The France Telecom Orange Labs (Beijing) Video Semantic Indexing Systems – TRECVID 2011 Notebook Paper

Yuan Dong[1, 2], Kun Tao[1], Xiaofu Chang[1], Shan Gao[2], Jiwei Zhang[2], Hongliang Bai[1], Wei Liu[1], Feng Zhao[1], Peng Li[1], Chengbin Zeng[1]

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

ABSTRACT

To accomplish the TRECVID2011 SIN task, some mild changes were made to our old systems. Considering the continuous growth of concept number and data amount, the number of visual features was reduced. Thus no more than 9 features were used in the final systems. Two early fusion runs and two late fusion runs were submitted. Although more time-consuming, the early fusion runs show better performances. Among them the L_A_FTRDBJ-SIN-1_1 run achieved our best MAP 0.051, which is based on 8 features and composite-kernel SVM. Our experiments also show that a proper selection of composite-kernel weights may be beneficial. In the comparison of two late fusion runs, the run with unified weights performs better.

1. INTRODUCTION

This year, the size of IACC data corpus became double, while the number of concepts increased even faster [1]. The problem of calculation efficiency has become extremely important, and some changes of research strategy must be made. For the second time, the labeling work didn't cover the whole development corpus. Over one hundred of the concepts were even discarded for the lack of positive samples. But there're still 346 concepts for a Full run. The unbalance of labeled data keeps making difficulties to all participants.

Orange Labs Beijing submitted 4 runs for the video semantic indexing (SIN) task. The first two runs are based on early fusion and the other two are based on late fusion. Limited by computing resources, only the last run is a "Full" run and the others are "Light". Some basic information of our runs is shown below:

- L_A_FTRDBJ-SIN-1_1: Composite Kernel of 8 features with equal weights. MAP = 0.051.
- L_A_FTRDBJ-SIN-2_2: Composite Kernel of 5 features with unequal weights. MAP = 0.051
- L_A_FTRDBJ-SIN-3_3: Late Fusion of 9 features. MAP = 0.043.
- F_A_FTRDBJ-SIN-4_4: Late Fusion of 9 features with unified weights. MAP_for_Light = 0.046. MAP_for_Full = 0.129.

Last year, there are 19 visual features were used in our systems [2]. Although we ever wished to get more information by using more features, it is really time-consuming. Considering the sharp increasing of concept number and data amount, we have to reduce the feature redundant significantly. The performance of each feature was tested separately, especially for some new features. Based on the testing results, 9 features were selected for late fusion, two sub-sets of 8 features and 5 features were used for early fusion.

For composite-kernel early fusion, a comparison between equal kernel weights and unequal weights was made. The comparison was only made on 5-feature models for the reason of time limitation, and the system using unequal kernel weights worked better. For late fusion, we kept using the 2-step fusion based on logistic regression. The same as last year, the run using unified weights worked better.

2. THE VISUAL FEATURES

This year, we tried some new features in our experiments, including Color Histogram (CLH), Dense Color Sift [3], MSER [4, 5] and Multi-resolution LBP (LBPmr) [6]. The Color Histogram is 360-D. It's a combination of 72 color bins and 5 segmented sub-regions. The Dense Color Sift and MSER are extracted by open source binary kits. Then as usual, they are transformed into BOW histograms. The vocabulary size of Dense Color Sift is 512, while the MSER's is 256.

We use a three-scale version of the $LBP_{8,R}$ to calculated the LBPmr features. Besides the original image, two low-pass filtered images are created and corresponding LBP operators are calculated. By concatenating the results of different resolutions, the final feature vector is built up, whose dimension is 3 times of the original one.

Besides above new features, the other features have been introduced in [2]. A total of 22 visual features were tested, and all results are shown in Fig. 1. “hist”, “hists” and “htg” means hard-assigned histogram, soft-assigned histogram and un-normalized histogram separately. The following numbers mean the dimensions of histograms. 70% labeled shots of development corpus were used for training, and 30% for testing. 50 concepts of Light corpus were tested and their MAPs were calculated.

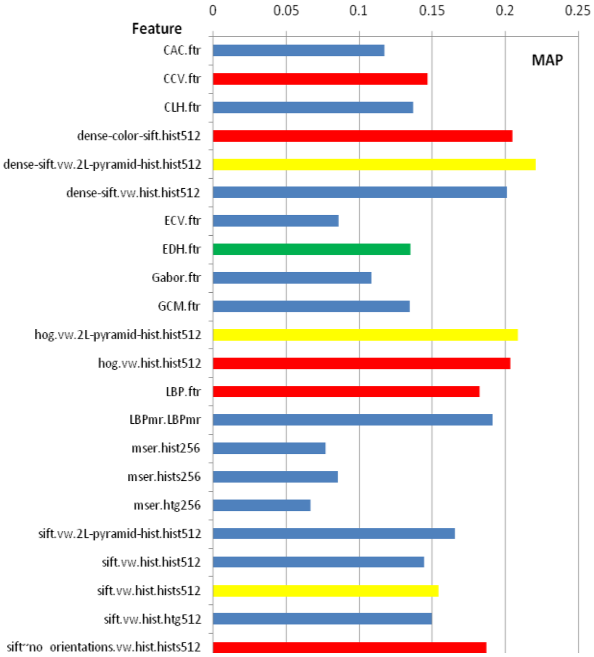


Fig. 1 Performances of Different Features

As shown in Fig. 1, Dense Color Sift shows slight advantage against Dense Sift, while LBPmr is also better than LBP. But the performance of MSER is far from expected. Based on observation, we found that there are too many key-frames in which no regions of interest can be found by the MSER detector. These key-frames include some transitional scenes, fine textures, fuzzy scenes et al. Their negative impacts on classification models caused the failure of MSER on TRECVID2011 dataset.

Considering the testing results, calculation costs and redundancy relationships, 5 features were selected as Group_1 (Red in Fig. 1):

- *CCV.ftr*
- *LBP.ftr*
- *hog.vw.hist.hist512*
- *dense-color-sift.hist512*
- *sift~no_orientations.vw.hist.hists512*

Then another 3 (Yellow in Fig. 1) were added to form up Group_2:

- *hog.vw.2L-pyramid-hist.hist512*
- *dense-sift.vw.2L-pyramid-hist.hist512*
- *sift.vw.hist.hists512*

The Group_1 and Group_2 were used for early fusion runs. Finally, 8 features in Group_2 and EDH.ftr were used for late fusion.

3. THE EARLY FUSION RUNS

In TREC10 evaluation, we only submitted one Light run based on early fusion, which archived the best MAP out of our 4 runs on 10 Light concepts. Although the training of composite-kernel SVM is time-consuming, many other researches made by us have all proved that it really performs better than late fusion models. Thus we paid more attention and computing resources for this method.

In 2009, we tried to use “multiple kernel learning” to learn the kernel weights, but it even did not achieve better performance than using equal weights [7]. This might be related to limited positive samples for some concepts and large intra-concept variability. This year the testing MAP of each feature is used to calculate the weights. Due to the time limitation, the unequal weights were only tested on Group_1 early fusion. The testing results show that it is better than using equal weights. The comparison results are shown in Table I.

