

Fast and Robust Short Video Clip Search for Copy Detection

Junsong Yuan^{1,2}, Ling-Yu Duan¹, Qi Tian¹, Surendra Ranganath², and Changsheng Xu¹

¹ Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{jyuan, lingyu, tian, xucs}@i2r.a-star.edu.sg

² Department of Electrical and Computer Engineering, National University of Singapore
elesr@nus.edu.sg

Abstract. Query by video clip (QVC) has attracted wide research interests in multimedia information retrieval. In general, QVC may include feature extraction, similarity measure, database organization, and search or query scheme. Towards an effective and efficient solution, diverse applications have different considerations and challenges on the abovementioned phases. In this paper, we firstly attempt to broadly categorize most existing QVC work into 3 levels: concept based video retrieval, video title identification, and video copy detection. This 3-level categorization is expected to explicitly identify typical applications, robust requirements, likely features, and main challenges existing between mature techniques and hard performance requirements. A brief survey is presented to concretize the QVC categorization. Under this categorization, in this paper we focus on the copy detection task, wherein the challenges are mainly due to the design of compact and robust low level features (i.e. an effective signature) and a kind of fast searching mechanism. In order to effectively and robustly characterize the video segments of variable lengths, we design a novel global visual feature (a fixed-size 144-d signature) combining the spatial-temporal and the color range information. Different from previous key frame based shot representation, the ambiguity of key frame selection and the difficulty of detecting gradual shot transition could be avoided. Experiments have shown the signature is also insensitive to color shifting and variations from video compression. As our feature can be extracted directly from MPEG compressed domain, lower computational cost is required. In terms of fast searching, we employ the active search algorithm. Combining the proposed signature and the active search, we have achieved an efficient and robust solution for video copy detection. For example, we can search for a short video clip among the 10.5 hours MPEG-1 video database in merely 2 seconds in the case of unknown query length, and in 0.011 second when fixing the query length as 10 seconds.

1 Introduction

As a kind of content-based video retrieval, Query by video clip (QVC) has posed many applications such as video copy detection, TV commercial & movie identification, and high level concept search. In order to implement a QVC solution, we have to solve the following challenges: 1) how to appropriately represent the video content and define similarity measure; 2) how to organize and access the very large dataset consisting of

large amounts of continuous video streams; and 3) the choice of a fast searching scheme to accelerate the query process. Towards an effective and efficient solution, diverse applications have different considerations and challenges on the abovementioned phases due to different search intentions. Different strategies and emphasis are thus applied. For example, the task of retrieving “similar” examples of the query at the concept level is associated with the challenge of capturing and modeling the semantic meaning inherent to the query [1] [2]. With an appropriate semantics modeling, those examples (a shot or a series of shots) with a similar concept as the query can be found. Here we are not concerned with search speed since the bottleneck against a promising performance is inherent to the gap between low-level perceptual features and high-level semantic concepts. In terms of video copy detection, an appropriate concept-level similarity measure is not required as the purpose is only to identify the presence or locate the re-occurrences of the query in a long video sequence. However, the prospective features or fingerprints are expected to be compact and insensitive to variations (e.g. different frame size, frame rate and color shifting) brought by digitization and coding. Particularly the search speed is a big concern. The reasons are twofold. Firstly, its application is usually oriented to a very large video corpus or a time-critical online environment; Secondly, the mostly used frame-based or window-based matching coupled with a shifting mechanism causes more serious granularity than the shot-based concept-level retrieval, wherein we have to quickly access much more high-dimensional feature points.

Based on the above discussions, we attempt to broadly categorize most existing QVC works into 3 levels, as illustrated in Fig.1. The production procedure of video content (left) is depicted and those associated QVC tasks at 3 different levels (right) are listed. Such categorization is expected to roughly identify common research issues, emphasis and challenges within different subsets of applications in diverse environments.

