THU and ICRC at TRECVID 2007

Jinhui Yuan, Zhishan Guo, Li Lv, Wei Wan, Teng Zhang, Dong Wang, Xiaobing Liu, Cailiang Liu, Shengqi Zhu, Duanpeng Wang, Yang Pang, Nan Ding, Ying Liu, Jiangping Wang, Xiujun Zhang, Xiaozheng Tie, Zhikun Wang, Huiyi Wang, Tongchun Xiao, Yinyu Liang, Jianmin Li, Fuzong Lin, Bo Zhang
Intelligent multimedia Group,
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing, P. R. China

JianGuo Li, WeiXin Wu, XiaoFeng Tong, DaYong Ding, YuRong Chen, Tao Wang, Yimin Zhang
Scalable Statistical Computing Group in Application Research Lab, MTL
Intel China Research Center, Beijing, P. R. China

Abstract

Shot boundary detection

The shot boundary detection system in 2007 is basically the same as that of last year. We make three major modifications in the system of this year. First, CUT detector and GT detector use block based RGB color histogram with the different parameters instead of the same ones. Secondly, we add a motion detection module to the GT detector to remove the false alarms caused by camera motion or large object movements. Finally, we add a post-processing module based on SIFT feature after both CUT and GT detector. The evaluation results show that all these modifications bring performance improvements to the system. The brief introduction to each run is shown in the following table:

Run_id	Description
Thu01	Baseline system: RGB4_48 for CUT and GT detector, no motion detector, no sift post-processing, only using development set of 2005 as training set
Thu02	Same algorithm as thu01, but with RGB16_48 for CUT detector, RGB4_48 for GT detector
Thu03	Same algorithm as thu02, but with SIFT post-processing for CUT
Thu04	Same algorithm as thu03, but with Motion detector for GT
Thu05	Same algorithm as thu04, but with SIFT post-processing for GT
Thu06	Same algorithm as thu05, but no SIFT processing for CUT
Thu09	Same algorithm as thu05, but with different parameters
thu11	Same algorithm as thu05, but with different parameters
Thu13	Same algorithm as thu05, but with different parameters
Thu14	Same algorithm and parameters as thu05, but trained with all the development data from 2003-2006

High-level feature extraction

We try a novel approach, Multi-Label Multi-Feature learning (MLMF learning) to learn a joint-concept distribution on the regional level as an intermediate representation. Besides, we improve our Video diver indexing system by designing new features, comparing learning algorithms and exploring novel fusion algorithms. Based on these efforts in improving feature, learning and fusion algorithms, we achieve top results in HFE this year.

Run_id	Description
B_tsinghua-icrc_Huanhuan[1]	MAP (0.125), rank based Borda fusion of all tsinghua + icrc basic run results of undersampling SVM, baseline SVM, protoline and combined training of both 05

[1] This run is mistakenly named as a type A run. However, it is a type B run since the protoline involves annotation from other datasets such as Labelme.

	and 07 training set with domain adaptation technique.
A_tsinghua-icrc_Olympic12008	MAP (0.131), baseline system, (26 features) x 5 round SVM, with feature selection by floating search.
B_tsinghua-icrc_Yingying	MAP (0.132), Type-B system (baseline of 26 feature + protoline feature + combline_0507) x 5 round SVM, Simple average fusion
A_tsinghua-icrc_Huanhuan	MAP (0.0837), the Undersampling SVM baseline run with 22 features. It outperforms general SVM classifiers
A_tsinghua-icrc_Jingjing	MAP (0.0875), SFFS run outperforms the USVM baseline about 5.7%. It shows that feature selection fusion is valid to slightly increase MAP performance.
A_tsinghua-icrc_Nini	MAP (0.0978), runs of TRECVID07 outperforms the USVM baseline about 16.5%

Search

Run_id	Description
A_1_tsinghua_1	Used only text-based search result
A_2_tsinghua_2	combined example-based search result and concept-based search result. The concept lexicon consists of 39 concepts from high-level feature extraction task.
A_2_tsinghua_3	combined text-based search result and example-based search result
A_2_tsinghua_4	combined all 3 modalities. The concept lexicon here consists of concepts selected from LSCOM
I_2_tsinghua_5	experts with default options
I_2_tsinghua_6	experts with manually adjusted options

This year, intelligent multimedia group in Department of computer science and technology, Tsinghua University and Scalable Statistical Computing Group in Application Research Lab, MTL, Intel China Research Center took part in TRECVID 2007 as a joint team and submitted the results for all of the four tasks. In this paper, the methods of shot boundary detection, high level feature extraction and search are presented while rushes summarization is excluded since it is reported in the workshop during ACM MM 2007.

1. Shot boundary detection

1.1 System overview

As shown in Fig. 1.1, the shot boundary detection system of 2007 is basically the same as that of previous years, the details of which can be found in [Tsinghua06, Yuan07]. In this paper, we only describe the modifications of this year, which are emphasized by the filled blocks Fig. 1.1.

1.2 Visual Representation by Block-based RGB Color Histogram

In the previous systems (i.e., 2005-2006), the content of each frame is represented by block-based RGB color histogram. For both CUT and GT detectors, we adopt the same grid partition for each frame, that is, each frame is partitioned into 4 by 4 blocks. However, according to the evaluation results of [Yuan07], to achieve best performance, CUT detector and GT detector may need different granularities. Therefore, in the implementation of this year, we adopt 16 by 16 blocks for CUT detector and 4 by 4 blocks for GT detector.

1.3 Motion Detector

Camera motion or large object movement is one of the major challenges to the GT detection [Yuan07]. In this year, we add a motion detector to the system so that we can remove the disturbances caused by camera motion or large object movement. The basic idea is, we evaluate the motion activity of each GT candidate yielded by GT detector. If the there exists strong motion activity within the candidate, we remove it from the shot boundary results. The method

consists of two steps. First, we calculate the main movement of each frame. We extract the motion vectors of each frame by block matching approach. According to the observation of [Att06], the motion estimation will be more robust with coarser block partition. In our implementation, each frame is partitioned into blocks in 48 by 48 pixels. Since there may exist noises in the produced motion vectors, we post-process them to get motion estimation with higher quality. Concretely, we firstly cluster the motion vectors of each frame into two groups by k-means approach. And then we use the centroid of the larger cluster to indicate the main movement in the current frame. Second, let mv_i indicate the main movement of the i -th frame, the motion activity ma of the GT candidate is obtained by averaging all the mv within it. If ma exceeds a pre-determined threshold, the GT candidate will be judged as a false alarm.