Table I. Early Fusion Results

Features	Kernel Weights	Testing MAP	Submitted System MAP
5 of Group_1	Equal	0.248	-
5 of Group_1	Unequal	0.250	0.051
8 of Group_2	Equal	0.262	0.051

The total dimension of Group_1 features is 3176, while that of Group_2 is 8808. The corresponding computing costs are several times different. But the difference between their testing MAPs is small. Considering the factor of generalization, their performances in submitted runs are even closer.

Although the total dimension of Group_2 is nearly 3 times of Group_1’s, it brings little improvement in MAP. Besides the reason of information redundancy, another factor should be considered: There’re two 2-level pyramid histogram features in Group_2. In our systems, the pyramid features is regarded as a combination of 5 separate histograms of different sub-regions. Thus we use composite-kernel SVM to train their corresponding models. Then in an early fusion system based on composite-kernel methods, there're two levels of composite-kernels in fact. Such complex structure will sometimes reduce the generalization ability of the models.

Anyway the existing experimental results mean that we can be more proactive to make a compromise between indexing precisions and calculation costs in real world applications. Such kind of change is also being encouraged by the organizers of TRECVID.

4. THE LATE FUSION RUNS

Although the testing performances are not as good as early fusion methods, late fusion methods have their own advantages such as low computing cost and combination flexibility. We submitted two late fusion runs for comparison. 9 features are put into 3 groups: CCV, EDH and LBP belong to "cel3" group, while 2 hog features and 4 sift features belong to "hog2" and "sift4" respectively. The intra-group fusion is equal weighted, and the inter-group fusion is based on logistic regression (Fig. 2).

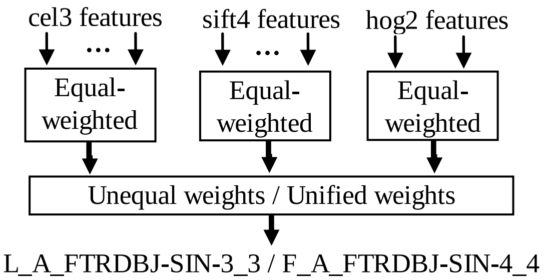


Fig. 2 2-step fusion structure

Last year, we used a group of unified weights trained on a small concept corpus to fusion the scores of 130 concepts, it showed surprisingly good generalization ability. In this year's evaluation, it kept working well.

First, the fusion weights were calculated for each of the 50 concepts of Light corpus. Using these weights the shot scores were combined into the results of our 3rd submitted run. Then the average weight of each feature was calculated. By re-normalizing the average weights, a group of unified weights could be got to accomplish the fusion of 346 concepts.

Although the 4th submitted run using unified weights is a Full run, its MAP on Light run concepts can be also seen (0.046), which is better than using different weights (0.043).

5. CONCLUSION

Although there is no amazing new element being added into our systems, the adjustment of our research strategy brought some encouraging results. Giving up superimposing more and more features, we are trying to find the balance between indexing precisions and computing costs. For one more time the composite-kernel early fusion was proved to be good, it will be widely used in our future works. At the same time, the unified-weight late fusion can also be a powerful complement in different applications. The testing results of new features are also very valuable for other research works. In the future, we wish to study the redundancy of existing features. By making more comparisons of different feature combinations on different datasets, some optimized combinations should be fixed and bring more benefits to our researches.

6. REFERENCES

[1] "Guidelines for the TRECVID 2011," http://www-nlpir.nist.gov/projects/tv2011/tv2011.html.

[2] K. Tao, etc. "The France Telecom Orange Labs (Beijing) Video Semantic Indexing Systems – TRECVID 2010," http://www-nlpir.nist.gov/projects/tvpubs/ tv.pubs.org.html, 2009.

[3] http://koen.me/research/colordescriptors/

[4] http://www.robots.ox.ac.uk/~vgg/research/affine/index.html

[5] J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. of British Machine Vision Conference, pages 384-396, 2002.

[6] T. Maenpaa, "The local binary pattern approach to texture analysis - extensions and applications," University of Oulu, 2003, PhD Dissertation.

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

FRANCE TELECOM ORANGE LABS (BEIJING) AT TRECVID 2011: CONTENT-BASED COPY DETECTION

Hongliang Bai[†]*,Yuan Dong*[‡]*, Wei Liu*[†]*, Lezi Wang*[‡]*, Chong Huang*[‡]*, Kun Tao*[†]

[†]France Telecom Research & Development - Beijing, 100190, P.R.China
[‡]Beijing University of Posts and Telecommunications,100876, P.R.China
{hongliang.bai,wei.liu,kun.tao}@orange.com
{yuandong,wanglezi,huangchong661100}@bupt.edu.cn

ABSTRACT

In this paper , the Content-based Copy Detection (CCD) system is described by France Telecom Orange Lab (Beijing). It is the first time that we participate in the CCD evaluation task. So we focus on building the initial benchmark and present some new ideas to implement the task. Four runs are submitted for the task, namely:

FT.m.NOFA.orange1: SIFT feature, feature selection, inverted table-based indexing and grouping, SVM-based result verification, fusion of energy difference feature and CEPS-like feature, flexible hashing-based searching, the fusion of audio and video querying results are introduced.

FT.m.NOFA.orange2AudioOnly: Fusion of energy difference feature and CEPS-like feature, flexible hashing-based searching.

FT.m.balanced.orange3: More loose parameter tuning is conducted compared with the **FT.m.NOFA.orange1** run.

FT.m.balanced.orange4VideoOnly: SIFT feature, feature selection, inverted table-based indexing and grouping, SVM-based result verification.

After experiments, in the NOFA profile, the **FT.m.NOFA.orange2AudioOnly** run has better performance than the **FT.m.NOFA.orange1** one. And in the balanced profile, the **FT.m.balanced.orange3** run is better than the **FT.m.balanced.orange4VideoOnly** one. In the NDCR metric of balanced profile, **FT.m.balanced.orange3** can have leading ranks.

Index Terms— TRECVID, SIFT, CEPS-like, Energy Difference Feature, Feature Selection, Inverted Table-based Matching, Content-based Copy Detection, SVM.

1. INTRODUCTION

With the growth of images and videos on the internet, the retrieving requirement from users has increased enormously. They can record videos or take photos by the mobile phones, video camcorders, or directly download from the video webs, and then distribute them with some modifications. More than 13 million hours of video were uploaded during 2010 and 35 hours of video are uploaded every minute, and YouTube reached over 700 billion playbacks in 2010 [1]. Among these huge volumes of images and videos, the large number of them are duplicate or near duplicate.

Based on a sample of 24 popular queries from YouTube, Google Video and Yahoo! Video, on average there are 27% redundant videos which are duplicate or nearly duplicate to the most popular version of a video in the search results [2]. Nearly 30% videos are duplicated in one-day Orangesport videos[1]. Users always feel frustrated when they find what they are interested,seeing many duplicate sequence. So the copy detection is one of very important techniques to retrieve and delete the videos. It also can reduce the large disk storage for the video website.

