

A Hierarchical Scheme for Rapid Video Copy Detection

Xiao Wu^{1,2}, Yongdong Zhang¹, Sheng Tang¹, Tian Xia^{1,2} and Jintao Li¹

¹Key Laboratory of Intelligent Information Processing

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

²Graduate University of Chinese Academy of Sciences

{wuxiao, zhyd, ts, txia, jtli}@ict.ac.cn

Abstract

Today with the rapid increasing popularity of web video sharing, digital copyright protection encounters many troubles. Video copy detection schemes are emerging to cope with the digital video piracy and illegal distribution problems. But the large amount of video data and diversity of copy attacks pose difficulties on copy detection. This paper presents a hierarchical scheme to detect video copies, especially the temporal attacked and re-encoded ones. Our algorithm which is based on the ordinal signature of intra frames and effective R-Tree indexing structure archives real time performance. Comparison experiments are conducted on the benchmarked database of CIVR 2007 copy detection showcase and demonstrate the promising results of the proposed approach.*

1. Introduction

With the increasing popularity of online digital video sharing (e.g. YouTube), a huge amount of digital video data are generated and distributed by users everyday. By using advantaged editing tools, the amateur users enjoy combining several interesting movie clips into their own videos, which are later distributed online and may lead to copyright violation.

Content based video copy detection (CBCD) is proposed these years [1] to provide a solution to digital copyright enforcement and usage tracking. An effective copy detection scheme is desired to face two challenging issues. One is coping with the variety of attacks and the other is finding the copies from innumerable suspicious candidates under tolerable computation cost. Previous works [2-7] [11] [16] propose various detection algorithms to cope with copy attacks. While most studies put more emphases on spatial variations, limit effort is made for temporal variations (e.g., fast and slow motion, frame rates change) [2]. Actually in most situations a video copy is produced by simply combining and re-encoding several famous or interesting video segments (e.g., exciting movie clips, popular short web videos) with slight retouch. Another issue is that the rapid increasing amount of web videos

created by amateur users and television station causes very large search range of copies. It is inevitable to employ an efficient indexing structure to speed up the detection process. Finally, one of the very essences of copy detection problem is to achieve a balance between robustness and efficiency.

2. Previous Work

A number of studies [3-6] [8] [9] [10], etc. attempt to use either global features or local invariant features to cope with various attacks, including changes in brightness, color, frame spatial shifts, and geometric transformations such as translation, scaling, and rotation, etc. In [6] audio features are used as a complement feature. [7] presents an approach using DC sequence signature. In [9] IBM research group proposes a new motion signature and the novel application of ordinal signature. They compare motion, intensity and color-based signatures for video sequence matching. In the CIVR 2007 copy detection showcase [17] their algorithm focus on accurate matching on frame level and achieved the highest precision under comparative high time consumption. [2] tends to use the dynamic time warping matching algorithm which also has a high complexity of n^2 .

On another hand, as it is in information retrieval applications, the indexing structure and similarity search scheme is actually one of the most difficult tasks of a CBCD scheme in large fingerprints reference database [11]. [3] uses the Hilbert space-filling curve as a space-partition method. [1] [12] discuss the use of inverted file techniques for indexing frame features.

In contrast to the previous efforts, we believe more attention should be paid on temporal and re-encoding attacks (e.g., fast and slow motion, format change) which are the most common attacks of online videos. Finally the time complexity should also be taken into consideration for the need of real-time detection for online videos. In this study we adopt a more scaleable scheme of combining two filters to distinguish copies from large video corpus and focus on gaining better performance in detecting copies under temporal and re-encoding attacks. Then we use the R*-Tree indexing structure [13] to combine the two level

hierarchical detection scheme for further improvement.

3. The proposed approach

Since only a small portion of online videos are copies, we deem it more important classifying copies from numberless clips than distinguishing copy attacks or copy location. In our work the system is designed to distinguish whether a query video is a copy or not. An overview of the proposed approach is shown in Fig. 1.