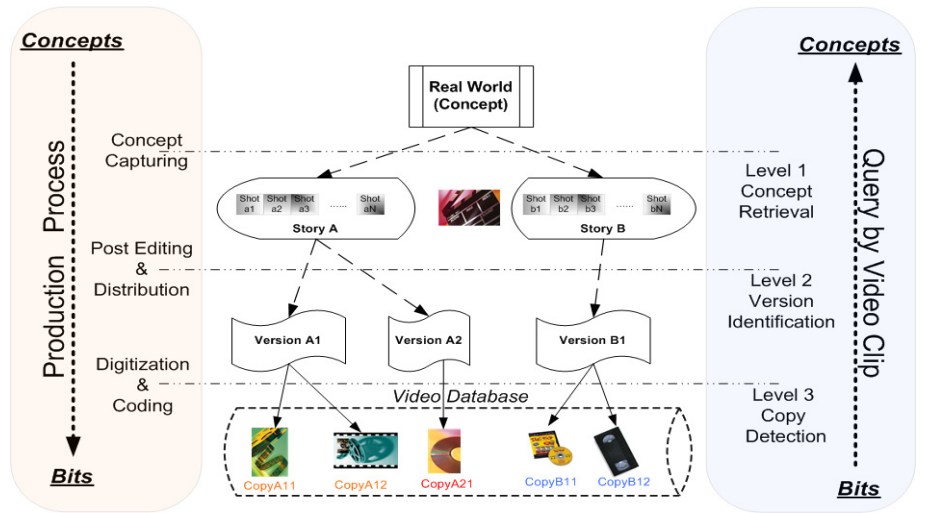


Fig. 1. A three layer framework for query by video clip.

Table 1. A Concretization of three-level QVC framework together with representative works.

Retrieval categories	Level 1 Video Copy Detection	Level 2 Video Title Identification	Level 3 Concept-based retrieval
Goals of retrieval operations	To retrieve copies of the query content, but variations with coding parameter changes are allowed, see below. (query and target instances share same content, same version, but may have coding variations) [15] [16][17][18][19][20][21][22][23][24] [25] [26] [27] [28] [29] [31] [32]	To retrieve contents with variations from the query content due to post editing. (query and target instances are from the same title, but are different versions. For instance, a movie and its trailers, an original advertisement and its reedited shortened version) [9] [10] [11][12] [13] [14] [20] [24] [28]	Concept search at semantic level. (query and target examples share “similar” concept or content, for instance, to retrieve video clips of a Boeing 747’s take off scene, to find out all the free-throw scenes in a basketball match) [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [19]
Variations allowed	Coding variations, including • Frame rate • Bit rate • Frame size • Color shift • Digitization noises etc.	Shot level editing (shot removal, insertion, summary etc.); Image frame editing (caption insertion, Logo insertion, frame shift, etc.); same movie but in different language editions, etc.	Many instances of the same concepts, but appearing in different setting and environments, including but not limited to camera angle variation, scale, acquisition time and place variations, etc.
Likely features used	Mainly low level features such as color, gray level features, independent of video structure	Low level features and features from image, shot or object levels; content modeling	Refined low level features; Object level features and high level concept features; Graph model, motion model, etc.
Research challenges	Precise similarity measure at concept-level is not required, but the signatures need to be compact and robust against various coding variations; [15] [16] [19] [20] [22] [24] [25] [26] The target database usually contains unlabeled video streams or are considering on-line application environment, therefore efficient feature extraction and fast search is required; [15] [17] [18] [20] [21] [22] [23] [27] [29] [31] [32] Database should be organized properly to support query with various length	Efficient detection and robust representation against various post-editing effects are needed; [10] [11] [12] [13] [14] [20] [24] [28]	Robust detection of high level concepts, such as in-door, out-door (TRECVID concepts); [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]; Effective organization and indexing and the large dataset to support fast access and efficient query [30]
Existing Works	NTT [17] [18] [32], I ² R [25] [26], RMIT [22], University of Kansas [21], IBM [16], Philips [29], Intel [15], NEC [23], INRIA [20]	Berkeley University [11] [12], NUS [13], Columbia University [10], Oakland University [14], INRIA [20], Fraunhofer Institute [24]	TRECVID [1], Michigan State University [3], Columbia University [9] [10], I ² R [6] [7] [8] etc.

Under this framework, our work in this paper is focused on video copy detection, the lowest search level (See Sections 3, 4, 5). We want to jointly take into account the robustness issue and the search speed issue to complete the efficient and effective detection. The experimental dataset includes 10.5 hours video collections and in total 84 given queries with the length ranging from 5 to 60 seconds are performed. Our experiments have shown that both fast search speed and good performance can be accomplished at the lowest retrieval level.