The copy detection is to find the corresponding copy sequences of one query from the video database, and the query may have different audio and video transformations. The video and audio information can be used to implement the copy detection. The video-based querying has more information and distinguish the copy from the reference database. The audio-based methods can well solve the difficulty, especially when the audio information is consistent with the variable video frames. Usually, the framework is composed by preprocessing, feature extraction, querying methods and fusion postprocessing.

The video feature can be classified into the global feature and local feature. The global features are generated from the whole gray or color frames, such as color histogram [3], DCT coefficient [4], and binary spatiotemporal feature [5]. The local features are extracted from the local points or regions, so it is more robust to complex background, occlusion, scaling, rotation. The local feature extraction basically has two steps [6]; one is feature detectors, such as Harris detector, Harris Laplace detector, Hessian Laplace, Harris/Hessian Affine detector, and the other is feature descriptors, such as Scale Invariant Feature Transformation (SITF) [7], Shape Context [8], Gradient Location and Orientation Histogram [9], Speeded Up Robust Features [10], DAISY [11]. In the TRECVID C-

[1] http://sports.orange.fr/

CD task, the querying videos have been generated by the complex transformations in the video and audio channels. The local features are mainly selected in the querying system.

For the audio feature extraction, a Weighted Audio Spectrum Flatness (WASF) is presented to extend the MPEG-7 descriptor-ASF by introducing human auditory system functions to weight audio data [12]. The feature is robust to several audio transformations, but tuning the parameters is one hard work. The HAAR filters are influenced by the training data [13]. Mel-Frequency Cepstral Coefficients (MFCC) is a feature used in the speech recognition and copy detection [14]. Energy Differences Feature (EDF) is widely used in [15, 16, 17], and the good performance is achieved in the large-scale video database. However, EDF only considers one scale property of the frequency.

For the video retrieving, the hash function [18] is used for the accurate searching with the high efficiency. But, the hash-based searching can not deal with the near duplicate video querying clips. Locality-Sensitive Hashing (LSH) [19] is not suitable in the low-dimensional video feature space.

The remainders of the paper are organized as follows. The system framework is described in Section 2. The detail video-based querying is proposed in Section 3. Section 4 presents the audio-based querying. The audio and video fusion is introduced in Section 5. Section 6 shows some experimental results. Finally, the conclusions and future works are listed.

2. SYSTEM OVERVIEW

Both the video and audio based copy detection system include two parts: reference database generation and query data retrieval, shown in Fig. 1.

The database generation for the video-based detection is described as follows. Uniform sampling method is applied to extract keyframes from the reference videos. SIFT features are used to describe the local keypoints of keyframes. Each SIFT descriptor is assigned to the visual word in the minimum Euclidean distance. Specific information of keypoints are hashed into an inverted table database, and keyed by the visual word ID. The inverted indexing is adopted for the efficient querying.

For a given query video, keyframes are extracted based on uniform sampling rate. Then, query videos go through the SIFT feature extraction, the visual word assignment and the invert indexing. The matching feature pairs of two frames are set into different groups based on the difference the frame ID. Hierarchical filtering based on the Hough transform is adopted to spatially verify the feature pairs and return the matching score of two frames. Scores of two frames within the same group are added to calculate the final group score. The group with the maximum score is selected. And the Support Vector Machine (SVM) is adopted to refine the video based copy detection results.

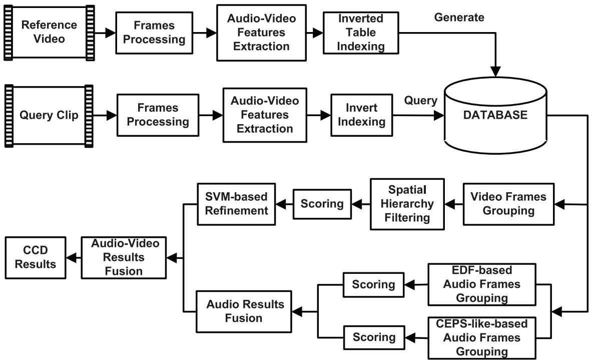


Fig. 1. TRECVID CCD system overview

The two audio fingerprints, EDF and CEPs-like, are extracted from the reference and query audio streams of 44100HZ. The reference fingerprints are stored by the inverted indexing table. Querying audio features hash the matches from the database and the group with the maximum score is selected as candidates. Then, the results generated after two features are fused.

Finally, the CCD querying output is generated by the fusion of video and audio based results.

3. VIDEO-BASED COPY DETECTION

3.1. Frame Processing

The first step is to extract keyframes from the video clips. The even sampling algorithm is used: two keyframes per second for query videos and four keyframes for references. The experience is that the system gives better detection performance in the higher sampling rate, but the more computing time is consumed.

3.2. Feature Extraction

The local features are robust to the camera coding, insertions of pattern, change of gamma, picture-in-picture transformation [20] and their combinations. The 128D SIFT features vectors are adopted to describe the local keypoints of the keyframes. A vocabulary of 50000 visual words are generated by clustering the SIFT descriptors of training images beforehand by the k-means clustering algorithm. Each descriptor of TRECVID2011 video data is mapped to one of the visual words in the minimum Euclidean distance. The keypoints' frame ID, video ID, position, scale, orientation and visual word ID are kept for the invert indexing and hierarchy filtering.

3.3. Indexing and Grouping

The reference video database is formed by hashing information of keypoints into an inverted table, keyed by the visual word ID. Each entry of table includes the reference keypoints'

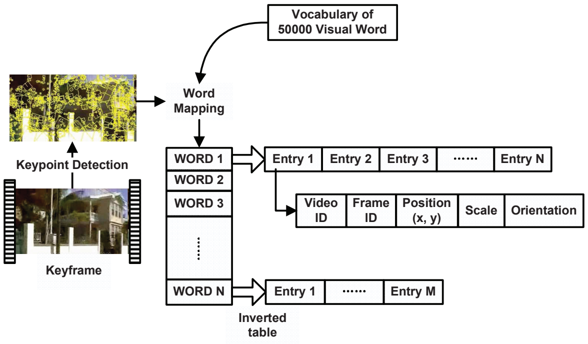


Fig. 2. Inverted table

video ID, frame ID, keypoints position, scale and orientation in Fig.2. And the memory usage of each entry is 11 bytes. A given query keypoint finds the matches in database by direct hashing with time complexity of $O(1)$. The matching pairs between two frames are grouped based on difference of frame IDs. As shown in Fig.3, a pair of frames with the matching kepoint pairs is one element of a group. A pair of frames is regarded as the matched if the similarity score is high enough after the hierarchy filtering. The sum of the similarity score of frame pairs within the same group is defined as the group score.