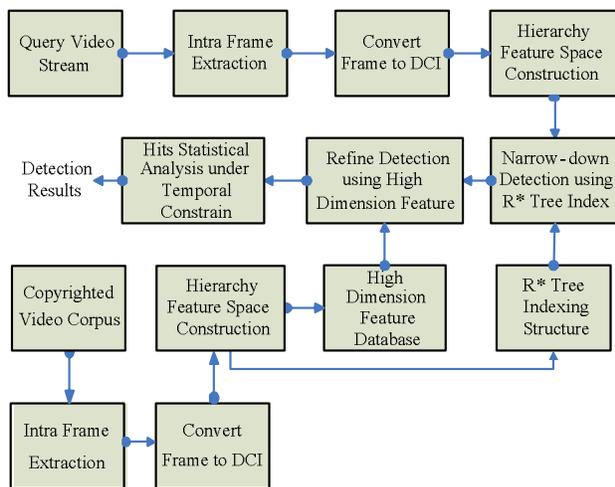


Fig. 1. Diagram of the proposed detection scheme

The proposed approach contains two steps: off-line step and online step. In the off-line step, we collect copyrighted videos to build a database which maintains the extracted features. In the online step, the classifiers process the query video, and determine which query is a copy. The details of the two steps are explained in the next subsections.

3.1. Down-sampled Intra frames extraction

In feature extraction step we make use of all intra frames of the video stream. The use of the intra frame shows two advantages. First the intra frame contains more stable visual content than B and P frames in the MPEG standard. Second the extraction will be very fast without decoding B and P frames from video stream. Similar to the works in [12], as it is showed in Fig.2 (a), we convert frames to grayscale and conduct 2D-DCT on the pixel grayscale matrix. Furthermore, we obtain the main components of every 8*8 DCT coefficients block to present the block and then construct a down sampled image which we called DCI, short for DC image. We use the DCI for two reasons. First, due to the linearity of DCT coefficients of intra frame, the DC value reconstruction without decomposing spends small computation cost. Second, The DCI preserves global visual context of the frames which is adequate for the sub-image feature extraction described later.

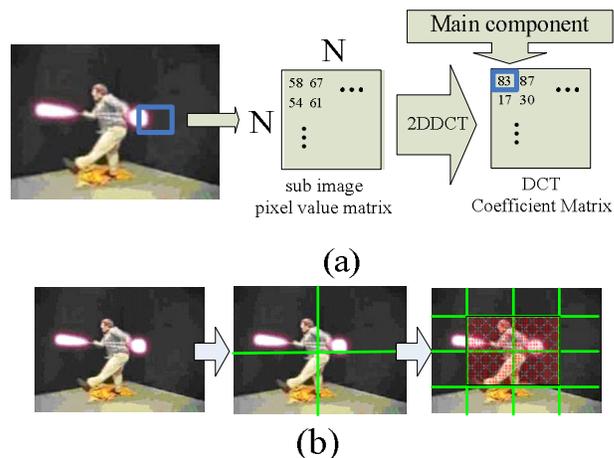


Fig. 2. sub image 2D-DCT and DCI extraction
(a) DCI extraction from the original frame
(b) Segment frame into blocks for feature extraction
The image comes from the StarWarKid,
one of the most popular online video

3.2. Multilevel feature spaces construction and search strategy

As discussed in section 3, a single feature scheme is inadequate for describing various copy appearance. We design a hierarchical copy detection system using multi classifiers to recognize certain copies. It extracts two kinds of features based from the DCIs to construct a two level feature space, and the classifier of every level is called as a Filter. Each filter consists of a special feature space and corresponding searching metric to file out different kind of copies.

3.2.1 First level detection using R*-Tree indexing

Similar to traditional image retrieval, in our work color moments feature is used to be the feature of the first level filter to reduce the number of potential copies before other features are used. We segment a DCI into M*N blocks and introduces the grayscale moments as a course feature for the first level filter. The first and second order moments (mean and stand variance of pixel grayvalues in fact) of every block are computed as (1) and (2),

$$\mu = \frac{1}{N} \sum_{i=1}^N P_i \quad (1)$$

$$\sigma = \left(\frac{1}{N} \sum_{i=1}^n (P_i - \mu)^2 \right)^{\frac{1}{2}} \quad (2)$$

in which N denotes pixel number of a block, P_i denotes the grayvalue of the pixel i . Similar to the ordinal intensity signature in [9] [4] [14], we use the sequence of the ordinal

signatures of each DCI as the feature vector. The sequence provides a more efficient way of video matching [4]. As it is showed in Fig.2, we adopt $M=2$, $N=2$, and for each DCI we obtain a feature vector of 8 dimensions ($2*2*2=8$).