2 Related Works

After a comprehensive literature review [1-32], we concretize the framework as listed in the Table 1. The references are roughly grouped around application intentions and their

addressed research challenges respectively. Due to limited space, no detailed comparison will be given here.

3 Feature Extraction for Video Copy Detection

In video copy detection, the signature is required to be compact and efficient with respect to large database. Besides, the signature is also desired to be robust to various coding variations mentioned in Table 1. In order to achieve this goal, many signature and feature extraction methods are presented for the video identification and copy detection tasks [11] [15] [16] [26] [28] [29].

As one of the common visual features, color histogram is extensively used in video retrieval and identification [15] [11]. [15] applies compressed domain color features to form compact signature for fast video search. In [11], each individual frame is represented by four 178-bin color histograms in the HSV color space. Spatial information is incorporated by partitioning the image into four quadrants. Despite certain level of success in [15] and [11], the drawback is also obvious, e.g. color histogram is fragile to color distortion and it is inefficient to describe each individual key frame using a color histogram as in [15].

Another type of feature which is robust to color distortion is the ordinal feature. Hampapur et al. [16] compared performance of using ordinal feature, motion feature and color feature respectively for video sequence matching. It was concluded that ordinal signature had the best performance. The robustness of ordinal feature was also proved in [26]. However, based on our experiments, we believe better performance could be achieved by combining ordinal features and color range features appropriately, with the former providing spatial information and the latter providing range information. Experiments in Section 5 support these conclusions. As a matter of fact, many works such as [3] and [14] also incorporate the combined feature in order to improve the performance of retrieval and identification.

Generally, the selection of ordinal feature and color feature as signature for copy detection task is motivated by the following reasons:

- (1) Compared with computational cost features such as edges, texture or refined color histograms which also contain spatial information (e.g. *color coherent vector* applied in [28]), they are inexpensive to acquire
- (2) Such features can form compact signatures [29] and retain perceptual meaning
- (3) Ordinal features are immune to global changes in the quality of the video and also contain spatial information, hence are a good complement to color features [26]

3.1 Ordinal Feature Description

In our approach, we apply Ordinal Pattern Distribution (*OPD*) histogram proposed in [26] as the ordinal feature. Different from [26], the feature size is further compressed in this paper, by using more compact representation of I frames. Figure 2 depicts the operations of extracting such features from a group of frames.

For each channel $c = Y, Cb, Cr$, the video clip is represented by *OPD* histograms as:

$$H_c^{OPD} = (h_1, h_2, \dots, h_l, \dots, h_N) \quad 0 \leq h_i \leq 1 \text{ and } \sum_i h_i = 1 \quad (1)$$

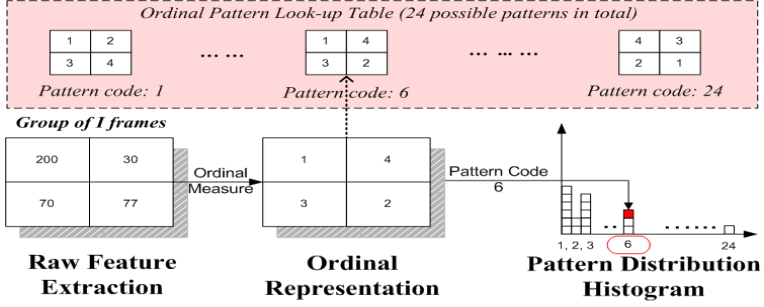


Fig. 2. Ordinal Pattern Distribution (OPD) Histogram.

Here $N = 4! = 24$ is the dimension of the histogram, namely the number of possible patterns mentioned above. The total dimension of the ordinal feature is $3 \times 24 = 72$.

The advantages of using *OPD* histograms as visual features are two fold. First, they are robust to frame size change and color shifting as mentioned above. And secondly, the contour of the pattern distribution histogram can describe the whole clip globally; therefore it is insensitive to video frame rate change and other local frame changes compared with key frame representation.

3.2 Color Feature

For the color feature, we characterize the color information of a GoF by using the cumulative color information of all the sub-sampled I frames in it. For computational simplicity, Cumulative Color Distribution (*CCD*) is also estimated using the DC coefficients from the I frames.