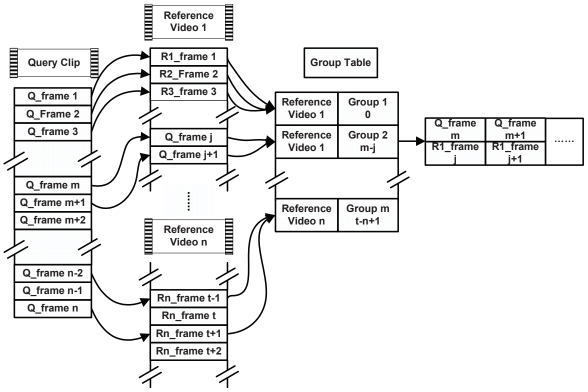


Fig. 3. Inverted table-based frame grouping

3.4. Hierarchical Filtering Based on Hough Transform

An assumption is that the most pairs between two frames satisfy the specific spatial consistency of correct matched pairs. Similar to the Hough transform, the hierarchical filtering algorithm is introduced to select the pairs of the largest number with the same predefined spatial parameters. Hence, the error matching pairs between two frames, generated by the direct indexing based on visual words, are filtered out. The hierarchical filtering contains two parts: consistency of the keypoint scale and orientation, and consistency of geometric transformation.

The pairs' consistencies of scale and orientation are measured by the difference of their scales and orientations, respectively. Based on the predefined assumption, the most consistent pairs between two keyframes are reserved for the next geometric verification.

The geometric verification is a voting strategy on a set of affine parameters $\begin{pmatrix} a1 & b1 \\ a2 & b2 \end{pmatrix}$ generated by solving the transform equations of two matching pairs, which is reserved by the scale and orientation consistency verification. The affine transform maps a query keypoint $Q(x_{qi}, y_{qi})$ to the matched reference point $R(x_{ri}, y_{ri})$, where x, y indicates the position of a keypoint located in the image. The transform is modeled as:

$$\begin{bmatrix} x_{qi} \\ y_{qi} \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{ri} \\ y_{ri} \\ 1 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ 1 \end{bmatrix} \quad (1)$$

where a_1, a_2, b_1, b_2 can be solved by substituting position values of two matching pairs. The total C_N^2 affine parameter groups are generated for N matching keypoint pairs.

For a given pair of frames, $id\langle \delta a_1, \delta a_2, \delta b_1, \delta b_2 \rangle$ represents one geometric transformation parameter obtained by combination of affine parameters, where $\delta a_1, \delta a_2, \delta b_1, \delta b_2$ are quantized affine parameters of a_1, a_2, b_1, b_2 respectively. And id_{max} represents the dominant transformation parameter where the number of inliers is the maximum. Only matching descriptor pairs are reserved if the corresponding transform parameters are equal to id_{max}. It should be time consuming to solve the affine parameters exhaustively. However, only a small parts are used to solve the Equation 1, and generally vary from 10 to 200, which depend on the image complexity. The initial system adopted the RANSAC-based algorithm. It is more consuming than the Hough Transform-based algorithm when the number of test data is small.

3.5. Scoring

Within a group, the similarity between query frame I_j and reference image I_i is defined as $S = M_{ij}/\sqrt{N_j}$, where M_{ij} denotes the number of matching pairs after the hierarchical filtering and N_j is the number of keypoints in the query frame. The two frames are matched if the similarity score S is larger than a specific threshold. One match frame pair generates a hypothesis $(rid, \delta t, S_f)$, where $\delta t = t_q - t_r$, t_q and t_r indicate the time stamps (frame ID) of query and reference frame respectively, and rid is the reference video index in the database. Hence, the query clip can be aligned to the reference video if a single parameter (temporal offset) has been determined. In this case, the frames are grouped via parameters δt and r. The aligned group score is defined as $S_g = \sum_{i=0}^{n} S_{fi}$, where n is the number of match frame pairs in one group and S_{fi} is the frame similarity score. The segment of reference video with the best group score is selected as the candidate video-based CCD output.

3.6. SVM-based Result Verification

We observed that some correct detections are missed by selection with the absolute high threshold and there exists both correct pairs and errors in the interval between high and low thresholds. This step aims to verify the detection results based on the machine learning algorithm. SVM is used to check the above candidate detection results. Ten dimensions histogram vectors are generated by counting the corresponding matching frames which have the specific numbers of matching pairs after hierarchy spatial verification and the ratio that the pairs taken in the total keypoints, shown in Fig. 4. The detection results of TRECVID 2010 video data are used as training data: correct and wrong detection results as the positive and negative samples, respectively. Lib-SVM with RBF kernel is used to trained the model and classify the results generated by the TRECVID 2011 data. We only focus on the positive video pairs which are not detected by previous scoring strategy. After experiments, some missed pairs are recalled.

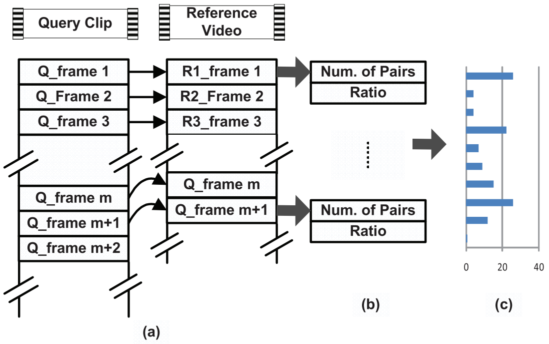


Fig. 4. (a) An example of the candidate matching video clip pair. (b) The number of matching keypoint pairs and the ration that the pairs taken in the total keypoints generated by each frame pair. (c) The histogram representation: first 5 dimensions indicates the number of keypoint pairs and the rest indicates the ratio.

4. AUDIO-BASED COPY DETECTION

The audio-based copy detection system framework is introduced in Fig. 5. Firstly, the querying audio signal is separated from the videos. Then the audio signal is processed by the Butterworth and Hamming window filtering. After the Fast Fourier Transform (FFT) analysis, the 17 sub frequent bands are selected in the mel-frequency space. 16-bit EDF and 16-bit CEPS-like feature are extracted respectively. The two types of features are used to query in the reference database. The different searching results from the above features are fusioned finally.

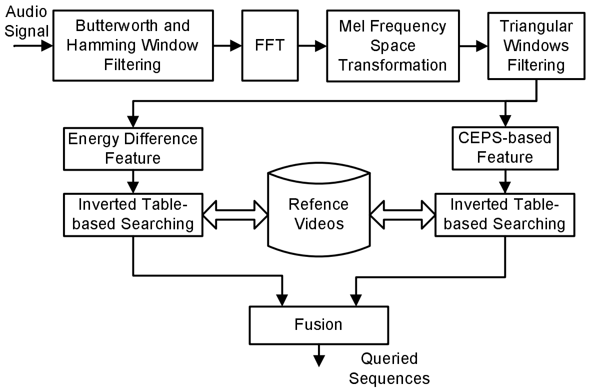

Audio Signal
Butterworth and Hamming Window Filtering
FFT
Mel Frequency Space Transformation
Triangular Windows Filtering
Energy Difference Feature
CEPS-based Feature
Inverted Table-based Searching
Refence Videos
Inverted Table-based Searching
Fusion
Queried Sequences

Fig. 5. Audio-based copy detection system framework

4.1. Feature Extraction

4.1.1. Butterworth and Hamming Window Filtering

The sampling rates of the internet videos vary in a large range. The first step is to normalize the sampling rates into a constant value F_N, here F_N is set 44100 Hz. Then the normalized signals S are lowpass filtered to 4000 Hz by a Butterworth filter. The magnitude-squared response of a N-order analog lowpass Butterworth filter is $|\mathbf{H}(j\Omega)|^2 = 1/(1+(\Omega/\Omega_c)^{2N})$, where the cutoff frequency Ω_c is 3dB. The top 100 coefficients is used to convolve with S in the time domain.