On another hand, in real cases it is inevitable that a practical industry copy detection system should hold huge number of original video data, while the time consumption spend on detection should be taken into consideration. In contrast to the frame by frame matching metrics using in previous works [9] [4], our scheme adopt an efficient technical for very large database retrieval, called the R*-Tree indexing structure [13]. The dimension of the block moments feature is low, which is suitable to construct an R*-Tree indexing for very fast feature matching. In additional, as we discussed in section 1 that a query is probably not a copy of the videos in database, the k-nearest neighbors (KNN) algorithm using visual similarity in information retrieval is not suitable to this problem [3].

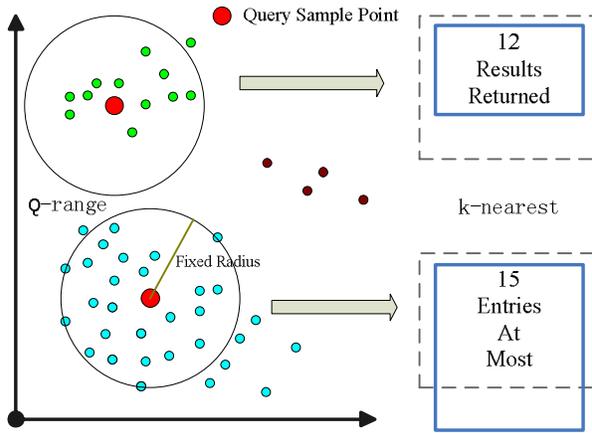


Fig.3. A combination of Q-range and k-nearest querying metrics. As a tradeoff between computation cost and robust, At most 15 candidates to be return for further detection for example

We adopt the combination of Q-range and k-nearest querying metrics for copy detection, as is shown in Fig.3. In this study, we set Q-range radius fixed to 0.12.

3.2.2 Second level detection using high dimension feature

As it is showed in Fig.2 (b), a weighted block grayscale histogram based feature is extracted. In this stage, first the DCI is divided into $M_x * M_y$ blocks and a grayscale histogram is computed for each block. Secondly a weighted block grayscale histogram is computed using the weighting function below. The weighting strategy pays more strength on the visual context of frame centre for the reason that a number of copy attacks (such as cropping, etc.) usually preserve the centre blocks of frames as original. The

weighted block grayscale histogram is used as the second level feature, denoted as WHIS.

$$WHIS_i = HIS_i * weight_i \quad (3)$$

Where $i = 1, 2, \dots, M_x * M_y$

As shown in the Algorithm 1, the Weighted Histogram (WHIS) is computed for each DCI.

Algorithm 1: Fine feature WHIS generation

For each DCI_i , $i = 1, 2, \dots, D$ ($D = \text{number of DCI}$)

 Divided DCI_i into $M_x * M_y$ blocks

 For each block B_{mn}

 Compute grayscale histogram H_{mn} of X bins

 End

 Construct HIS with the length of $M*N*X$ by splicing each H_{mn}

 Compute weighted WHIS using weighting function as (3)

End

In our experiments, the $M_x = 4$, $M_y = 4$, $X = 32$ and weighting function is as (4):

$$weight_i = \begin{cases} 0.1, & \text{where } i = 6, 7, 10, 11 \\ 0.05, & \text{otherwise} \end{cases} \quad (4)$$

which keeps the WHIS normalized. To match the WHIS, we use L2 distance which is suitable for matching histogram based features.

3.3. Hits statistics under temporal constrains

To design a algorithm robust to temporal attacks we introduce the temporal constrains in online process of queries. QV denotes the input video stream, and the extracted intra frames are denoted as $QV[v][i]$, $i = 0, 1, \dots, N$; $v = 0, 1, \dots, V$. Similarly, we denote videos in the database as DV, and denote the i th frame of v th video as $DV[v][i]$. We call a frame matching between $QV[v][i]$ and $DV[w][j]$ as a Hit. After the second level filtering, a batch of Hits is obtained for each query video. An example of Hits statistics is given in Fig.4, as the order of corresponding matching contains intrinsic temporal information.

Essentially it is a procedure of searching for global optimal in sequence matching. First we adopt a temporal constrain mechanism to refine Hits of each query video. As

is shown in Fig.4, only those continual Hits are preserved by using a slide window with fixed length of WL.

