DCT-Based Videoprinting on Saliency-Consistent Regions for Detecting Video Copies with Text Insertion

Rong Yang[1], Yonghong Tian[2], and Tiejun Huang[2]

[1] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[2] National Engineering Laboratory for Video Technology, Peking University,
Beijing 100871, China
ryang@jdl.ac.cn, {yhtian,tjhuang}@pku.edu.cn

Abstract. Ideal video fingerprinting should be robust to various practical distortions. Conventional fingerprinting mainly copes with natural distortions (brightness change, resolution reduction, etc.), while always gives poor performance in case of text insertion. One alterative way is to apply a weighting scheme based on the probability of text insertion for feature similarity calculation. However, the weights must be learned with labeled samples. In this paper, we propose a method that first addresses valid regions where the saliency values keep consistent between the query and original frames, namely saliency-consistent regions. Other regions, probably the inserted ones, are discarded. Then a DCT-based hamming distance is calculated on those saliency-consistent regions. Besides, the saliency-based distance is also considered and a further weighted linear distance is evaluated. The proposed algorithm is tested on the MPEG-7 video fingerprint dataset, achieving a false rate of 0.7% in case of text insertion and 0.32% in average for other 8 distortions.

Keywords: Text insertion, saliency-consistent region, saliency map, discrete cosine transform (DCT), video copy detection.

1 Introduction

With growing broadcasting of digital video, the video copy detection has received a further attention recently. Generally, the video copy detection technique aims at deciding whether a query clip is a copy of a supervised video or not. There are two techniques relevant to the copy detection: watermarking and content-based copy detection (CBCD). The watermarking technique inserts discernable information into the supervised videos prior to distribution. However, effective watermarking algorithms, that can keep fidelity and meanwhile be robust against transmission attacks, are not available yet. While in CBCD, a brand new idea is brought forward, that is “the media itself is the watermark” [1]. Generally speaking, the CBCD technique consists of two main modules: feature extraction and feature matching modules. In the feature extraction module, a compact feature, which can identify the video and also be robust to various signal distortions, is extracted as the video fingerprint; in the feature matching one, a distance measurement is defined to quantify the similarity between two video fingerprints.

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 797–806, 2009.

There have been many fingerprint extraction algorithms proposed in the literature. All these algorithms can be broadly classified into two categories: The former considers a video clip as group-of-frames. In [2], the AC coefficients of a video clip are computed by 3-D DCT transformation and further quantified to a binary vector. The methods consider over the spatiotemporal quality of the video in its totality and always yield a more compact and fast-matching feature. However, it may fail if the video suffers from attacks of frame dropping or frame rate descending; besides, due to lose of detail temporal information of each frame, it is impossible to address the extract position of a copy clip. The latter category believes that a video clip is a set of continuing frame sequences. Individual feature is extracted for each frame; then features of continuing frames compose the indentified fingerprints of the whole video. Mohan [3] introduced an ordinal measure: Each frame is subdivided into blocks, then the average value of each block is calculated, and the rank of each block by the average is assigned as the fingerprint. Furthermore, Chen and Stentford [1] proposed to use temporal ordinal fingerprint. The difference lies in that the rank is performed between the same spatial locations along the temporal frames. Some other popular block-based approaches concentrate on differential luminance evaluation. Oostveen [4] developed a spatial-temporal fingerprint based on the block differential of luminance in spatial and temporal regions synchronously. In [5], a frame is partitioned into blocks and luminance gradient of each pixel is computed, then for each block, the centroid of gradient orientations (CGO) weighted by the amplitude is calculated. Besides the global block-based features above, another exciting research trend involves descriptors around local points of interest. [6] utilized the scale-invariant feature points (SIFT) to catch the most significant and stable quality of the whole frame. In [7], observing the dynamic behavior trends of points of interest along the video sequence, the author further assigned a label of behavior to each local descriptor, which provides more rich high-level semantic information. However, addressing points of interest always requires high computation complexity.