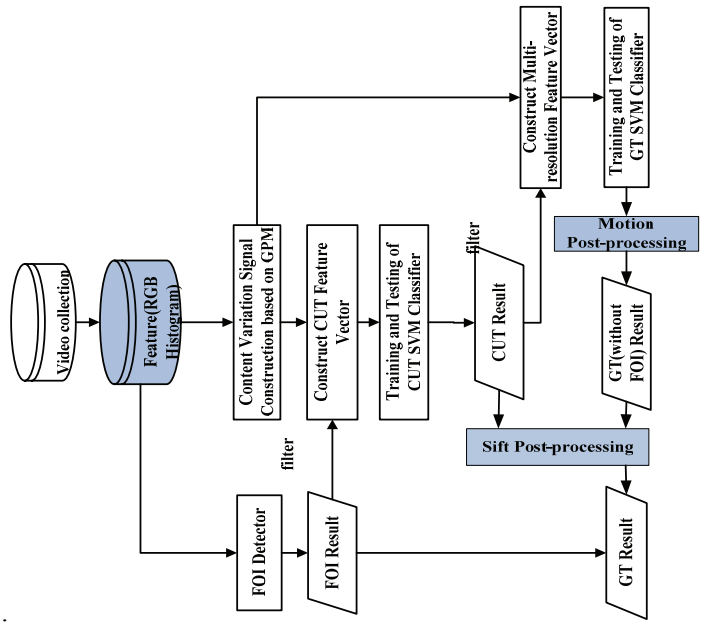


Fig. 1.1 The flowchart of SBD system 2007.

1.3 SIFT Post-processing

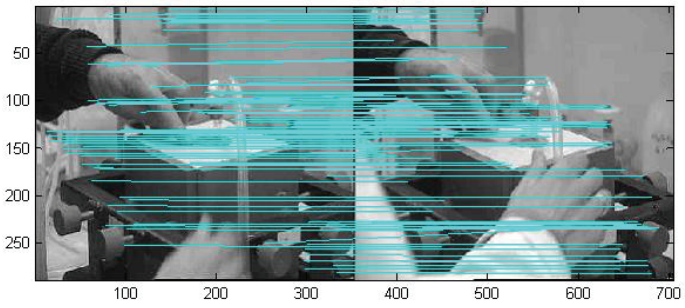


Fig. 1.2. The SIFT matching result of two frames.

Scale invariant feature transform (SIFT) is a kind of novel feature extraction approach, which is broadly used in many applications such as object recognition and image registration [Lowe04]. SIFT has many appealing properties, including scale invariant, direction invariant, robust to light variation, highly distinctive, etc, compared to traditional color and texture feature. That is why SIFT is good at find the same objects in pictures of different scale or view. However, though SIFT is popular in other applications, it is seldom used in shot boundary detection task [Park06]. In the implementation of this year, we add a post-processing module based on SIFT feature to both CUT and GT detectors. The basic idea is, with the shot boundary candidates (CUT or GT), we examine whether the frames before

and after the candidate contain the same objects by the matching of SIFT feature. If the match ratio is high enough, there may exist the same objects in the frames before and after the candidate. The system will determine the candidate as a false alarm. As shown in Fig. 1.2, part of the second frame is covered by human body. Since the color feature of the two frames will differ a lot, the CUT detector based on color histogram may determine it as a CUT. However, with SIFT feature, the system can infer that the two frames belong to the same shot by strong SIFT matching.

1.4 Experiment and Evaluation

We submitted 10 runs this year. The detailed performance of each run is show in Table 1.1. The evaluation result shows that our system is among the best. In the next, we will show that all the modifications of this year bring performance improvement to the system: (a) comparing *thu01* and *thu02*, the finer block partition color histogram improves the CUT F-measure from 0.964 to 0.969; (b) comparing *thu03* and *thu02,* SIFT post-processing for CUT detector improves the CUT F-measure from 0.969 to 0.973; (c) comparing *thu04* and *thu03*, motion detector improves the GT F-measure from 0.542 to 0.680; (d) comparing *thu05* and *thu04*, the SIFT post-processing for GT improves the GT F-measure from 0.680 to 0.718; (e) comparing *thu14* and *thu05*, we find the increase of the size of training set may not lead to performance improvement.

1.5 Conclusions

The evaluation results show that our system is one of the best. The experiments also validate that our modifications to the system lead to the performance improvement. Especially, the motion detection and SIFT feature play important role in the system. However, in our current implementation, all the above modifications are rule-based, which include several heuristically chosen thresholds. This may limit the system generalizing to the videos from other domain. In [Acmmm07], we show that conditional random fields (CRFs) may be an effective way to automatically fusing the various clues useful for shot boundary detection. We expect that CRFs may address the shortcomings of rule-based methods in the future systems. Finally, we find that there is an old archival video in the test set of this year, i.e., *BG_11362.mpg*. Such videos with low quality bring significant challenges to our system. Developing effective shot boundary detection system for such videos will also be an interesting problem.

Table 1.1: Evaluation results of the 10 submissions

Sysid	All Transitions			Cuts			Gradual Transitions			Frame GT Accuracy		
	Rcl	Prc	F#	Rcl	Prc	F#	Rcl	Prc	F#	Rcl	Prc	F#
thu01	0.955	0.877	0.914	0.969	0.96	0.964	0.791	0.406	0.537	0.755	0.799	0.776
thu02	0.958	0.884	0.920	0.973	0.966	0.969	0.791	0.411	0.541	0.755	0.799	0.776
thu03	0.954	0.893	0.922	0.968	0.979	0.973	0.796	0.411	0.542	0.755	0.798	0.776
thu04	0.95	0.943	0.946	0.968	0.98	0.974	0.757	0.617	0.680	0.77	0.848	0.807
thu05	0.949	0.956	0.952	0.968	0.982	0.975	0.743	0.695	0.718	0.787	0.852	0.818
thu06	0.953	0.944	0.948	0.973	0.969	0.971	0.733	0.689	0.710	0.786	0.856	0.820
thu09	0.947	0.958	0.952	0.968	0.982	0.975	0.714	0.7	0.707	0.785	0.865	0.823
thu11	0.947	0.962	0.954	0.968	0.982	0.975	0.718	0.733	0.725	0.783	0.854	0.817
thu13	0.948	0.956	0.952	0.968	0.982	0.975	0.733	0.689	0.710	0.786	0.856	0.820
thu14	0.948	0.955	0.951	0.969	0.981	0.975	0.728	0.685	0.706	0.793	0.855	0.823

2. High level feature extraction

This year, we try a novel approach which warps the labeling information of many concepts with many features to learn a joint-concept distribution on the regional level as an intermediate representation. We call it as Multi-Label Multi-Feature learning (MLMF learning), and system using this kind of feature as the protoline system. This approach has the following advantages: 1) It improves over the early fusion approach by selecting a few discriminative feature

dimensions in the Multi-class Boosting algorithm; 2) It improves the late fusion approach by counting the feature correlations properly and greatly reduces the expensive SVM testing operations at run-time; 3) It provides a natural mechanism for inductive transfer (or domain adaptation) by analyzing the basic multi-concept at regional level which is much more invariant to domain changes.

