot to the concepts. Then, the similarity measurement of a shot and a query can be converted to the correlation measurement of their corresponding vectors. Firstly we normalize the concept vectors by subtracting the average concept vector of the dataset. Then we adopt the simple dot product of vectors to finish this work. This method is essentially the weighted-sum of concept scores. The weight of each concept is the relevance of the query to the concept subtracted by the average relevance.

For automatic retrieval, the relevance of the query is calculated by the average relevance of all the query examples. In interactive retrieval, the relevance of the query is calculated by the average relevance of all the positive shots. Note that information of negative shots is not needed for this method.

3.3 Result Expansion

The positive results often cluster together in the time dimension or in the image feature dimension. However, there are also many negative shots around the positive results. So we think it may be helpful to allow users to expand the result optionally in the interactive retrieval.

3.3.1 Time Expansion

The successive shots often have similar contents, so exploring nearby shots of positive results are expected to be helpful. The users are allowed to explore the nearby 15 shots of a specific shot, 7 prior shots and 8 posterior shots. The number of a specific shot and its nearby shots is 16, which is just the number of shots in a display page of our system. We assume the users only do time expansion for the positive results, as it is meaningless to explore the nearby shots of a negative result. To make the users do expansion more simply, we give an independent interface for time expansion. It allocates a page for each time expansion result of a positive shot. Users can browse the time expansion results in sequence by clicking a "Next" button, without switching between the interfaces of search results and time expansion results. Furthermore, it is no use to explore the similar expansion results of two or more neighboring shots. For example, once the time expansion results of the shot "shot3_40" has been explored, there is no need to do time expansion for shot "shot3_38" and "shot3_42".

For more effective time expansion, the exact temporal structure of shots should be utilized. The DVMM group of Columbia enabled exploring full stories and found that a significant number of additional shots could be found this way, especially for named person topics [Chang05]. We will try to use the information of stories in the future.

3.3.2 Image Expansion

The image content is expressive for some topics, for example, soccer, ship boat, snow, etc. So we also enable users to explore the most similar 16 shots of a specific shot. The image similarity of two shots is exactly the similarity of key frames of the two shots. Again we assume the users only do image expansion for the positive results. Similar to time expansion, we give an independent interface for image expansion and allow users to browse the image expansion results in sequence.

The similarity matrix of all shots in the dataset can be computed offline. So it is realistic for online retrieval, which is an advantage compared to the dense computing methods such as SVM classification. However, the performance is not well enough for many topics. Thus, it doesn't play an important role in our system.

3.4 Experiments

3.4.1 Automatic Experiments

We submitted 3 automatic search runs:

F_B_1_HU01_1 uses only Text (ASR) Model

F_B_2_HU02_2 uses Text (ASR) Model, Image Model and Concept Model. The models are combined linearly with query-class-dependent weights trained on data of TRECVID2005.

F_B_2_HU03_3 uses Text (ASR) and Concept Model. The models are models are combined linearly with equal weights

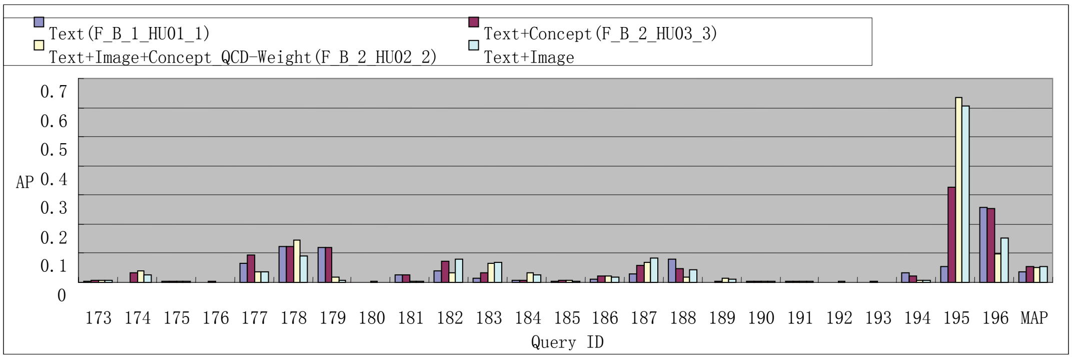


Figure 8. Automatic search results

The evaluation result is illustrated in Figure 8. We can see that although the result of text model (run1: F_B_2_HU01_1) seems not very good, when combined with concept model, the result (run3: F_B_2_HU03_3) improved nearly 44%. But when all of the three models combined together, the final MAP did not increase obviously. In order to compare with this result, we add another run uses Text Model and Image Model which is not submitted. It shows Image Model does well in topics about sports and special objects when combined with Text Model, for instance, "0195: find shots of one or more soccer goalposts" (reaches a MAP of 0.6059 compared with a MAP of 0.054 using only Text Model), "0187: find shots of one or more helicopters in flight", "0183: find shots of water with one or more boats or ships" and so on. However, the performances of nearly half topics stay in the same level or increase indistinctively, and 7 topics show notable decrease in MAP while fusing Text Model and Image Model. When all of the three models fused together, the increase or decrease of some topics in Image Model and Concept Model counteract each other so that the MAP of three models does not increase markedly. A possible reason of this fusion result is the unsuitable selection of fusion weight allocated to each model.