The cumulative histograms of each channel ($c=Y, Cb, Cr$) can be defined as:

$$H_c^{CCD} = \frac{1}{M} \sum_{i=b_k}^{b_k+M-1} H_i(j) \quad j = 1, \dots, B \quad (2)$$

where H_i denotes the color histogram describing an individual I frame in the segment. M is the total number of I frames in the window and B is the color bin number. In this paper, $B = 24$ (uniform quantization). Hence, the total dimension of the color feature is also $3 \times 24 = 72$, representing three color channels.

4 Similarity Search and Copy Detection

For visual signature matching, Euclidean distance $D(\cdot, \cdot)$ is used to measure distance between the query Q (represented by H_Q^{OPD} and H_Q^{CCD} , both are 72-d signatures) and the sliding matching window SW (represented by H_{SW}^{OPD} and H_{SW}^{CCD} , both are 72-d signatures). The integrated similarity S is defined as the reciprocal of linear combination

of the *average* distance of *OPD* histograms and the *minimum* distance of *CCD* histograms in the Y, Cb, and Cr channels:

$$D^{OPD}(H_Q^{OPD}, H_{SW}^{OPD}) = \frac{1}{3} \sum_{c=Y,Cb,Cr} D(H_Q^{OPD}, H_{SW}^{OPD}) \quad (3)$$

$$D^{CCD}(H_Q^{CCD}, H_{SW}^{CCD}) = \underset{c=Y,Cb,Cr}{Min} \{D(H_Q^{CCD}, H_{SW}^{CCD})\} \quad (4)$$

$$S(H_Q, H_{SW}) = \frac{1}{w \times D^{OPD} + (1 - w) \times D^{CCD}} \quad (5)$$

Let the similarity metric array be $\{S_i; 1 \leq i \leq m + n - 1\}$ corresponding to similarity values of $m + n - 1$ sliding windows, where n and m are the I frame number of the query clip and the target stream respectively. Based on [17] and [32], the search process can be accelerated by skipping unnecessary steps. The number of skipped steps w_i is given as:

$$w_i = \begin{cases} \text{floor}(\sqrt{2}D(\frac{1}{S_i} - \theta)) + 1 & \text{if } S_i < \frac{1}{\theta} \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

where D is the number of I frames of the corresponding matching window. θ is the predefined skip threshold.

After the search, potential start position of the match is determined by a local maximum above the threshold, which fulfills the following conditions:

$$S_{k-1} \leq S_k \geq S_{k+1} \text{ and } S_k > \max\{T, m + k\sigma\} \quad (7)$$

where T is the pre-defined preliminary threshold, m is the mean and σ is the deviation of the similarity curve; k is an empirically determined constant. Only when similarity value satisfies (7), is it treated as the detected instance. In our experiments, w in (5) is set to 0.5, and θ in (6) is set to 0.05, and T in (7) is set to 6.

5 Experimental Results

All the simulations were performed on a P4 2.53G Hz PC (512 M memory). The algorithm was implemented in C++. The query collection consists of 83 individual commercials which varied in length from 5 to 60 seconds and one 10-second long news program lead-out clip (Fig. 3). All the 84 given clips were taken from ABC TV news programs. The experiment sought to identify and locate these clips inside the target video collection, which contains 22 streams of half-hour broadcast ABC news video (obtained from TRECVID news dataset [1]). The 83 commercials appear in 209 instances in these half-hour news programs; and the lead-out clip appears in total 11 instances. The re-occurrence instances usually have color shifting, I frame shifting and frame size variations with respect to the original query. All the video data were encoded in MPEG1 at 1.5 Mb/sec with image size of 352×240 or 352×264 and frame rate of 29.97 fps. It is compressed with the frame pattern IBBPBBPBBPBB, with I frame temporal resolution around 400 ms. Fig. 3 and Fig. 4 give two examples of extracted features.