We also improve our Video diver indexing system by designing new features, comparing learning algorithms and exploring novel fusion algorithms. Some new features are shown to be effective, such as Edge Coherence Vector, Segmentation based color and shape statistics, etc. Under-sampling SVM (USVM) and SVM with RankBoost, direct Boosting algorithms are compared. For fusion, sequential forward floating feature selection (SFFS), simulated annealing fusion (SA) and Borda based rank fusion approaches are designed and compared respectively. Some specific detectors are built for person relevant concepts such as "crowd", "marching", "person" and "walking-running". We implemented some parallel optimization technology both on the task-level and the data-level to get linear speed up on multi-core systems for the feature extraction module.

Based on the new fusion approach and persistent effort in improving feature and learning algorithms, we achieved top results in HFE this year. We are in the 1st to 3th position in the highest in MAP of all submitted runs, with a margin of 30% MAP gain over the next run. At the same time, a top result for 9 out of 20 and a top-3 result for 15 out of 20 concepts are obtained.

2.1 Annotation

This year, we examine the community based annotation and change a small portion of the positive labels for a few specific concepts based on our observation to achieve consistency. In addition, we also annotate a few concepts related to scene and object with a bounding box on the past and current TRECVID development dataset for refinement. Additional annotations of related scenes are also extraction from the Labelme dataset and converted from polygon to maximum bounding box. The external annotations are only used for the MLMF learning approach.

2.2 Baseline

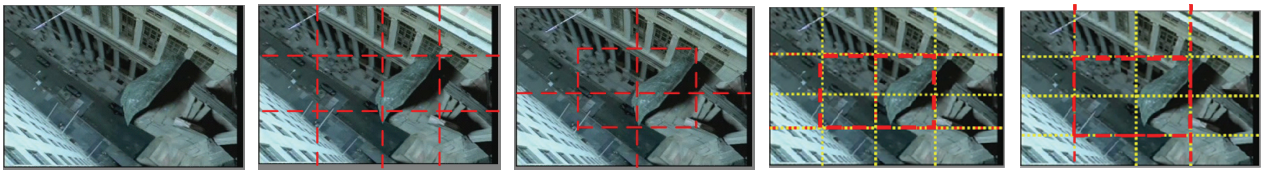
Our baseline follows the same framework in the TRECVID 2006 benchmark [Tsinghua06]. Three types of features are used, namely color, texture and edge. Besides the effective Color Moment, Haar Moment and Edge histogram features with a grid-layout partition, and the local keypoint features based on the SURF detector / descriptor [Liu07], a few new features are employed and shown to be effective. Given all these features, five Support Vector Machine (SVM) classifiers are trained with sub-sampled positive and negative examples. The SVM kernels are chosen with respect to the feature type. A five round RankBoost re-sampling procedure is carried out to generate balanced examples for each SVM classifier. Then a simple average fusion is taken afterwards. Please see [Wang07] for further details. Such visual-only baseline technique has been shown to be effective in the past TRECVID experiments [Tsinghua06].

2.3 New Features to explore the shape and spatial information

It has been shown that features are vital for the success of vision, both for human and for computers. Though the color and texture features are easy to extract and robust to view changes, edge based features are believed to have the closet relation with shape, which is very important for the human vision system. Simple edge extracted from canny or other operators is susceptible to view and illumination changes. The Edge histogram feature gains robustness at the cost of losing spatial information. However, we know from past experience that the spatial layout and different spatial granularities are both useful cues for capture concepts. So we systematically extend the edge features to account for both spatial position distribution and different granularities. Different partition schemes of the full image are evaluated and new features are proposed to represent spatial correlations. Connected and smoothed lines instead of raw edges are also extracted to assess the power of intermediate edge features. We also evaluate many kinds of existing features which are useful for many applications.

Concatenated features on the region level leverage on the spatial success. In addition to the traditional global feature

and grid-layout features shown in Fig. 2.1. (a) - (b), we implement three kinds of spatial partitions as shown in Fig. 2.1. (c) - (e) to fully explore the spatial features. Please notice that in (d), the first row and the last row are two horizontal regions and the left and the right part of the central two rows are two other regions, and the middle red rectangle is the fifth one. So there are five regions altogether. (e) is a transposed version of (d). The last three partitions are effective for basic features with larger dimensions since they can control the overall result feature dimension.



(a) global (b) gr4x3 (c) grsh (d) grh5 (e) gri5

Fig. 2.1. Different spatial layout partition schemes

Beyond the edge histogram, we borrow idea from the color auto-correlogram and color coherence vector to design edge auto-correlogram (EAC) and edge coherence vector (ECV) features. The edges are first detected with a Canny operator, and then quantized by both magnitude and orientation. These two features capture the local edge distribution by considering neighboring edge of the same magnitude and orientation, or by considering the connected components of similar edges, respectively. Also Hough transform is applied on the edge map and straight lines are obtained and global statistics based on the line parameters are gathered and used.

A new feature of SegShapeColor (SSC) combines shape context [Belongie02] and color moment on a segmented image. One advantage of this feature over existing segmented features is its fixed feature output dimension. The feature is extracted in three steps. First, we put the segmented image into a log-polar space by a reference point. Second, shape context is extracted from the segmented boundary. Third, the original image is quantized according to segmented regions. Finally, we compute color moments of the quantized image by the log-polar regions.

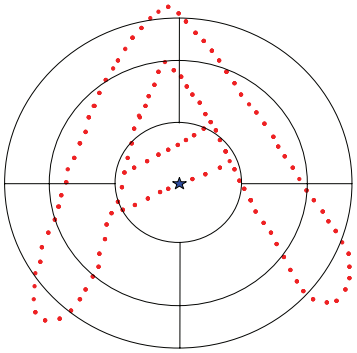


Fig. 2.2. Extract SegShapeColor feature in 9 log-polar regions.

These features of EAC, ECV and SSC can all be seen as a combination of two successful operations which are originally from independent effort, e.g. ECV combines edge and local characteristics and SSC combines segmentation and color with the shape context. They are part of a continuous effort towards imitating some hierarchical buildup of human vision invariance [Riesenhuber99].