3.4.2 Interactive Experiments

We submitted 3 interactive search runs:

I_B_2_HU06_6 Interactive search results, based on text search, story browsing and text+concept+LLF (low level feature) feedback, done by novices

I_B_2_HU04_4 Interactive search results, based on text search, story browsing and text+concept+LLF (low level feature) feedback, done by experts

I_B_2_HU05_5 Interactive search results, adding components of "concept query input" and "similar image browsing", done by experts

I_B_2_HU06_6 and I_B_2_HU04_4 are done on the same system by novices and experts respectively. The system is similar with our current system described above but no input of concept query and no similar image browsing. I_B_2_HU05_5 are done by experts on our current system. The topics are assigned to different novices and experts randomly while at last the results are gathered together. Note the differences between novices and experts are the familiarity to our system but not the familiarity to the topics.

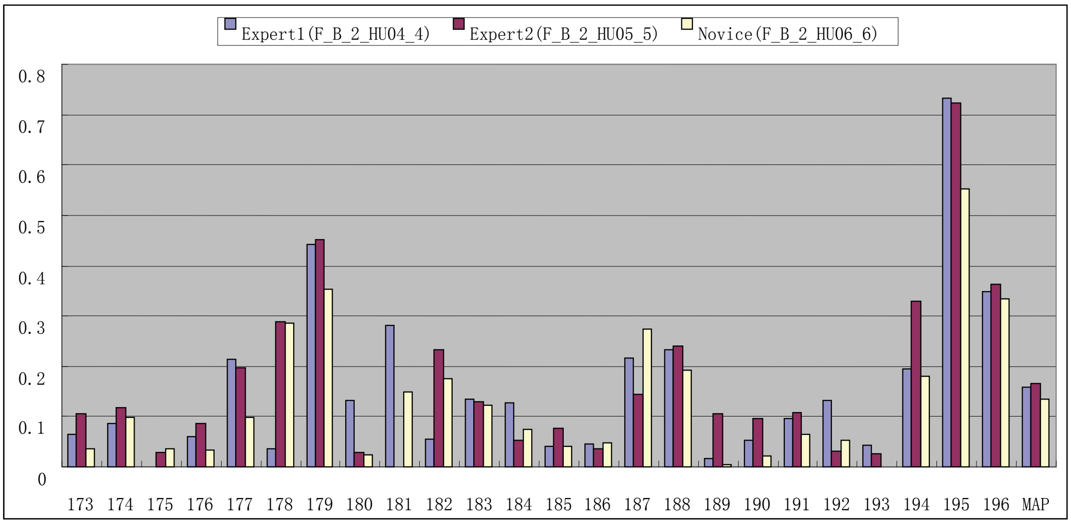


Figure 9. Interactive search results

The evaluation result is illustrated in Figure 9. From the results, we can find that experts do much better for some topics such as 0179 (Saddam) and 0195 (soccer goalposts). The may be because experts are more familiar with time expansion and image expansion, which should be helpful for 0179 (Saddam) and 0195 (soccer goalposts) respectively. However, the average results given by novices are comparable to results of experts, which was unexpected. For some topics such as 0178 (Cheney) or 0182 (soldiers or police), the AP of result done by some expert is quite low. Especially for the topic 0181 (Bush), one of our expert return a result with AP 0.0! That is probably because he/she doesn't know how Bush looks like. And it is more common for other topics to be ambiguous to users. It seems that the comprehension of the users to the topics is even more important than the familiarity to the search system. That is an important issue which should be considered carefully for topic design and experiment design.

4 Exploration of Rushes Exploitation

The data of rushes consist of raw material used to produce videos. Because the speech of the rushes is French, it is only visual analysis and indexing we did. The videos of BBC rushes first are segmented into shots by the shot boundary detection algorithm [Yuan05]. Then image features based on color, texture, and camera motion are extracted from shots. A hierarchical clustering algorithm is applied on the set of features of shots to hide redundancy of as many kinds as possible. The number of clusters is determined by users interactively or by thresholds automatically. A concept detection procedure is performed on these cluster centroids to present the non-redundant materials according to the semantic concepts: interview, fixed camera, person present, urban and others that we implemented in high-level feature extraction task of TRECVID 2006.