Although the fingerprint algorithms above have shown a significant effect for nature attacks (brightness change, resolution reduction, etc.), they always fail to cope with the malicious attacks like text insertion. The reason is that text insertion always partially changes the frame content and the feature values may also be changed. There are a few algorithms in the literature that take robustness against the distortion of text insertion into account. In [8], a block-based local feature representing the dominant type of edge direction was extracted, and then a weighting scheme based on the probability of caption superimposition for each block was applied to the similarity calculation. In more details, some samples of edited materials containing captions were selected as the training samples, and the probability of caption superimposition for each block was manually counted; then a linear model and a nonlinear logistic model were introduced to determine the final weights respectively. The disadvantages of the algorithm are that the weights totally depend on the training set; besides, it is difficult to obtain an accurate and appropriate weights distribution for a large amount of real distorted data by sample materials selection and manually counting.

This paper proposes a DCT-based fingerprinting on saliency-consistent regions, which aims at providing robustness to text insertion and also other video attacks. The rest of the paper is organized as follows: section 2 gives a detail description of the proposed algorithm: first, we give a broad overview of the whole framework, then the notion of saliency-consistent regions is explained; next, a DCT-based fingerprint is put forward, and in the fourth subpart, the similarity definition for the DCT-based fingerprinting on saliency-consistent regions is formulized. Experimental results on MPEG-7 video signature database are presented in section 3, and final conclusions and future work are addressed in section 4.

2 Proposed Algorithm

2.1 Overview

To be robust to the attack of text insertion, we propose a DCT-based fingerprint on the saliency-consistent regions for video copy detection. In [8], a weighting scheme based on the probability of text insertion is applied to the similarity calculation. However, the weights are acquired through an offline, learning process with labeled samples and manual counting. But in real situation, the distorted data is of large amount, and the insertion locations are distributed randomly. Therefore it is difficult to acquire the probability weights by samples selection and then offline labeling and counting.

We use the saliency consistency to estimate the potential locations with text inserted, mainly based on two facts: on one hand, the malicious inserted texts always differ from the original content in aspect of color, intensity or orientation. On the other hand, saliency map represents the conspicuous quality of the local units. As a result, once a region of the query frame is of saliency quality while the corresponding region of the original is un-salient, we infer that a text insertion may appear and the

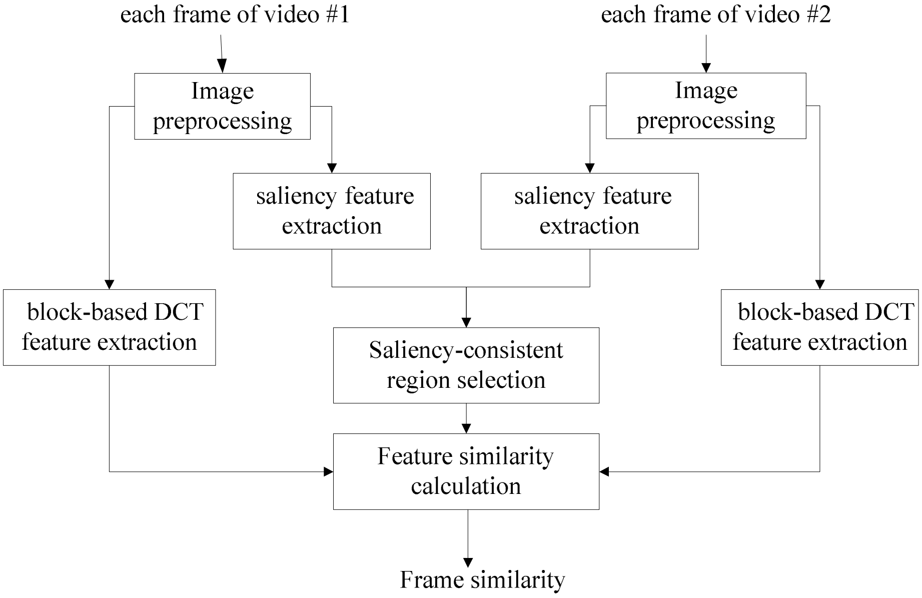

each frame of video #1
each frame of video #2
Image preprocessing
Image preprocessing
saliency feature extraction
saliency feature extraction
block-based DCT feature extraction
Saliency-consistent region selection
block-based DCT feature extraction
Feature similarity calculation
Frame similarity