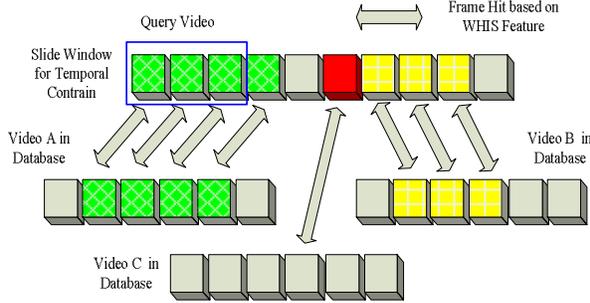


Fig.4. Hits of a query video. The temporal constrain window length (WL) is fixed to 3

Second, to determine the final copy “similarity” between query and the original videos, we conduct Hits statistics by simply summing all preserved Hits together as follow,

$$DI[QV][DV] = \sum_{i=1}^m \# Hits_i \quad (5)$$

where m denotes the number of matched segments. $\#$ denotes the number of Hits and DI denotes the Degree of Identity, which is used as the final index.

4. Experiments

In the discussion of previous works we mention that Chiu’s scheme [2] is dedicated to temporal copy attacks. The algorithm based on DTW allows comparison of high tolerance to tempo variation of video sequences and declaimed to be superior to [4] and [15]. In this section first we present how we design the test data for experiments. Then a comparison evaluation is conducted between Chiu’s scheme [2] and ours.

4.1. Test data setup

In our research, we use a large benchmarked dataset [18] from CIVR 2007 video copy detection showcase [17] as database, which contain 101 videos, approximately 5,220,000 frames, totally 100G data lasts more than 58 hours. All the videos are in Video-CD standard, MPEG1, PAL, 352x288, 1150Kb/s, 25 fps (MPEG-1 Audio Layer II, 44100 Hz, Stereo, and 224 Kb/s). This corpus includes commercials, sports, news, movies, TV shows, documentary films, home videos, landscapes, music TVs and edited clips supplied by video amateurs. Video length ranges from 10 seconds to 2 hours. In the off-line step, our system extracts 365058 intra frames from the videos in the database. The query videos are consists of 67 copies and 38 non-copies. As discussed before, we believe a large portion

of queries is non-copies. In the dataset, the duration of copy and non-copy is respectively 3h40min and 5h28min42sec. We prepare many short copies for the reason that in practice the majority of copies has short duration. The 67 segments include 57 clips extracted randomly from original database and 10 videos which are the copies of Task ST1 in CIVR showcase [17]. The 10 queries are clustered as spatial-temporal attacked copies. Each segment contains clips from one certain original video and re-encoded into one of the formats in Table.1 (b). Both spatial and temporal attacks are posed upon these segments.

Table 2 lists the used attacks. The 38 segments were randomly selected from TRECVID2007 corpus as non-copies.

Table.1. query attacks design using the video edit tool Moviemaker on windows platform. # denotes the number of clip or frame

Index Query set	# clips	Total duration	#Intra frames	Attack methods
TRECVID 2007 videos	38	5h28min42sec	49305	Non-copy
Fast and slow forward	9	8min38sec	1554	Double and half speed
Video format change	7	3min35sec	602	in Table.1 (b)
Temporal hybrid attack	21	18min	3240	swapping and combination
Spatial-temp oral hybrid	30	3h8min17sec	16201	gray, blurring, zoom, etc

(a) queries and the corresponding attacks

Index Format	bit rate	resolutio n	frame rate
Pocket PC (218 Kbps)	218 Kbps	208*160	20
Pocket PC (143 Kbps)	143 Kbps	208*160	8
Pocket PC (full screen 218 Kbps)	215 Kbps	320*240	15
Wind band (340 Kbps)	340 Kbps	320*240	25
Local play (2.1 Mbps PAL)	2.1 Mbps	720*576	20

(b) format changes

4.2. Evaluation using ROC curves

In the experiments we plot Receive Operating Characteristics (ROC) curves to evaluate the performance. The false negative rate FNR and false positive rate FPR are computed as equation (6),

$$FNR(k) = \frac{FN}{N_C} \quad \text{and} \quad FPR(k) = \frac{FP}{N_T} \quad (6)$$

where FP denotes number of false positives (non-copy clips that are detected). FN denotes number of false negatives (copy clips that are not detected). N_T denotes

the number of non-copy clips. N_C denotes the number of copy clips. k denotes the threshold for $K(h_1, h_2)$ (in section 3.2). We compare the results of the proposed approach against Chiu's algorithm [2], which is based on dynamic time warping matching strategy and claim to be outperform previous works [4] [15]. Fig.5 illustrates the ROC curves of the two approaches under the attacks in Table.1. These curves are draw as the threshold k changes. In experiments we try 0.69, 0.7, 0.71, 0.725 and 0.75 as the threshold.