Other features like Gabor [Lee96], MRSAR [Mao92] (multi-resolution simultaneous auto-regression), LBP [Ojala02], and Shape Context [Belongie02] are also implemented and optimized for speedup. Overall, we have 26 kinds of different configuration of feature and spatial partition combinations, some are old and some are new.

Our experimental results show that our local keypoint (SURF [Bay06]) based representation does not work as well as last year. We are trying to figure this out. For detailed feature results, see the result section.

2.4 Learning

A different approach is implemented to counter-balance the imbalanced data problem. The under-sampling SVM (USVM) [Akbani04] under-sample the negative samples into K (for example) non-overlapped subsets, and combine every negative subset with positive samples to form K training sets. Then, we train probabilistic SVM model on each of the K training sets, and obtain K models for each low-level feature on one concept. The parameters of SVM (C, i.e., complexity; γ, i.e., the coefficient in RBF kernel) are determined in a coarse-to-fine searching manner by cross-validation in the training phase. The sampling value K ($K \in \{1,3,5,7,9\}$) is determined by the ratio of negative samples and positive samples for each concept. The larger the ratio is in a concept, the larger the value of K is. In the testing phase, K models output K predictions for each concept and each feature. The baseline run fuses the prediction from all low-level features. The benefit is that USVM can handle the imbalance between large negative samples and small positive samples [Akbani04] and outperforms general SVM classifiers.

In this approach, 22 low-level visual features with possibly different spatial partitions are processed with USVM. This serves as another baseline this year.

For the keypoint based representation, we implement our Feature and Spatial Covariant (FESCO) kernel in [Liu07] to combine the spatial information at the learning level. This approach augments is motivated to augment the holistic histogram representation with implicit spatial constrains. This solution is both accurate and fast. Partly adapted from the spatial pyramid match kernel (SPM) [Lazebnik06], this scheme achieves better match accuracy than SPM [Lazebnik06] and PM [Grauman06]. Please refer to [Liu07] for details. The combined result verifies our findings again with a 33% performance gain over the best single level keypoint result.

Recently, researchers are interested in domain adaptation [Sugiyama07] (inductive transfer) which aims at utilizing dataset of different distributions for a common data set. This year we also try a few approaches such as Covariant Shift. But the datasets are so different, that the approach does not work very well.

2.5 Intra-concept Fusion

Intra-concept fusion deals with multiple features for a given concept. It is proven by many past experiments that fusing multiple features altogether will do much better than any single feature result. However, no one has been able to answer how many features are enough for fusion, what's the correlation between different concepts, either experimentally or theoretically. We had pointed out, through our past and current experimental results, that the fusion problem is inherently difficult [Wang07]. Even for the simplest weighted linear fusion, it is hard to choose the proper weights for each feature across data sets.

This year, we tried four different kinds of methods to learn the weight for each feature. The first one is a boosting algorithm which approximately expresses Average Precision on a per-sample weight basis and tries to optimize it directly. The second approach uses advanced Genetic Algorithm by sampling from a parameter space distribution which is described by a histogram. This Histogram-based Estimation of Distribution Algorithm learns the weights by mutation and evolution. The third method is based on Simulated Annealing. Simulated annealing (SA) is a heuristic optimization algorithm for difficult combinatorial optimization problems, especially ones where a desired global extreme is hidden among many poor local extremes [Kirkpatrick83]. SA fusion approach brute-force search optimal weights to linearly fuse results from different models. The last one method is Probabilistic Model Supported Rank Aggregation (PMSRA) [Ding07]. PMSRA is a robust and effective rank fusion approach. For the PMSRA method, order statistics are used to model the probability that a shot of a given rank is likely to be relevant. According to the probabilistic models built from training lists, the probability for shots to be relevant given fusing lists can be estimated and thus become the criterion for fused ranks.

Based on the observation that not all low-level features are useful for every concept, the selective fusion run employed feature selection from all low-level visual features to further improve the performance. The sequential forward

floating selection algorithm (SFFS) [Pudil94] selects effective low-level features for each concept during the training phase, and gets final results based on the fusion of the model from selected low-level features.
Also a simple Rank based Borda fusion method is employed for easy comparison and implementation.

2.6 Inter-concept Fusion

Exploiting relationships between multiple semantic concepts, the Inter-concept fusion tries to find related concepts in a given concept lexicon to enhance the concept detection performance. We employ a Bayesian Dirichelet Metric to learn a concept relation and then use simple one-layer Neural Network for modeling the pair-wise correlations.
Taken as a unified feature selection problem where each feature is a concept detector, SFFS can also be used in this part. We also developed it here.

2.7 MLMF Learning

As stated above, fusion has been a central problem for video indexing and tremendous efforts had been put on it. However, it is still far from a solved problem. A well-known dichotomy splits fusion algorithms into feature fusion and classifier fusion, or equivalently, early fusion and late fusion. It is generally accepted that late fusion is better than early fusion in that the former avoids the notorious curse of dimensionality. However, the latter also has the potential to count for correlations between different features. In addition, the concept context which takes different concept labeling information into account can be seen as a further fusion step after each concept detector. This year, we also designed a new framework which is different from the de facto pipeline framework significantly.
The pipeline system follows the late fusion scheme of one-feature-one-concept (OFOC) at a time, with/without a concept fusion step. In sharp contrast, we warp the labeling information of many concepts with many features to learn a joint-classifier as the very first step. Given a specific region with many-feature-many-concept (MFMC) input, the joint-classifier is learned to output multi-class probabilities. This approach takes a joint boosting algorithm [Torrabala04] as a region-based front-end feature extractor. Then a standard SVM classifier with the RankBoost Resampling technique is applied to learn a target concept based on the concatenated regional multi-class probabilities. This approach has the following advantages: 1) It improves over the early fusion approach by selecting a few discriminative feature dimensions in the Multi-class Boosting algorithm; 2) It improves the late fusion approach by counting the feature correlations properly and greatly reduces the expensive SVM testing operations at run-time; 3) It provides a natural mechanism for inductive transfer (or domain adaptation) by analyzing the basic multi-concept at regional level which is much more invariant to domain changes. The drawback is that regional annotations are required to run the Multi-class Boosting algorithm. Further extension to this rudimental exploration can be expected.