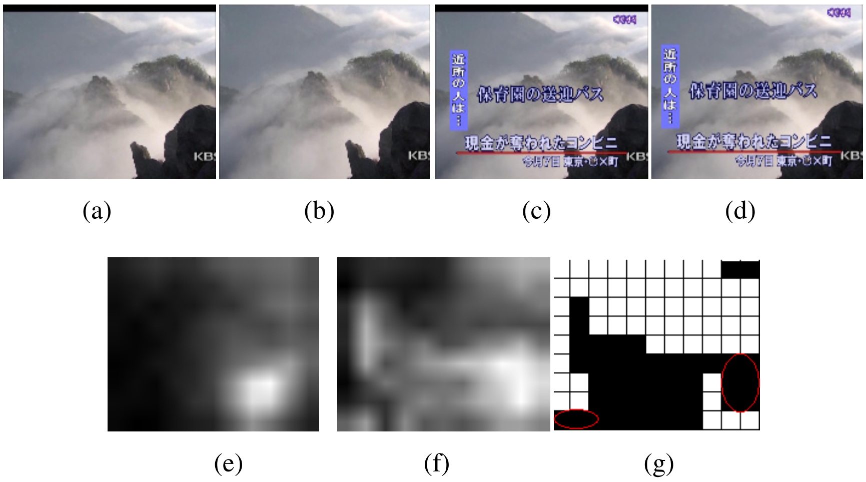
Fig. 1. Framework of the proposed frame similarity computation

region is marked invalid. For regions that keep consistent in saliency quality (salient or un-salient) between the query and original frames, there is a great probability that most original contents are reserved in the query. Those regions are marked as valid regions or so called saliency-consistent regions. DCT feature, which has been shown effective [1] [9] for copy detection, is computed on those saliency-consistent regions as the measurement of frame similarity.

Figure 1 illustrates the framework of the frame similarity computation. In the image preprocessing step, a sequence of operations are applied to eliminate the difference brought by various display formats, like conversions to resolutions, frame rate, letter-box and pillar-box formats, etc. Then the saliency maps of two compared frames are calculated and utilized to address the saliency-consistent regions. Finally the DCT-based hamming distance on those saliency-consistent regions are computed as the frame similarity. Moreover, considering the discrimination and robustness of saliency, the saliency-based hamming distance is evaluated and appended to the total frame similarity.

2.2 Detecting Saliency-Consistent Regions

The essence of saliency lies in that only locations whose local visual statistics significantly differ from their surrounding locations statistics can be "focus of attention". Itti [10] proposed a visual attention system driven by saliency, bottom-up mechanism. In the framework, an image is first analyzed at multiple spatial scales, giving rise to a number of feature maps referring to sub-elements of color, intensity and orientation. Within each of the feature maps, locations that significantly differ from their neighbors are highlighted. Finally, all highlighted locations from all feature maps are combined into a master saliency map, which represents the conspicuity quality over the whole image.



(a) (b) (c) (d)

(e) (f) (g)

Fig. 2. Saliency-consistent region selection: (a) the original frame; (b) the preprocessed frame of original; (c) the query frame with text insertion; (d) the processed frame of query; (e) saliency map of (b); (f) saliency map of (d); (g) saliency-consistent regions (white)

In this paper, saliency is utilized to address the potential text insertion regions. To convert the continuous values of saliency map to a binary quality (salient or un-salient), we select a certain proportion of the most salient regions as the potential inserted regions and mark them 1,while others are set as 0. An intuitive idea is to just consider un-salient regions that are definitely not text inserted regions for further computation. However regions which themselves are salient due to the original content wound be discarded. So a saliency-consistent region selection is introduced. That is, if the corresponding regions in both the query and original frames are marked the same value, they are selected as saliency-consistent regions.

Figure 2 represents a demonstration of saliency-consistent region selection. It shows that the regions with text inserted are marked (black) correctly. But some regions (red ellipse) that contain useful information are also marked as potential inserted regions. The phenomenon happens in regions that are close to the inserted ones, because the text inserted region would also influence the saliency of its neighbors. But still, the saliency-consistent regions marked can provide enough information about where the DCT-based hamming distance should be considered.

2.3 DCT-Based Feature

The fingerprint based on DCT was originally developed by Kim [9] for image copy detection. First, an image is divided into 8*8 equal-sized blocks; then the average intensity of each block is derived. The resulting 64 intensity values are transformed into a series of coefficients by performing an 8*8 2D DCT, and finally the coefficients are ranked by the AC magnitudes. The rank matrix is the image signature.