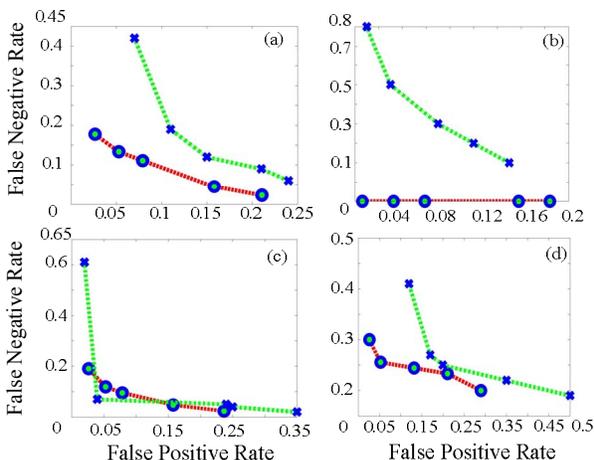


Fig.5. the ROC curves of pressure experiments.

Circle are our results and cross are Chiu's [2]

(a) fast and slow forward; (b) video format change; (c) temporal hybrid attacks; (d) spatial-temporal hybrid attacks

Fig.5 (a) shows the result that the performance of our approach is better than Chiu's under fast and slow motion attacks. The reason is that fast and slow motion attacks do not change the order or location of intra frames. Fig.5 (b) shows the result by combing three common variations: bit rate change, resolution change and frame rate change (in Table.1 (b)). Our approach has immunity to frame rate change. This is because the frame rate changing (from 25fps to 15fps, e.g.) do not change the order and appearance of intra frames. Since the ordinal signature of image block grayscale moments and histograms are adopted, the two approaches are robust to resolution change. Fig.5 (c) shows the result of detecting some short copies, which are embedded into long TRECVID videos. Since there is no entire length segment for time warp matching, the false negative rate of Chiu's result is high. In contrast, our method adopt flexible temporal constrains to avoid this situation and maintain a more smooth curves. Fig.5 (d) shows that our approach is also robust to common spatial attacks. The spatial attacks include gray, blurring, zoom in and out and even special effects such as watercolor and poster effects. In our scheme, the primary information of videos is extracted and preserved in the DCIs, i.e. the down sampled images of video intra frames. This

transformation facilitates the usage of indexing structure and maintains certain invariance to common spatial attacks.

4.3. Computation cost analysis

As discussed in section 3, copy detection algorithm must take the time consumption into consideration. We use a standard PC (Celeron 3.06GHz with 1G ram) to conduct the experiment to evaluate the time complexity of the proposed method and the Chiu's. The result is shown in Fig.6 as follow.

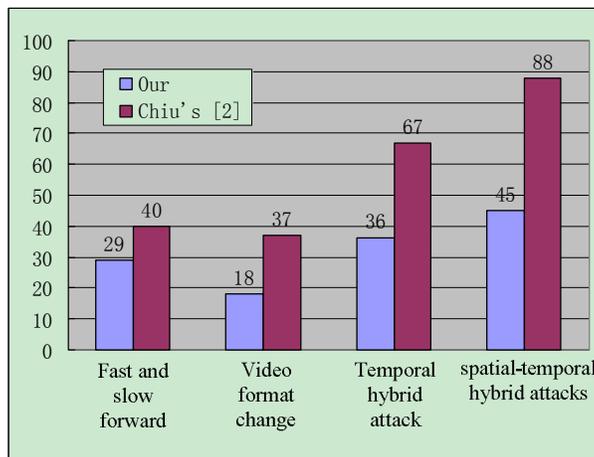


Fig.6. time consumption (seconds) of the two approaches.

The time unit of detecting spatial-temporal hybrid attacks is minute. Time consumption increases as the duration of queries increases (in Table 1(a)).