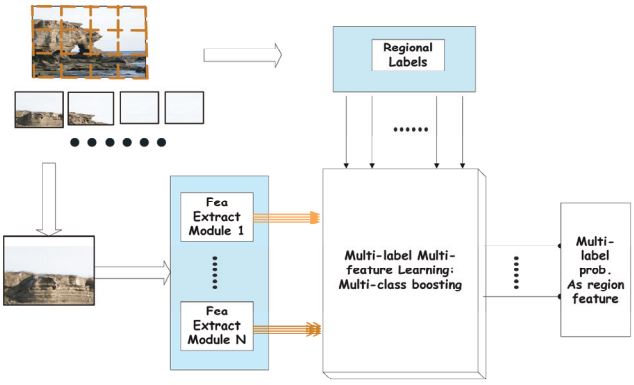


Fig. 2.3. Multi-label Multi-feature Learning as a feature extraction front end

This approach is somewhat similar to the Proto-concept idea which is proposed by the MediaMill group in TRECVID

2006. However, our approach differs from theirs in the following ways: We aim at learning a feature prior for many concepts while they focus on clustering each concept in one feature space; We use a discriminative multi-label Boosting framework which can easily handle the multi-label problem while they adopt a generative approach; Furthermore, we are interested in designing a new fusion algorithm which could take advantage of multi-label and multi-feature at the same time while as they are separated into two parts in the past algorithms.

2.8 Specific detectors for face and person related concepts

In the HFE task, there are many object relevant concepts, e.g. face, person, crowd, car, airplane and computer screen concepts. For face and person concepts, we tried object detector to get better performance than global features.

2.8.1 Face detection

We used sparse feature based boosting face detector to detect faces [Huang05]. The face detector is robust with high precision. The ratio of total face area to image area is calculated as the confidence score.

2.8.2 Person detection

We detected person concept by integrated boosted histograms of oriented gradients (IBHOG) approach. The algorithm includes four modules: 1) face detection; 2) boosted histograms of oriented gradients, 3) color-texture segmentation, and 4) probabilistic SVM score. According to our observations, the images with detectable faces (size > 20x20) are labeled as "person". The other images (no person, or has person but with small/blur face, or has person with back-view, etc) are examined through a boosted histograms detector. To remove some possible false alarms, we use a color-texture segmentation to verify the detected rectangles. If a rectangle-area belongs to a single segmented region, there is no person in this rectangle with high probability. In order to rank the probability of a detected person region, we use a probabilistic SVM model trained by the concatenated histograms of oriented gradients in the whole detection window to predict the confidence score.

2.9 Results

2.9.1 Per run result

As shown in the Fig. 2.4, the runs from our group is in the 1st to 3th position in the highest in MAP of all submitted runs, with a margin of 30% MAP gain over the next run. At the same time, a top result for 9 out of 20 and a top-3 result for 15 out of 20 concepts are obtained.

Give the large possible variations in the system, we fuse or select only a few representative methods from our experiments on the training data set, and many others remains to be explored in the future.

The baseline run combines 26 kinds of features and achieves an MAP of 0.131. Further experiments show that the MLMF learning is nearly as effective with only 9 kinds of simple features, a close MAP of 0.123. So MLMF learning is worth further exploration.

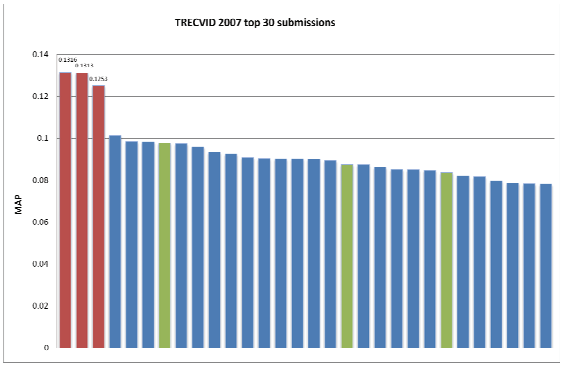


Fig. 2.4. Multi-label Multi-feature Learning as a feature extraction front end

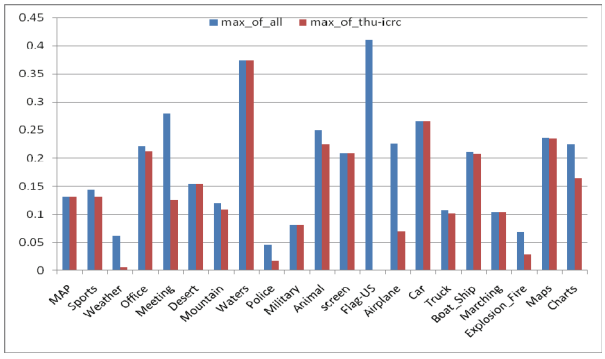


Fig. 2.5. Multi-label Multi-feature Learning as a feature extraction front end

From the runs submitted, we can see that the USVM run outperforms general SVM classifiers for imbalanced data. A

further SFFS run outperforms the USVM baseline about 5.7%. It shows that feature selection fusion is valid to slightly increase MAP performance. The B_tsinghua-icrc_Nini run of TRECVID07 outperforms the USVM baseline about 16.5%. The two fusion methods of SA and SFFS are switched also by testing on a validation dataset in a pre-located training dataset. On 16 concepts, SA approach outperforms USVM baseline about 16.9% and the sports and explosion concepts got the best performance across all runs. In addition, another main reason may be from that it fuses THU detection results besides ICRC results.

2.9.2 Per-concept result

The per-concept best result is shown in Fig. 2.5. Clearly, we did not work well on the concepts like Weather, Meeting, Police, Flag-US, and Airplane. However, the weather concept is ambiguous for our team members.

Also, we select a few good features to evaluate the single feature performance and report their performance in Table 2.1. We can see that the first few features are all excellent and edge based features are very useful. And some newly proposed features are among the best few.

Table 2.1. Some feature names, descriptions and performances

Feature Name	Description	MAP
grsh_eh64	Edge Histogram with grsh partition	0.075
grsh_ecv64	Edge Coherence Vector with grsh partition	0.063
gri5_ct48	GLCM feature with gri5 partition	0.056
grsh_gabor48	Gabor feature with grsh partition	0.054
gr4x3_hm10	Haar Moment feature with gr4x3 partition	0.047
g_shaperef981	Shape Reference feature with global partition	0.045
gri5_ccv72	Color Coherence Vector feature with gri5 partition	0.045
g_shapecont965	Shape Context feature with global partition	0.043
g_jseg_shapecolor	Segment Shape Color feature with global partition	0.043
g_mrsar	MRSAR feature with global partition	0.041
g_lbph	LBP feature with global partition	0.034
grh5_hline_q160	Line feature based on Hough Trans. with grsh partition	0.023
g_ac64	Color AutoCorrelegram feature with global partition	0.022
g_eac64	Edge AutoCorrelegram feature with global partition	0.020

It is also very interesting to compare the most useful features in the last three years. In the year 2005, Color Moment and Edge Histogram are good. In 2006, keypoint based features are proven to be useful. In 2007, it seems that too much gray images are incorporated in the dataset so that color based features are ineffective. The edge based features are very strong.