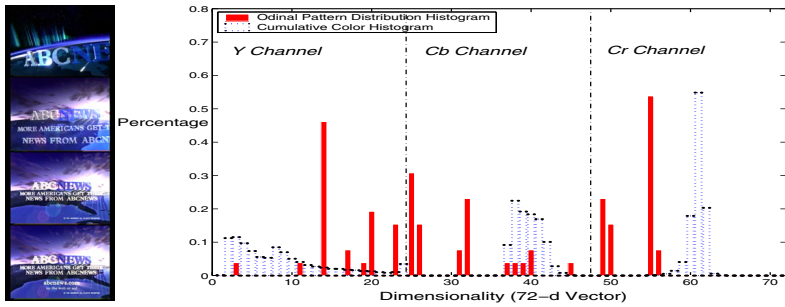


Fig. 3. ABC News program lead-out clip (left, 10 sec) and its *CCD* and *OPD* signatures (right).

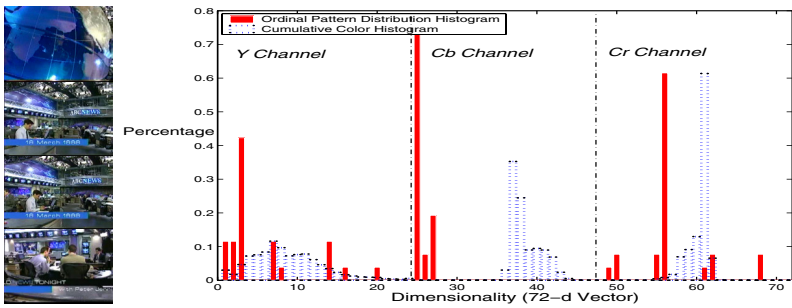


Fig. 4. ABC News program lead-in clip (left, 10 sec) and its *CCD* and *OPD* signatures (right).

We note here that the identification and retrieval of such repeated *non-news* sections inside a video stream helps to reveal the video structure. These sections include TV commercials, program lead-in/lead-out and other Video Structure Elements (*VSE*) which appear very often in many types of video to indicate starting or ending points of a particular video program, for instance, news programs or replay of sports video.

Table 2 gives the approximate computational cost of the algorithm. The task is to search for instances of the 10 second long lead-out clip (Fig. 3) in the 10.5 hour MPEG-1 video dataset. The *Feature Extraction* step includes DC coefficient extraction from the compressed domain, the formation of color histogram ($3 \times 24-d$) of each I frame (H_i histogram in (2)). This step could be done off-line for the 10.5-hour database. On the other hand, *Signature Processing* consists of the procedures to form *OPD* and *CCD* signatures for the specific matching windows during the active search. Therefore its cost may vary according to the length of the window, namely the length of the query. If the query length is known or fixed beforehand, signature processing step could also be done off-line. In that case, the only cost of active search is *Similarity Calculation*. In our experiment, similarity calculation through a video database of 10.5 hours needs only 11 milliseconds.

The performance of searching for the instances of the given 84 clips in the 10.5 hour video collection is presented in Fig. 5. From the experimental results we found that a

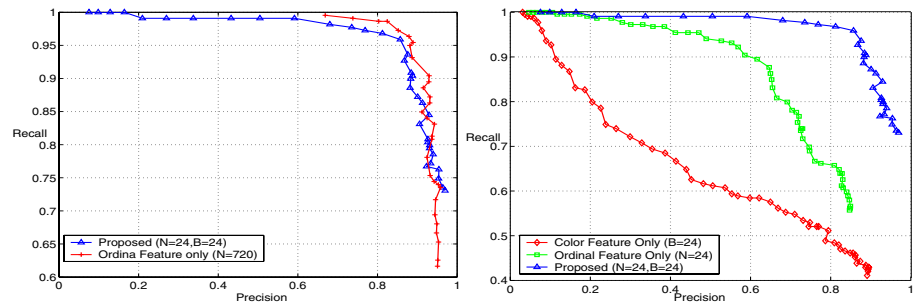


Fig. 5. Performance comparison using different feature: proposed features vs. 3×720 -d *OPD* feature (left); proposed features vs. 3×24 -d *CCD* feature and 3×24 -d *OPD* feature respectively (right); the detection curves are generated by varying the parameter k in (7) ($Precision = detects / (detects + false\ alarms)$) ($Recall = detects / (detects + miss\ detects)$).

large part of the false alarms and missed detections are mainly caused by the I frame *shifted matching* problem, when the sub-sampled I frames of the given clip and that of the matching window are not well aligned in temporal axis. Although the matching did not yield 100% accuracy using the proposed signatures (72-d *OPD* and 72-d *CCD*), it still obtains performance which is comparable with that of [26], where only *OPD* with $N=720$ is considered. However, compared with [26] whose feature size is $3 \times 720=2160$ dimension, our proposed feature is as small as a $(3 \times 24+3 \times 24)=144$ dimensional vector, *15 times* smaller than that of [26]. Besides, in terms of Fig. 5, it is obvious that better performance can be achieved by using the combined features than using only *CCD* (color feature) or only *OPD* (ordinal feature) respectively.