In the future, the basic video units should be sub-shot to remove the contents with intermediate motion [Ngo05]. We will add a component of interactive feedback by users for correcting the clustering results. Then a classifier is to be trained for each cluster respectively and to be applied to test data of the rushes. Visual and audio features as many kinds as possible should be extracted and indexed so users can be explore the collection of rushes data by a search engine based on these indexes.

Acknowledgements

This work is partially supported by National Key Basic Research Project of China (2004CB318108). We would also like to thank Intel China Research Center and Tsinghua National Laboratory for Information Science and Technology for their help and support.

References

[Amir03] A. Amir, et al., .IBM research TRECVID-2003 video retrieval system, in Proc. TRECVID Workshop, ser. NIST Special Publication, Gaithersburg, USA, 2003.

[Amir05] A. Amir et al, IBM Research TRECVID-2005 Video Retrieval System, In: Proceedings of TRECVID 2005 Workshop.

[Bay06] H. Bay, T. Tuytelaars, L.V. Gool. SURF Speeded Up Robust Features, in ECCV 2006

[Bouthemy99] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. IEEE Transaction on Circuits and Systems for Video Technology, vol. 9, pp. 1030-1044, 1999.

[Chang05] S. Chang et al, Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction, In: Proceedings of TRECVID 2005 Workshop.

[Chua04] T.-S. Chua, S.-Y. Neo and et al, TRECVID 2004 Search and Feature Extraction Task by NUS PRIS, TREC Video Retrieval Evaluation Online Proceedings, 2004

[Dumais88] Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Using Latent Semantic Analysis to Improve Access to Textual Information, Bell Communications Research, 1988

[Fan04] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref, and L. Wu, ClassView: hierarchical video shot classification, indexing, and accessing,. IEEE Trans. Multimedia, vol. 6, no. 1, pp. 70-86, 2004.

[Fellbaum98] C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press. 1998

[Freund03] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer, An Efficient Boosting Algorithm for Combining Preferences, Journal of Machine Learning Research 4, pp 933-969, 2003.

[Google] Google Web Search. http://www.google.com, 2006

[Heng03] Wei.Jyh. Heng and King. N. Ngan, High accuracy flashlight scene determination for shot boundary detection, Signal Processing: Image Communications, vol. 18, no. 3, pp. 203-219, Mar. 2003.

[Huang05] C. Huang, H. Ai, Y. Li, and S. Lao, Vector Boosting for Rotation Invariant Multi-View Face Detection, The IEEE International Conference on Computer Vision (ICCV-05), pp.446-453, Beijing, China, Oct 17-20, 2005

[Iyengar03] G. Iyengar, H.J. Nock, and C. Neti, Discriminative model fusion for semantic concept detection and annotation in video, ACM MM 2003.

[Jing04] Feng Jing, Mingjing Li, Hong-Jiang Zhang, Bo Zhang: An efficient and effective region-based image retrieval framework, IEEE Transactions on Image Processing, Vol. 13(5) pp 699-709, May 2004

[Kim02] S.-H. Kim and R.-H. Park, Robust video indexing for video sequences with complex brightness variations, in Proceeding of IASTED International Conference Signal and Image Processing, Kauai, Hawaii, USA, Aug. 2002, pp. 410--414.

[Kim04] N. W. Kim, T. Y. Kim, and J. S. Choi. A Probability-Based Flow Analysis Using MV Information in Compressed Domain. in Mexican International Conference on Artificial Intelligence, pp. 592-601, 2004.

[Landauer98] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. In Discourse Processes, 25, 259-284. 1998

[Lowe04] D.G. Lowe Distinctive image features form scale-invariant keypoints, International Journal of Computer Vision Vol. 60(2) pp 91-110, 2004

[Lienhart99] Rainer Lienhart. Comparison of automatic shot boundary detection algorithms. in SPIE Image and Video Processing VII, vol. 3656, Jan. 1999, pp. 290--301.

[Lienhart01] Rainer Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics, vol. 1, no. 3, pp. 469--486, 2001.

[LSCOM] LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.

[Natsev05] Apostol Natsev, Milind R. Naphade, and Jelena Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In ACM Multimedia, Singapore, November 2005.

[Ngo05] Chong-Wah Ngo, Zailiang Pan, Xiaoyong Wei, Xiao Wu, Hung-Khoon Tan, Wanlei Zhao. Motion Driven Approaches to Shot Boundary Detection, Low-Level Feature Extraction and BBC Rushes Characterization at TRECVID 2005. In Proc. of the TRECVID Workshop, Gaithersburg, USA, 2005.