Then the hamming window filtering is applied to every frame in order to keep the continuity of the first and the last points in the frame before FFT. The hamming window filtering is $\mathbf{H}(i) = 0.54 - 0.46 * \cos(2\pi i/(N-1))$, where N is the sample number in each frame and set 2048. The inter overlapping is 1024 samples ($23.2ms$).

4.1.2. FFT and Mel-frequency Space Transformation

After the Hamming window filtering, the 1-D audio signals are transformed into 2-D spectrograms by FFT. The spectrum between 300 Hz and 4000 Hz is equally divided into 17 sub bands in the mel-frequency space. The mel-frequency can reflect similar effects in the human's subjective aural perception. The relation of the mel-frequency and natural frequency is $Mel(f) = 2595 * \log(f/700+1)$, where f is the natural frequency.

4.1.3. Energy Difference Feature

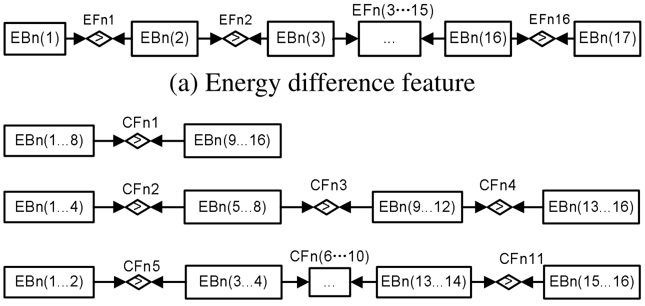
A triangular filtering is used to compute the energy of each sub band in the magnitude frequency response. The number of the filters is equal to that of the sub bands. The coefficients of the filter are defined by

$$w(n) = \begin{cases} \frac{2n}{N-1} & n = 0, 1, ..., \frac{N-1}{2} \\ 2 - \frac{2n}{N-1} & n = \frac{N-1}{2}, ..., N-1 \end{cases} \qquad (2)$$

EDF features between the sub-bands are used to generate the fingerprint of each frame, which are calculated by Equation 3.

$$EF_n(m) = \begin{cases} 1 & EB_n(m) > EB_n(m+1) \\ 0 & otherwise \end{cases} \qquad (3)$$

where $EB_n(m)$ represents the energy value of the n-th frame at the m-th sub-band, and $m \in [1 \cdots 16]$. The 15-bit and 32-bit fingerprints are used in [15, 16] respectively. After considering the storage size of $short\ int$ and robustness of the searching algorithm, the 16-bit fingerprint $EF_n(m)$ is selected. The feature is demonstrated in the Fig. 6(a).



(a) Energy difference feature

(b) CEPS-like feature

Fig. 6. Extraction of two types of audio features, which describe the energy property of the different scales

4.1.4. CEPS-like Feature

The cepstrum denotes the rate of the change in the different spectrum bands and the result of taking the Fourier Transform (FT) of the log spectrum. The EDF feature only considers the energy difference in the low level. The CEPS-like feature is proposed to combine the multi-scale energies into one feature.

In Fig.6(b), $CF_n(1)$ is the highest-scale feature, which used all information of 16 sub bands. $CF_n(2 \cdots 4)$ are in the second level and the difference of four adjacent sub bands. $CF_n(5 \cdots 11)$ are in the third level. $CF_n(12 \cdots 16)$ are the same with $EF_n(1)$,$EF_n(4)$,$EF_n(7)$,$EF_n(10)$ and $EF_n(13)$ respectively. $EB_n(m_1 \cdots m_2)$ is the energy sum from the m_1-th sub band to the m_2-th sub band.

4.2. Flexible Hash-based Searching

The hash-based searching is a very important and widely used technology. The searching performance is improved by two aspects: (1)one-bit modification in the hash matching. If the hamming distant of a querying and reference feature is one, they are regarded as a matching pair; (2)matching time can tolerate some time errors because of the frame losing or noise interference. The above algorithms can improve the searching flexibility.

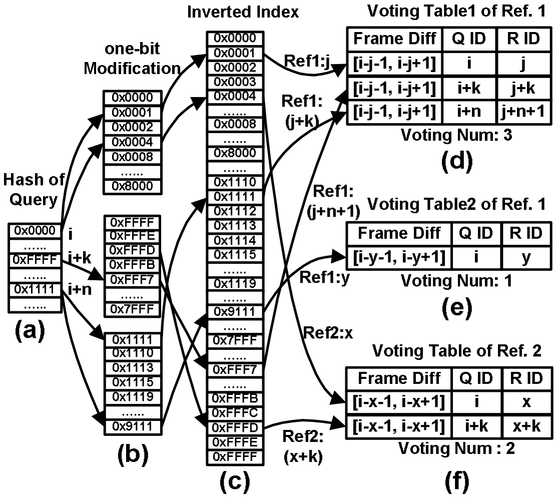


Fig. 7. Flexible hash-based searching with one-bit modification and the time redundance

The audio fingerprint matching strategy is shown in Fig. 7. Fig. 7(a) shows the sequence of hash values in a querying clip. And Fig. 7(b) can make the hash matching more robust by modification of any one bit of a hash value. Seventeen different values are generated for a 16-bit feature. These modified hash values are matched with the ones from reference data in the inverted index table, shown in Fig. 7(c). The voting tables are used in references, which is related to the matched hash values from the inverted table. The voting number are the hitting value in some time difference between indexes of the matched reference and querying. The voting strategy is illustrated in Fig. 7(d)(e)(f). The largest voting result N_{vote}(Voting Number 3) occurs in Fig. 7(d). The time duration of the queried sequence is $[j, j+n+1]$ in the reference database.

$$N_{vote} \triangleq \arg\max_{\tau} \sum_{r,q \in N} \delta(\tau - |r-q|) \qquad (4)$$

where r and q are the time indexes of the matching sequence of the querying and reference. If N_{vote} is greater than the predefined threshold T, the queried reference sequences will be regarded as the querying results.

4.3. Result-based Fusion from Different Features

The fusion algorithm can be used in the stages of the feature extraction or searching results. The fusion of the searching results are proposed from the different features, shown in Fig. 8. For the retrieving results from every feature, the higher precision is generated if the threshold T is set with higher values. In Fig. 8, G_1 and G_2 are the above reliable querying results, and G_3 is the logical “AND” operation results from ED-

F and CEPS-like features. Both the advantages of EDF and CEPS-like are taken in the G_3. The querying results are more reliable if the outputs of above two features are same. The final results are the logical "OR" of G_1, G_2 and G_3. The parameters TH_1 and TH_2 will be discussed in the experimental section.