6 Conclusion and Future Work

In this paper, we have presented a three-level QVC framework in terms of how to differentiate the diverse “similar” query requests. Although huge amounts of QVC research have been targeted in different aspects (e.g. feature extraction, similarity definition, fast search scheme and database organization), few work has tried to propose such a framework to explicitly identify different requirements and challenges based on rich applications. A closely related work [28] has just tried to differentiate the meanings of “similar” at different temporal levels (i.e. frame, shot, scene and video) and discussed various strategies at those levels. According to our experimental observation and comparisons among different applications, we believe that a better interpretation of the term of

Table 2. Approximate Computational Cost Table (CPU time).

Task: Search for a 10 second long query clip	Feature Extraction	Active Search		
		Signature Processing		Similarity Calculation of Signatures
		Ordinal Feature	Color Feature	
10.5 h MPEG1 Video (93,725 I frames)	1178.034 sec	0.969 sec	0.688 sec	0.011 sec

“similar” is inherent to the user-oriented intentions. For example, in some circumstances, the retrieval of “similar” instances is to detect the exact duplicate or re-occurrences of the query clip. Sometimes, the “similar” instances may designate the re-edited versions of the original query. Besides, searching “similar” instances could also be the task of finding video segments sharing the same concept or having the same semantic meaning as that of the query. Different bottlenecks and emphasis exist at these different levels.

Under the framework, we have provided an efficient and effective solution for video copy detection. Instead of the key frames-based video content representation, the proposed method treats the video segment as a whole, which is able to handle video clips of variable length (e.g. a sub-shot, a shot, or a group of shots). However, it does not require any explicit and exact shot boundary detection.

The proposed *OPD* histogram has experimentally proved to be a useful complement to the *CCD* descriptor. Such an ordinal feature can also reflect a global distribution within a video segment by the accumulation of multiple frames. However, the temporal order of frames within a video sequence has not yet been exploited sufficiently in *OPD*, and also in *CCD*. Although our signatures are useful for those applications irrespective of different shot order (such as the commercial detection in [13]), the lack of frame ordering information may make the signatures less distinguishable. Our future work may include how to incorporate temporal information, how to represent the video content more robustly and how to further speed up the search process.

References

- [1] <http://www-nlpir.nist.gov/projects/trecvid/>. Web site, 2004
- [2] N.Sebe et al., “The state of the art in image and video retrieval,” In *Proc. of CIVR’03*, 2003
- [3] A. K. Jain et al., “Query by video clip,” In *Multimedia System*, Vol. 7, pp. 369-384, 1999
- [4] D. DeMenthon et al., “Video retrieval using spatio-temporal descriptors,” In *Proc. of ACM Multimedia’03*, pp. 508-517, 2003
- [5] Chuan-Yu Cho et al., “Efficient motion-vector-based video search using query by clip,” In *Proc. of ICME’04*, Taiwan, 2004
- [6] Ling-Yu Duan et al., “A unified framework for semantic shot classification in sports video,” To appear in *IEEE Transaction on Multimedia*, 2004
- [7] Ling-Yu Duan et al., “Mean shift based video segment representation and applications to replay detection,” In *Proc. of ICASSP’04*, pp. 709-712, 2004
- [8] Ling-Yu Duan et al., “A Mid-level Representation Framework for Semantic Sports Video Analysis,” In *Proc. of ACM Multimedia’03*, pp. 33-44, 2003
- [9] Dong-Qing Zhang et al., “Detection image near-duplicate by stochastic attribute relational graph matching with learning,” in *Proc. of ACM Multimedia’04*, New York, Oct. 2004
- [10] Alejandro Jaimes, Shih-Fu Chang and Alexander C. Loui, “Detection of non-identical duplicate consumer photographs,” In *Proc. of PCM’03*, Singapore, 2003
- [11] S. Cheung and A. Zakhor, “Efficient video similarity measurement with video signature,” In *IEEE Trans. on Circuits and System for Video Technology*, vol. 13, pp. 59-74, 2003
- [12] S.-C. Cheung and A. Zakhor, “Fast similarity search and clustering of video sequences on the world-wide-web,” To appear in *IEEE Transactions on Multimedia*, 2004.
- [13] L. Chen and T.S. Chua, “A match and tiling approach to content-based video retrieval,” In *Proc. of ICME’01*, pp. 301-304, 2001
- [14] V. Kulesh et al., “Video clip recognition using joint audio-visual processing model,” In *Proc. of ICPR’02*, vol. 1, pp. 500-503, 2002