[NIST02] NIST, shot boundary evaluation guide, http://www-nlpir.nist.gov/projects/t2002v/sbmeasures.html

[Mutch06] J. Mutch and D.G. Lowe, Multiclass Object Recognition with Sparse, Localized Features, in CVPR 2006

[Paul05] Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F. Smeaton. TRECVID 2005 – An Overview. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[Peker01] K.A. Peker, Video indexing and summarization using motion activity, Ph. D. Dissertation, New Jersey Institute of Technology, 2001

[Platt00] J. Platt. Probabilities for SV machines. In Advances in Large Margin Classifiers, pages 61{74. MIT Press, 2000.

[Rath99] G. B. Rath and A. Makur. Iterative least squares and compression based estimations for a four-parameter linear global motion model and global motion compensation. IEEE Transactions on Circuits & Systems for Video Technology, vol. 9, pp. 1075-1099, 1999.

[Riesenhuber99] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. Nature Neuroscience, 2(11): 1019-1025, 1999

[Riesenhuber02] M. Riesenhuber and T. Poggio, How Visual Cortex Recognizes Objects: The Tale of the Standard Model.

[Robertson95] S.E. Robertson, S. Walker, and M. Sparck Jones, et al.. Okapi at TREC-3. Proc. Second Text Retrieval Conf. (TREC-3), 1995.

[Sequeira93] M. M. de Sequeira, and F. Pereira. Global motion compensation and motion vector smoothing in an extended H.261 recommendation. in Video Communications and PACS for Medical Applications, Proc. SPIE, pp. 226-237, 1993.

[Serra05] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In CVPR, 2005.

[Serra06] T. Serre, Learning a Dictionary of Shape-Components in Visual Cortex, Ph.D. thesis, MIT, 2006.

[Snoek04] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, F.J. Seinstra, The MediaMill TRECVID 2004 Semantic Video Search Engine, TREC Video Retrieval Evaluation Online Proceedings, 2004

[Snoek06] C.G.M. Snoek, M. Worring, and A.G. Hauptmann. Learning rich semantics from news video archives by style analysis. ACM Trans. Multimedia Computing, Communications Application, vol. 2, no. 2, May 2006, in press.

[Snoek06a] C.G.M. Snoek, M. Worring, J. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders, The Semantic Pathfinder: Using an Authoring Metaphor for Generic Multimedia Indexing, accepted by PAMI.

[Sorwar03] G. Sorwar, M. Murshed, and L Dooley. Fast global motion estimation using iterative least-square estimation technique. Fourth International Conference on Information, Communications & Signal Processing and Fourth IEEE Pacific-Rim Conference On Multimedia

15-18 December 2003, Singapore.

[Tan95] Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge. A new method for camera parameter estimation. in Processing of International Conference Image Processing, vol. 1, pp. 405-408, 1995.

[TRECVID] TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/trecvid/

[Truong00] Ba Tu Truong, Chitra Dorai and Svetha Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. in Proceedings of ACM Multimedia, Los Angeles, USA, 2000, pp. 219--227.

[Voorhees94] Ellen M. Voorhees , Query expansion using lexical-semantic relations, In Proceedings of ACMSIGIR. Dublin, Ireland, pages 61--69. ACM/Springer, 1994

[Wang06] D. Wang, J. Li, B. Zhang. Relay Boost Fusion for Learning Rare Concepts in Multimedia. CIVR 2006

[Wolpert92] D.H. Wolpert, Stacked Generalization, Neural Networks, Vol. 5, pp. 241-259, 1992.

[Yan04] Rong Yan, JunYang, Alexander G. Hauptman, Learning Query-Class Dependent Weights in Automatic Video Retrieval, Proceedings of the 12th annual ACM international conference on Multimedia, 2004

[Yuan04] J.Yuan et al. Tsinghua University at TRECVID 2004: Shot Boundary Detection and High-level Feature Extraction. In: Proceedings of TRECVID 2004 Workshop.

[Yuan05] Jinhui Yuan, et al. Tsinghua university at trecvid 2005, in Proceeding of TRECVID Workshop 2005, Nov. 2005.

[Yuan07] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin and Bo Zhang. A Formal Study of Shot Boundary Detection. IEEE Transactions on Circuits and Systems for Video Technology, 17(2):168-186

[Zabih95] Ramin Zabih, Justin Miller and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks, in Proceeding of ACM Multimedia 95, San Francisco, CA, Nov. 1995, pp. 189--200.

[Zhai01] Zhai, C. Notes on the Lemur TFIDF model. Unpublished report. 2001