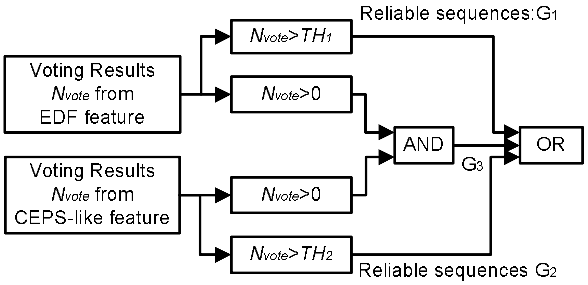


Fig. 8. Fusion of the searching results from EDF and CEPS-like features

5. FUSION OF VIDEO AND AUDIO BASED DETECTION RESULTS

The final querying results are generated by the fusion of audio and video detection results. There are five conflict cases for a specific the query data and the fusion strategies are different: the video-based results are regarded as submissions when the video-based algorithm gives the specific results but audio does not, and vice versa; the video-based results are regarded as submissions if the detected video IDs are the same; the audio-based results are the submissions when the detected video IDs are different; there is no submissions if neither audio nor video based give the results, shown in Fig. 9.

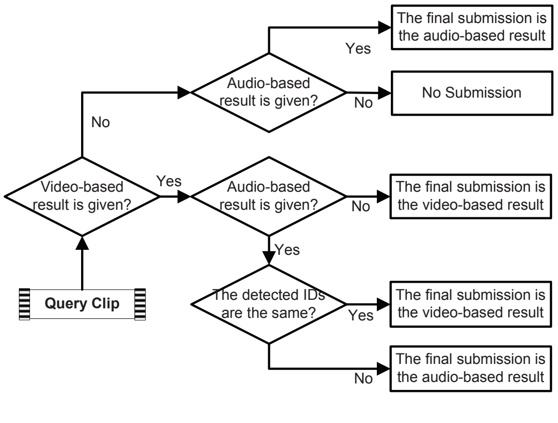

The final submission is the audio-based result
Yes
Audio-based result is given?
No
No Submission
No
Video-based result is given?
Yes
Audio-based result is given?
No
The final submission is the video-based result
Yes
Query Clip
The detected IDs are the same?
Yes
The final submission is the video-based result
No
The final submission is the audio-based result

Fig. 9. Fusion strategy of audio and video based results

6. EXPERIMENTS

In this section, the experiments are conducted to demonstrate the effectiveness of the proposed method.

6.1. Database Description

The reference data is identical to 400 hours and 12000 files in the TRECVID [21] 2011 test and training data. Each query has 8-type video and 7-type audio transformations.

In the audio-related task, the original audio clips are transformed into the following seven types, namely, **TA1.** do "nothing"; **TA2.** mp3 compression; **TA3.** mp3 compression and multiband companding; **TA4.** bandwidth limit and single-band companding; **TA5.** mix with speech; **TA6.** mix with speech, then multiband compress; **TA7.** bandpass filter, mix with speech, compress.

In the video-related task, the original video clips are transformed into the following seven types, namely, **TV1.** Simulated camcording; **TV2**. Picture in picture; **TV3.** Insertions of pattern; **TV4.** Compression; **TV5.** Change of gamma; **TV6.** Decrease in quality; **TV8.** Post production; **TV10.** change to randomly choose 1 transformation from each of the 3 main categories.

The final audio+video queries will be various transformation combinations of the audio+video querying, denoted as $T1, T2, \cdots, T70$.

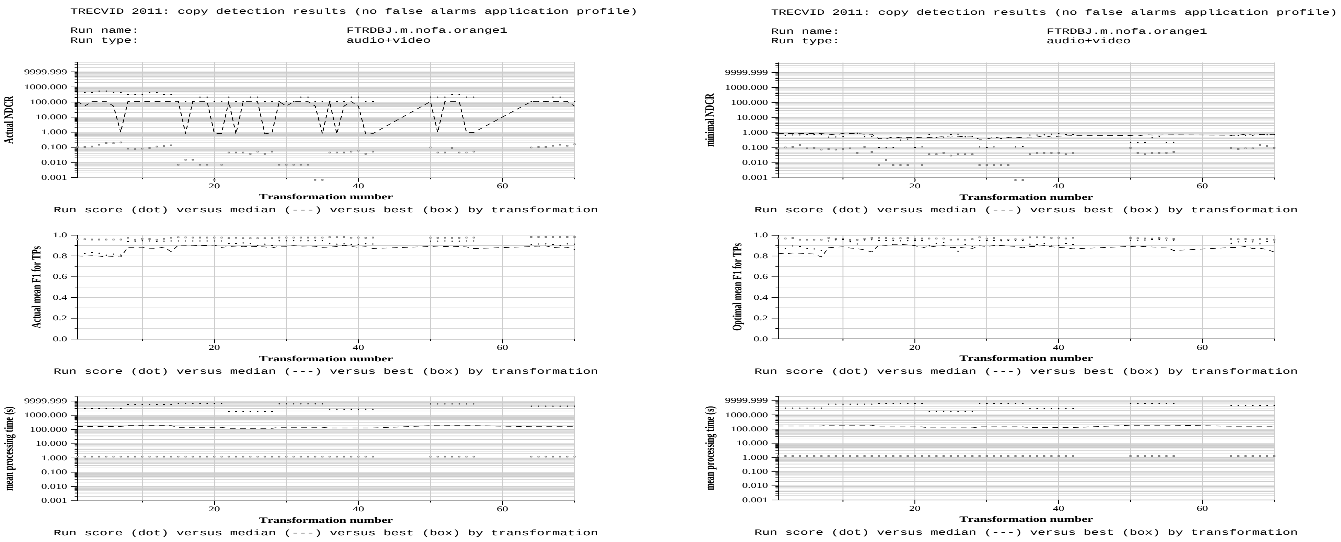
Many evaluation metrics are used in the CCD task. The Actual Normalized Detection Cost Rate (NDCR) and F1-Measure are selected in following.

6.2. System Performance

6.2.1. NOFA Profile

In **FT.m.NOFA.orange1** run , the actual NDCR is worse than the median level in all the submitted runs, shown in Fig. 10(a). The values varies in $106.883 \sim 534.225$. The reason is that many false alarms are generated in the matching algorithm. The F1-Measure performance is higher than the median level, which changed in $0.808 \sim 0.946$. After the parameter tuning, the optimal NDCR is shown in Fig. 10(b), and the NDCR and F1-Measure are improved to $0.097 \sim 0.925$ and $0.847 \sim 0.959$ respectively.