Table 2.2. The keypoint feature names, descriptions and performances

Feature Name	Description	MAP
Combined fesco	The combined result of three keypoints features	0.053
g_hsurf_kmlocal_q288	SURF keypoint with 288 code book, with global basis	0.036
gr2x2_hsurf_kmlocal_q72	SURF keypoint with 74 code book, with gr2x2 basis	0.040
gr4x4_hsurf_kmlocal_q18	SURF keypoint with18 code book, with gr4x4 basis	0.036

For the keypoint based representation, we also run multiple level matches and combine them in a FESCO framework. The performance increases significantly with over 30% gain. So we once again verified the effectiveness of this approach [Liu07].

2.10 Key frame extraction and the influence of sampling rate

In this year, the testing key frames are extracted by each attendee. In our system, we extract key frames in each shot by a leader-follower clustering algorithm. Assuming a shot S_i is a frame set $S_i=\{f^a, f^{a+1}, \ldots, f^b\}$. The shot key frames can be efficiently extracted by the following algorithm:

Step 1: Select middle frame as the first key frame

$K_i \leftarrow \{f^{[(a+b)/2]}\}$

Step 2: for j = a to b

If $\max(ColSim(f^j, f^k)) < T_h \quad \forall f^k \in K_i$

Then $K_i \leftarrow K_i \cup \{f^j\}$

where *ColSim*( ) is the similarity of color histogram feature. T_A is the similarity threshold, 0.85 in our system.

By this algorithm, 67,662 keyframes are generated for 17,986 shots in the test data set, 3.76 keyframes per shot. This is a rather dense sampling scheme compared with the train data set. We also decreased the sampling rate to see whether the sampling rate is crucial for indexing performance. As the sampling rate decreased to only 1/5 (corresponds to 0.2 on the X-axis in Fig. 2,6) of the original sampling rate by random selecting the keyframes, the performance drops only a tolerable 10%. So it seems that we can safely downsampling the keyframe selected for each shot when building the shot level index. However, further experiments on different keyframe sampling schemes are needed to fully answer this question.

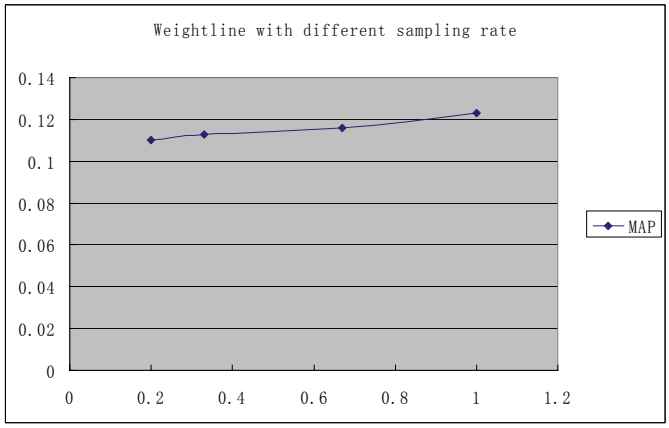


Fig 2.6. The testing set sampling rate's influence on MAP

2.11 Parallel computing

The HFE application is highly compute intensive, which requires tera-scale computing power. To speed up time-consuming HFE task, we use performance optimization and parallel computing techniques to accelerate the feature extraction, SVM classifier training and testing modules. Both Tsinghua and ICRC decomposed the large-scale SVM training and testing task into many small jobs, and adopted a cluster to execute the small tasks in parallel. Additional techniques, all from the ICRC side, include:

- Serial performance optimization by loop unrolling, lookup table, data alignment, SIMD optimization and Cache-conscious optimization.
- Use highly (parallel) optimized Intel's library such as IPP, MKL, and especially OpenCV. OpenCV contains a machine learning sub-library, which is much more efficient than other known open source implementations.
- Parallelization of the low-level feature extraction both in task-level and data-level parallelism and parallel performance optimization of these tasks to obtain almost linear speedup on multi-core systems.

2.12 Conclusions

From the experiments in Trecvid 2007, we found that:

- New features are vital to success

- Spatial information is of additional value
- Rankboost Resampling is good enough
- Simple fusion works pretty well
- As two sides of one coin, fusion and dataset adaptation remains difficult.

3. Search

We took part in both automatic and interactive retrieval tasks this year. Our system framework was mostly inherited from previous work, but many new methods were adopted.

3.1 Automatic Search

3.1.1 Building a Training Dataset

To conquer the lack of prior knowledge, we built a topic set by combining the topics from search task 2005 and 2006 and then excluding the duplicated ones and those about named person. The ground truth of those topics was manually annotated and collected in training video set, which was used to evaluate the performance of different methods.

3.1.2 Text-based Search

For English text search, keywords in a query were first extracted before query expansion, using the method inherited from last year. Then we look to add additional words to the query which are closely related to the keywords. WordNet was involved in the query expansion progress. Specifically, synonyms of each keyword were extracted and refined by removing words with many senses which were considered to cause ambiguity.

Speech text is translated from Dutch to English, and segmented in different levels: video level, story level and shot level. The "story level" here is partitioned simply by speaker information and visual clues such as color histogram. We use Lucene to index those translated speech transcripts and calculate the scores of each videos/stories/shots. Specially, for shot level retrieval, we adopted a temporal spread method as what we did before. In this method, the score of a certain shot will propagate and affect its temporal neighbors. This method was proven to be useful in experiments on the training dataset. Linear combination was used for fusion of the results from three levels, while the weights of each level were derived from the training data.

It is almost the same for Dutch text search, except that the speech transcripts were directly derived from videos and keywords from topic description were translated into Dutch. The text baseline was the average of two languages.

3.1.3 Example-based Search

This year we had a richer visual feature set instead of only edge histograms and color moments that we used before. Those new features, including Gabor, gri3_ctn48, grh5_ceh64, etc., which were described in High-Level Feature Extraction session, exhibited beautiful performance upon the training dataset. For each feature, SVM models were built in order to classify images corresponding to each topic. We still used radial basis function in our system, which performs better than other kernel functions, and selected parameters for each visual feature on the training dataset.

Since a much larger feature set was available, we faced a problem of choosing appropriate features for a certain topic. We introduced a measure for each feature on each topic: the ratio of average distance under this feature space among image and video examples (from this topic's description) and average distance among TRECVID 2007 video documents. The intuition is that, for certain topic, feature with smaller ratio will describe the topic better. Besides this theoretic measure, we also consider a practical measure, by setting up an experiment to evaluate the performance of each feature in the training dataset. We believed it was most convincing to take both measures into account. Simply, we took the product of the above two measure as the final criterion of performance of features, for certain topic, and choose the features which rank top 4 on this criterion. In TRECVID 2007 search topics, the most frequently selected features were gr6x5_protoconcept_r2600, grh5_ceh64, gri3_ctn48, grh5_gabor48 and gr4x3_cm9, they were outstanding for both measures in most cases.