- [15] M.R. Naphade et al., "A Novel Scheme for Fast and Efficient Video Sequence Matching Using Compact Signatures," In *Proc. SPIE, Storage and Retrieval for Media Databases 2000*, Vol. 3972, pp. 564-572, 2000
- [16] A. Hampapur, K. Hyun, and R. Bolle., "Comparison of Sequence Matching Techniques for Video Copy Detection," In *SPIE, Storage and Retrieval for Media Databases 2002*, vol. 4676, pp. 194-201, San Jose, CA, USA, Jan. 2002.
- [17] K. Kashino et al., "A Quick Search Method for Audio and Video Signals Based on Histogram Pruning," In *IEEE Trans. on Multimedia*, Vol. 5, No. 3, pp. 348-357, Sep. 2003
- [18] K. Kashino et al., "A quick video search method based on local and global feature clustering," In *Proc. of ICPR'04*, Cambridge, UK, Aug. 2004
- [19] A.M. Ferman et al., "Robust color histogram descriptors for video segment retrieval and identification," In *IEEE Trans. on Image Processing*, vol. 1, Issue 5, May 2002
- [20] Alexis Joly, Carl Frelicot and Olivier Buisson, "Robust content-based video copy identification in a large reference database," In *Proc. of CIVR'03, LNCS 2728*, pp. 414-424, 2003
- [21] Kok Meng Pua et al., "Real time repeated video sequence identification," In *Journal of Computer Vision and Image Understanding*, vol. 93, pp. 310-327, 2004
- [22] Timothy C. Hoad, et al., "Fast video matching with signature alignment," In *SIGIR Multimedia Information Retrieval Workshop 2003 (MIR'03)*, pp. 263-269, Toronto, 2003
- [23] Eiji Kasutani et al., "An adaptive feature comparison method for real-time video identification," In *Proc. of ICIP'03*, 2003
- [24] Nicholas Diakopoulos et al., "Temporally Tolerant Video Matching", In *SIGIR Multimedia Information Retrieval Workshop 2003 (MIR'03)*, Toronto, Canada, Aug. 2003
- [25] Junsong Yuan et al. "Fast and Robust Short Video Clip Search Using an Index Structure," in *ACM Multimedia Workshop on Multimedia Information Retrieval (MIR'04)*, 2004
- [26] Junsong Yuan et al., "Fast and Robust Search Method for Short Video Clips from Large Video Collection," in *Proc. of ICPR'04*, Cambridge, UK, Aug. 2004
- [27] Sang Hyun Kim and Rae-Hong Park, "An efficient algorithm for video sequence matching using the modified Hausdorff distance and the directed divergence," in *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 12 pp. 592-596, July 2002
- [28] R. Lienhart et al., "VisualGREP: A Systematic method to compare and retrieve video sequences," In *SPIE, Storage and Retrieval for Image and Video Database VI*, Vol. 3312, 1998
- [29] J. Oostveen et al., "Feature extraction and a database strategy for video fingerprinting," In *Visual 2002, LNCS 2314*, pp. 117-128, 2002
- [30] Jianping Fan et al., "Classview: hierarchical video shot classification, indexing and accessing," In *IEEE Trans. on Multimedia*, Vol. 6, No. 1, Feb. 2004
- [31] Chu-Hong Hoi et al., "A novel scheme for video similarity detection," In *Proc. of CIVR'03, LNCS 2728*, pp. 373-382, 2003
- [32] Akisato Kimura et al., "A Quick Search Method for Multimedia Signals Using Feature Compression Based on Piecewise Linear Maps," In *Proc. of ICASSP'02*, 2002