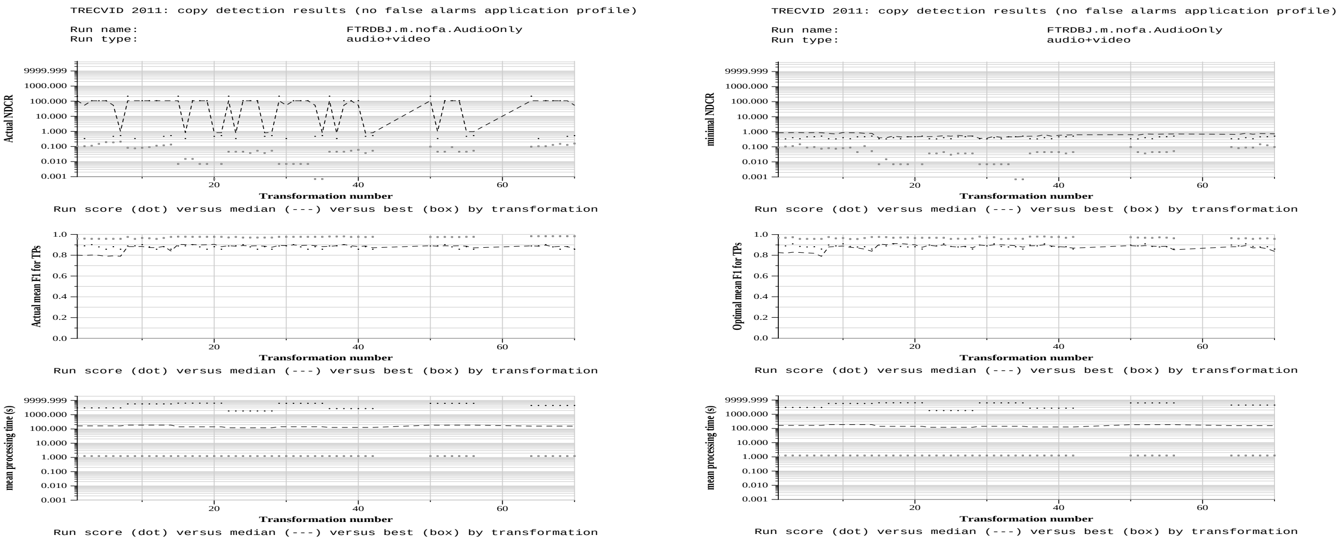
In **FT.m.NOFA.orange2AudioOnly** run , the actual NDCR is also worse than the median level in all the submitted runs, shown in Fig. 11(a), but is better than our **FT.m.NOFA.orange1** run. The values are $0.336 \sim 213.905$. The F1 performance is higher than the median level and changed in $0.858 \sim 0.902$, but the F1-Measure is worst than our **FT.m.NOFA.orange1** run because the audio feature is simple and not robust in estimating the time boundary. After the optimization, the optimal NDCR and F1-Measure are improved to $0.336 \sim 0.515$ and $0.860 \sim 0.910$ respectively, shown in Fig. 11(b). From **FT.m.NOFA.orange1** and **FT.m.NOFA.orange2AudioOnly** runs, the F1-Measure performances change little after the parameter optimization.



(a)Actual audio-video performances in the NOFA profile

(b)Optimal audio-video performances in the NOFA profile

Fig. 10. Performances of the **FT.m.NOFA.orange1** run in the NOFA profile



(a)Actual audio-only performances in the NOFA profile

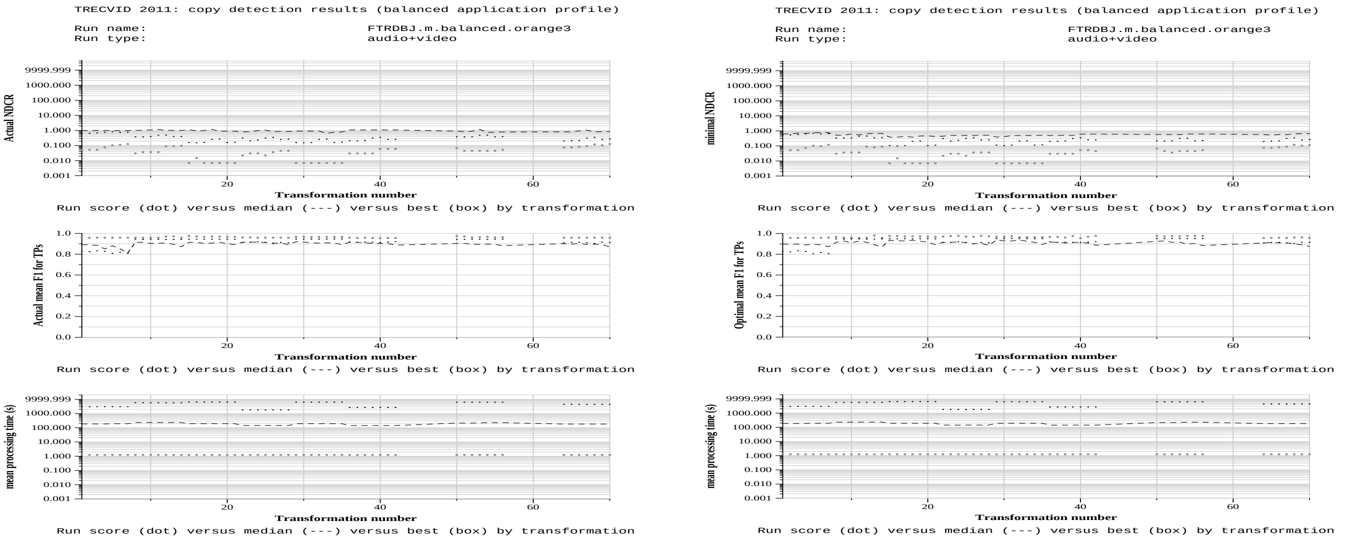
(b)Optimal audio-only performances in the NOFA profile

Fig. 11. Performances of the **FT.m.NOFA.orange2AudioOnly** run in the NOFA profile