Fusion of results from different features is also important. We tested several fusion methods and finally chose linear combination due to its simplicity and consistence performance. The weight of each feature is obtained from experiments on training dataset.

3.1.3 Concept-based Search

Two set of lexicons in this year were adopted by our system. The first one was the sets of concept annotations made publicly available as part of LSCOM. Concepts in these lexicons were chosen based on extensive analysis of video archive query logs and related to program categories, setting, people, objects, activities, events, and graphics. We filtered out concepts with less than 20 positive examples in the training set and get a number of 374 concepts left. Given the LSCOM concept annotation on a training set, we followed the state-of-the-art concept detection system to build the semantic concept indices. The second set of lexicons was consists of concepts in high-level feature extraction task. These two set of lexicons were used by same method, and different combinations were tried in different submitted runs.

We found concepts related to a topic in two ways, namely text-concept mapping and image-concept mapping. We collected text description for the concepts and then built an index. The text-concept mapping was using a text retrieval method to calculate scores that present the relation between concept and topic. On the other hand, the image-concept mapping followed the algorithm introduced in. The topic-concept mapping is a fusion of two ways mentioned above.

The topic-concept mapping result for each topic and the concept detection result for each shot are both kept as a concept vector. Retrieval in this modality is simply calculating the dot product of two concept vector.

3.1.4 Multi-modality fusion

The query classification method was not used this year since we have no enough samples to train respective model for each query class due to limited training dataset. Our system simply chooses linear combination to fuse those 3 modalities, and tried different kinds of combination. We finally submitted 4 automatic results:

- **run1**: used only text-based search result.
- **run2**: combined example-based search result and concept-based search result. The concept lexicon consists of 39 concepts from high-level feature extraction task.
- **run3**: combined text-based search result and example-based search result.
- **run4**: combined all 3 modalities. The concept lexicon here consists of concepts selected from LSCOM.

Although there still had other combinations, those 4 were considered more promising due to experiments set up on the training dataset.

3.1.5 Result

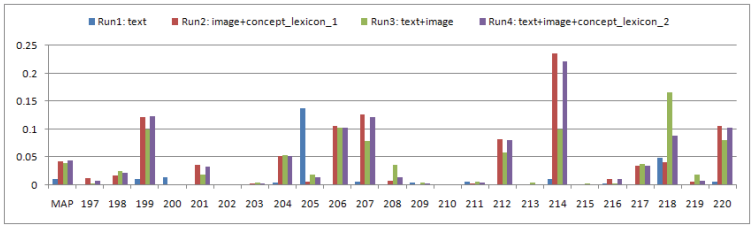


Fig. 3.1 Automatic search result

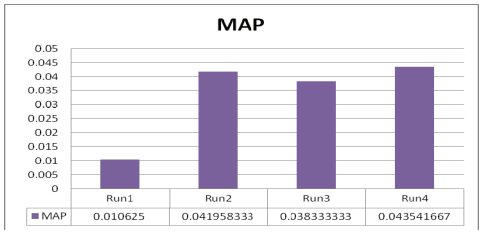


Fig. 3.2. MAP for each automatic run

The evaluation result is shown in Fig. 3.1 and Fig. 3.2. From the result, it is found:

- Text baseline drops significantly comparing with that of previous years. In our opinion, the main reason is the data this year is no longer news video. The search strategy in last year is not suitable again.
- Image baseline is very close to that of run4 which fusion of all modalities. Run4 adopted a larger concept lexicon. However, those concepts outside high-level feature extraction task were trained using data from other dataset. Thereby, run4 suffers from low quality of their extra concept detectors and shows little advantage.

- By comparing run1, run2 and run4 against run3, we can find how example-based search and concept-based search respectively play a role in each topic. For those topics with significantly corresponding concepts in our lexicon, concept-based search is preponderant.

3.2 Interactive System

3.2.1 System

This interactive system was improved upon the framework of SmartV, which was presented in VideoOlympics showcase of CIVR.

At the server end we included richer visual features, e.g., Garbor, ProtoConcept, etc. We also prepared two concept lexicons with 374 concepts chosen from LSCOM and 39 concepts from HLF task, respectively, accompanying with corresponding retrieval models. Based on them, richer retrieval options, consisting of plenty of retrieval models and parameters for each visual feature and concept lexicon, are available. The new retrieval models contain SVM classifier for each feature and two concept-based models.

For concept-based model, we mine a few concepts relevant to the query in a two-step process: 1) collect a set of relevant shots by the relevance feedback techniques, 2) and rank concepts in the shot set in light of their c-tf-idf scores. The c-tf-idf measurement, proposed by, can be viewed as the amount of information a concept bears, given the query. Specifically the tf counts for the concept popularity and the idf counts for the concept specificity. The method is similar as the image-concept mapping in automatic search; however, the text-concept mapping is discarded since we thought that query-by-concept can entirely substitute it. Those selected concepts form a concept subspace, then our system scores shots in this subspace, and fuses these scores with other search models to determine the coming result. For the sake of efficiency, we adopt the Vector Model for relevance ranking and a linear combination strategy for fusion. Simulative experiments conducted on the TRECVID 2006 search dataset show that search performance can significantly improve by leveraging the concept feature in this scenario.

For this complex server end, our system supplied a default options. However, we also provided another mode to experienced experts which allow users to adjust options, including weights for models and feedback processing procedure, based on specific topic, using their knowledge and experience. We tried these two modes, and submitted the results. The run5 was done by experts with default options, and the run6 was done also by experts with manually adjusted options. On the other hand, to compare the difference of expert and novice, we submitted a supplementary run which was done by novices.

A standalone user interface was designed and developed to replace the web-based interface of SmartV, for accomplishing higher effectiveness and efficiency. Besides tried-and-true temporal expansion thread and visual neighbors thread, we supplied a frame-level browsing thread, based on the keyframes extracted by Intel, to help user easily identify the specific events in the videos. Considering the inconveniency caused by multi-thread browsing, we added the FORWARD, BACKWARD and BOOKMARK functions. With the help of those functions, users can go along one thread and then jump back to the start point and choose another potential thread that was proven to be helpful in the user study. To achieve higher view, user interface were developed as a double-screen program, as shown in Fig. 3.3.