Our approach is much faster than Chiu's

According to [2], a 1,000 frames clip costs 3 seconds for querying the target collection which contains 712,060 frames (29.97fps). However in our experiments the DTW based method is not suitable for large-vocabulary applications because of the massive computation cost ($m*n$ complexity) with almost no efficient indexing algorithms for nearest neighbor search. The optimal performance of Chiu's method is nearly 11 times faster than real-time play (it cost the system 1 hour to analyses videos of 11 hours), while experiments shows the average time complexity of the proposed method is 35 times faster than real-time play. This is because we employ an efficient indexing structure to speed up the process and use the multilevel classifiers to file out noises, as discussed in the future work of [2].

5. Conclusion and future works

In this study we propose a hierarchical copy detection scheme robust to common temporal and re-encoding attacks, such as fast/slow forward, clip swapping and frame rate change, etc. The detection process achieves rapid speed on a large video collection. Comparison experiments with the state-of-art algorithm prove the efficiency and effect of

our approach. The features used in our scheme are relative simple, but the hierarchical framework based on advanced indexing structure is very suitable for large scale copy detection applications.

Our future works will be focus on finding novel machine learning algorithm combining more precise visual features, such as the local feature trajectory and motion feature. And the framework can expand to three or more levels for better performance.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (60773056), the Beijing New Star Project on Science & Technology (2007B071), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416) and the National Basic Research Program of China (973 Program, 2007CB311100).

References

- [1] Arun Hampapur and Rudolf M. Bolle, "VideoGREP: Video Copy Detection using Inverted File Indices," Technical Report, IBM T.J Watson Research Center IBM Research Division, 2001.
- [2] Chih-Yi Chiu, Cheng-Huang Li, and Hsiang-An Wang, "A Time Warping Based Approach for Video Copy Detection", The 18th International Conference on Pattern Recognition, 2006.
- [3] A. Joly, O. Buisson, and C. Frelicot, "Content-Based Copy Retrieval Using Distortion-Based Probabilistic Similarity Search", IEEE Transactions on Multimedia, 2007.
- [4] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No. 1, pp. 127-132, 2005.
- [5] J. Law-To, O. Buisson, V. Gouet-Brunetand, and N. Boujemaa, "Robust voting algorithm based on labels of behavior for video copy detection", in ACM Multimedia, pages 835-844, 2006.
- [6] P. Indyk, G. Iyengar, and N. Shivakumar, "Finding pirated video sequences on the internet", Technical report, Stanford University, 1999.
- [7] M. Naphade, M. Yeung, and B. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures", The SPIE Conference on Storage and Retrieval for Media Databases, 2000.
- [8] E. Chang, J. Wang, C. Li, and G. Wilderhold, "Rime: a replicated image detector for the world-wide web", in Proc. of SPIE Symp. of Voice, Video, and Data Communications, 1998, pp. 58-67.
- [9] A. Hampapur, K.H. Hyun, and R. M. Bolle, "Comparison of sequence matching techniques for video copy detection", The SPIE Conference on Storage and Retrieval for Media Databases, 2002.
- [10] B. Coşkun, B. Sankur, N. Memon, "Spatial-Temporal Transform Based Video Hashing", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 8, NO. 6, DECEMBER 2006.
- [11] A. Joly, O. Buisson, and C. Frelicot, "Statistical similarity search applied to content-based video copy detection", Proceedings of the 21st International Conference on Data Engineering Workshops, 2005.
- [12] J. Oostveen, T. Kalker, and J. Haitisma, "Feature extraction and a database strategy for video fingerprinting", in Proc. of Int. Conf. on Visual Information and Information Systems, 2002, pp. 117-128.
- [13] N. Beckmann, H-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: an efficient and robust access method for points and rectangles", in Proc. of ACM SIGMOD Int. Conf. on Management of Data, 1990, pp. 322-331.
- [14] S Lee, CD Yoo, "VIDEO FINGERPRINTING BASED ON CENTROIDS OF GRADIENT ORIENTATIONS", Acoustics, Speech and Signal Processing, 2006.
- [15] XS Hua, X Chen, HJ Zhang, "Robust video signature based on ordinal measure", International Conference on Image Processing, 2004.
- [16] A. Jaimes, S. F. Chang, and A. C. Loui, "Duplicate detection in consumer photography and news video", in Proc. of ACM Int. Conf. on Multimedia, 2002, pp. 423.424.
- [17] <http://staff.science.uva.nl/~civr2007/videocopy.php>
- [18] The origin of the video corpus is MUSCLE-VCD-2007 <http://www-rocq.inria.fr/imedia/civr-bench/index.html>