6.2.2. Balanced Profile

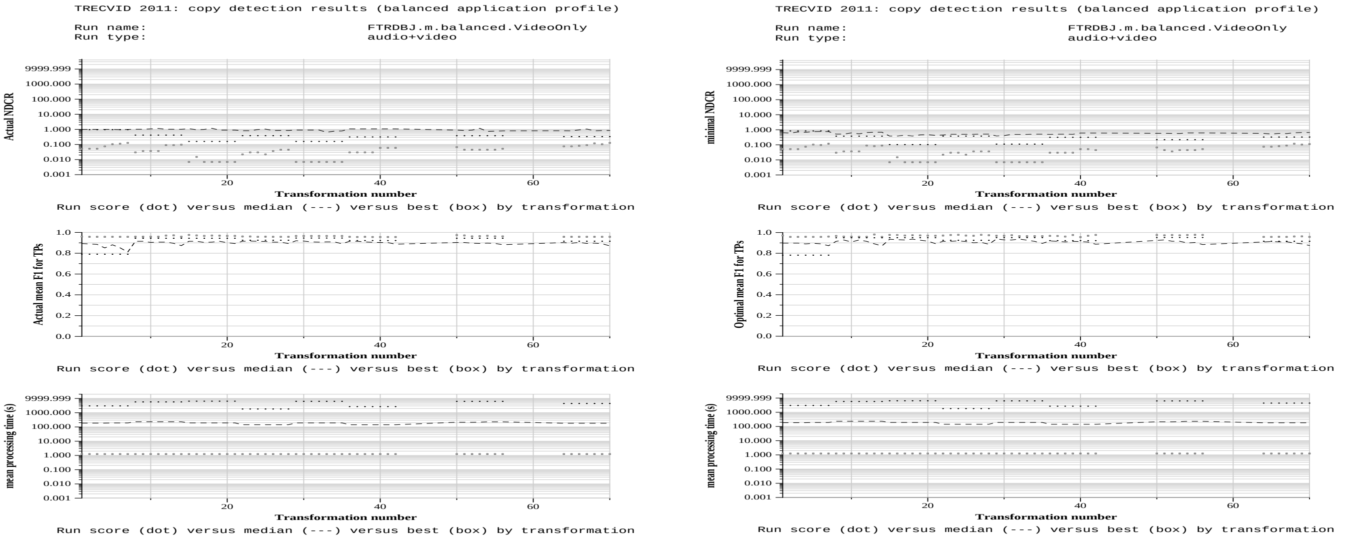
In **FT.m.balanced.orange3** run , the actual NDCR is better than the median level in the all submitted runs, shown in Fig. 12(a). The values are in 0.152∼0.818. The worst and best results occur in the **TV1** and **TV5** respectively. The F1-Measure performance is higher than the median level, and has 0.808∼0.946. After optimization, the optimal NDCR and F1-Measure are improved to 0.097∼0.701 and 0.805∼0.952 respectively, shown in Fig. 12(b).

In **FT.m.balanced.orange4VideoOnly** run, the actual NDCR is also better than the median level and has 0.159∼0.957, shown in Fig. 13(a). The NDCR of the run is worst than **FT.m.balanced.orange3**, and we can conclude the audio feature can make up the disadvantages of the video features. The F1-Measure is higher than the median level with 0.791∼0.946. After some parameter tuning, the optimal NDCR and F1-Measure are improved to 0.104∼0.840 and 0.782∼0.952 in Fig. 13(b).



(a)Actual audio-video performances in the balanced profile (b)Optimal audio-video performances in the balanced profile

Fig. 12. Performances of the **FT.m.balanced.orange3** run in the balanced profile



(a)Actual video-only performances in the balanced profile (b)Optimal video-only performances in the balanced profile

Fig. 13. Performances of the **FT.m.balanced.orange4VideoOnly** run in the balanced profile

6.3. Conflict Strategy in the Audio and Video Fusion

In TRECVID CCD 2011 submission, the video-based querying result is selected as the final one when the reference ID is same from the video and audio-based querying results. It can not deal with the conflict case when the starting and ending time in the reference video has little superposition from the audio and video-based querying results. No problems occur in the TRECVID CCD 2010 database. But in TRECVID C-CD 2011 database, the NDCR is negatively influenced. If the largest spanning time of the audio-video querying results is used as the new conflict strategy, the NDCR can be further improved. In the Fig. 14, the NDCR of T_1 is improved from 0.636 to 0.293, which nearly the best performance of all the submitted runs.

7. CONCLUSIONS AND FUTURE WORKS

In the TRECVID 2011, we submitted the CCD system for the first time. The audio-video fusion algorithm is presented in this paper. Some new ideas is introduced in the feature extraction, feature selection and matching. After the submission,

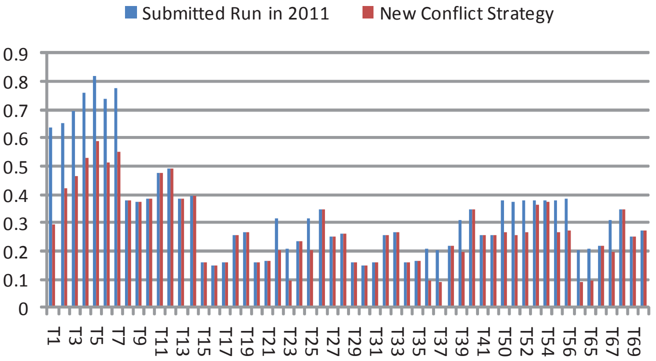


Fig. 14. Conflict strategy influence

the conflict strategy is very important problem. The SVM learning algorithm is proposed to solve the single threshold problem. In the future, the learning algorithm will be deeply discussed and used in the fusion of the video and audio-based query results.

8. REFERENCES

[1] http://www.youtube.com/t/press_statistics, ,” .

[2] X. Wu, C. Ngo, A. G. Hauptmann, and H. Tan, “Real-time near-duplicate elimination for web video search with content and context,” *IEEE Tran. on Multimedia*, vol. 11, no. 2, pp. 196–207, 2009.

[3] X. Yang, Q. Tian, and P. Xue, “Efficient short video repeat identification with application to news video structure analysis,” *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 600–609, 2007.

[4] X. Naturel and P. Gros, “A fast shot matching strategy for detecting duplicate sequences in a television stream,” in *CVDB*, 2005, pp. 21–27.

[5] L. Shang, L. Yang, F. Wang, K. Chan, and X. Hua, “Real-time large scale near-duplicate web video retrieval,” in *ACM MM*, 2010.

[6] S. Maji, “A comparison of feature descriptors,” .

[7] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[8] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *PAMI*, vol. 24, pp. 509–522, 2001.

[9] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Surf: Speeded up robust features,” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] E. Tola, V. Lepetit, and P. Fua, “Daisy: an Efficient Dense Descriptor Applied to Wide Baseline Stereo,” May 2010, vol. 32, pp. 815–830.

[12] J. Chen and T. Huang, “A robust feature extraction algorithm for audio fingerprinting,” in *Pacific Rim Conference on Multimedia(PCM)*, 2008, pp. 887–890.

[13] Y. Ke, D. Hoiem, and R. Sukthankar, “Computer vision for music identification: Video demonstration,” in *CVPR*, 2005.

[14] H. Jegou, Ma. Douze, G. Gravier, C. Schmid, and P. Gros, “Inria lear-texmex: Video copy detection task,” 2010.

[15] M. Heritier, V. Gupta, L. Gagnon, and P. Cardinal, “Crim’s content-based copy detection system for trecvid,” in *Content-Based Multimedia Indexing (CBMI)*, 2010.

[16] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in *Music Information Retrieval(ISMIR)*, 2002.

[17] J. H. Ton and T. Kalker, “Robust audio hashing for content identification,” in *Content-Based Multimedia Indexing(CBMI)*, 2001.

[18] I. Döhring and R. Lienhart, “Mining tv broadcasts for recurring video sequences,” in *Conference on Image and Video Retrieval(CIVR)*, 2009, pp. 1–8.

[19] P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” in *symposium on Theory of computing(STOC)*, 1998, pp. 604–613.

[20] L. Wang, Y. Dong, H. Bai, W. Liu, and K. Tao, “A word-based approach for duplicate picture in picture sequence detection,” in *Int. Conf. Broadband Network & Multimedia Technology*, 2011.

[21] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *Multimedia Information Retrieval(MIR)*, 2006, pp. 321–330.