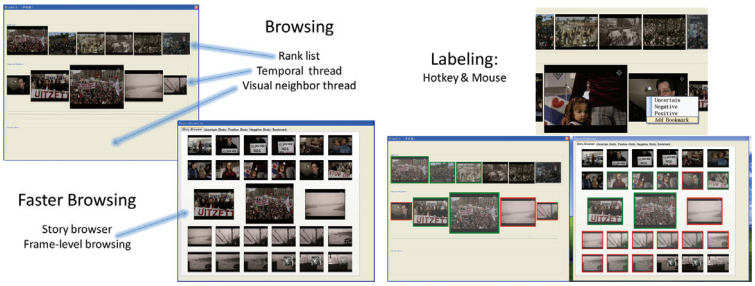


Fig. 3.3. Double-screen user interface

3.2.2 Result

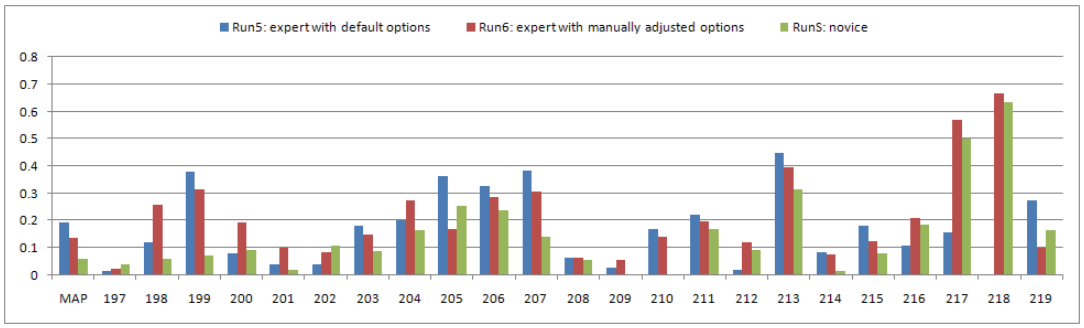


Fig. 3.4. Interactive search result

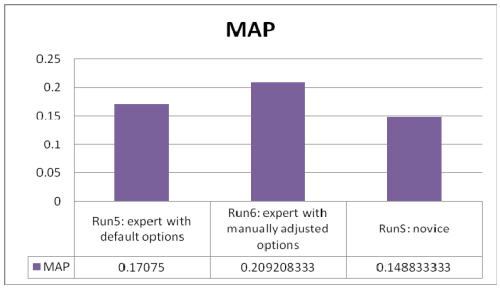


Fig. 3.5. MAP for each interactive run

Different results prove the positive effect of manual options. And small gap between expert and novice shows our system is not difficult to get started.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under the grant No. 60621062 and 60605003, and the National Key Foundation R&D Projects under the grant No. 2003CB317007 and 2004CB318108.

The authors would like to acknowledge the following organizations and people for their supports: Yang Tang, Rui Zhang, and Zhihu Ren for data preparation, NLIST for usage of THPCC, annotation community of TRECVID 2007, D. Lowe for SIFT binary, H. Bay for SURF binary and C.-J. Lin for LIBSVM.

References

[Acmmm07] Jinhui Yuan, Jianmin Li, Bo Zhang. Gradual Transition Detection with Conditional Random Fields. In Proceedings of ACM Multimedia 2007.

[Akbani04] Akbani R., Kwek S., Japkowicz N. Applying support vector machines to imbalanced datasets. ECML, 39-50, 2004.

[Att06] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, P. Haffner. AT&T Research at TRECVID 2006. In: Proceedings of TRECVID 2006 workshop.

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

[Belongie02] Belongie, S., Malik, J., and Puzicha, J. Shape Matching and Object Recognition Using Shape Contexts. IEEE Trans. PAMI, 24(4), pps.509–522, 2002.

[Ding07] Ding, D. and B. Zhang (2007). The Probabilistic Model Supported Rank Aggregation for Semantic Concept detection in Video. ACM International Conference on Image and Video Retrieval (CIVR07).

[Grauman06] K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image

features (version 2). Technical Report CSAIL-TR-2006-020, MIT, 2006.
[Huang05] Chang HUANG, Haizhou AI, Yuan LI, Shihong 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.
[Kirkpatrick83] S Kirkpatrick, CD Gelatt, and MP Vecchi, Optimization by Simulated Annealing. Science, Vol.220, 1983.
[Lazebnik06] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. of CVPR 2006.
[Lee96] Lee, T. S., "Image representation using 2D Gabor wavelets, IEEE Trans. PAMI, 18(10), pps. 959–971, 1996.
[Liu07] X. Liu, D. Wang, J. Li, and B. Zhang. The feature and spatial covariant kernel: Adding implicit spatial constraints to histogram. In Proc. of CIVR 2007.
[Lowe04] David G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2): 91-110 (2004).
[Mao92] Mao, J. and Jain, A.K. Texture classification and segmentation using multi-resolution simultaneous autoregressive models. Pattern Recognition, 25(2):173-188, 1992.
[Mikolajczyk05] Mikolajczyk, K. and Schmid, C. A performance evaluation of local descriptors. IEEE Trans. PAMI, 27(10):1615-1630, 2005.
[Ojala02] Timo Ojala, Matti Pietikainen, Topi Maenpaa, Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns, IEEE Trans. PAMI, July 2002 (Vol. 24, No. 7) pp. 971-987
[Park06] Min-Ho Park, Rae-Hong Park, and Sang Wook Lee Shot boundary detection using scale invariant feature matching. In: Proceedings of VCIP 2006.
[Pudil94] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters. 1994(15):1119-1125.
[Riesenhuber99] M. Riesenhuber and T. Poggio. Hierachical models of object recognition in cortex. Nature Neuroscience, 2(11): 1019-1025, 1999
[Sugiyama07] Masashi Sugiyama, Matthias Krauledat, Klaus-Robert Müller, Covariate Shift Adaptation by Importance Weighted Cross Validation; Journal of Machine Learning Research, 8(May):985--1005, 2007.
[Tsinghua06] Jie Cao, et al. Intelligent Multimedia Group of Tsinghua University at TRECVID 2006. In: Proceedings of TRECVID 2006 workshop.
[Torrabala04] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In Proc. CVPR, 2004.
[Wang07] D. Wang, X. Liu, L. Luo, J. Li, B. Zhang. Video Diver: Generic Video Indexing with Diverse Features, MIR workshop at ACM Multimedia, 2007.
[Yuan07] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, Bo Zhang. A Formal Study of Shot Boundary Detection. IEEE Trans. Circuits Syst. Video Techn. 17(2): 168-186. (2007)