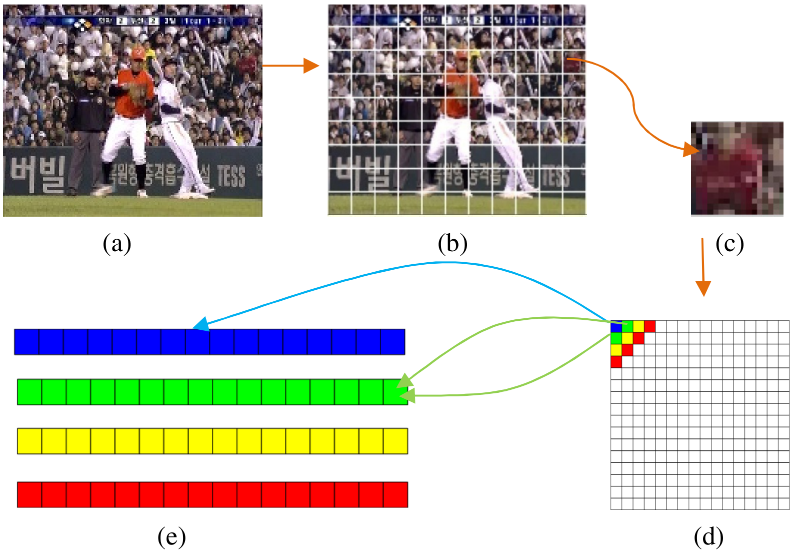


Fig. 3. The main framework of DCT feature extraction: (a) the enhanced frame by histogram equalization. (b) the frame with N blocks. (c) a block example with size of 16*16. (d) the DCT coefficients of one block with blue, green, yellow, red marking the first to fourth lower order coefficients respectively. the average of each order (the same color) is calculated. (e) N averages of each order are ranked respectively to generate the fingerprint.

An improved DCT feature is developed: First, a frame is enhanced by histogram equalization and then partitioned into N blocks. Next, a 2D DCT is performed to each block and the first 4th lower order coefficients (shown in Figure 3) are averaged to 4 coefficients accordingly. Then N coefficients of the same order are quantified to a binary vector with a median threshold. As a result, for each frame, an N*4-dimension binary vector is generated as the DCT fingerprint.

2.4 Frame Similarity Definition

Let x , y denote two compared frames, and the corresponding features are represented as:

$$\mathrm{d_x = [d_{x1}, d_{x2}, \cdots d_{xM}]}\ ,\ \mathrm{d_y = [d_{y1}, d_{y2}, \cdots d_{yM}]}\ .$$

$$\mathrm{a_x = [a_{x1}, a_{x2}, \cdots a_{xM}]}\ ,\ \mathrm{a_y = [a_{y1}, a_{y2}, \cdots a_{yM}]}\ .$$

where M is the number of blocks, and symbol "d" and "a" indicate the DCT and saliency features respectively. The frame similarity $\mathrm{s(x,y)}$ is calculated by simply comparing the DCT feature $\mathrm{d_{xi}}$ and $\mathrm{d_{yi}}$ on the saliency-consistent regions as shown below:

$$\mathrm{s(x,y)} = \textstyle\sum_{i=1}^{M} \delta(\mathrm{a_{xi}, a_{yi}}) * \mathrm{I}(\mathrm{d_{xi}, d_{yi}}) \tag{1}$$

where $\mathrm{I}(\mathrm{d_{xi}, d_{yi}})$ calculates the simple DCT hamming distance for block i , and

$$\delta(\mathrm{a_{xi}, a_{yi}}) = \begin{cases} 1 & \text{if } \mathrm{a_{xi} = a_{yi}} \\ 0 & \text{if } \mathrm{a_{xi} \neq a_{yi}} \end{cases} \tag{2}$$

where the weight value $\delta(\mathrm{a_{xi}, a_{yi}})$ indicates whether the block i is a saliency-consistent one or not. A further normalization is performed on $\mathrm{s(x,y)}$ as equation 3 shows:

$$\mathrm{s(x,y)} = \frac{M}{K}\textstyle\sum_{i=1}^{M} \delta(\mathrm{a_{xi}, a_{yi}}) * \mathrm{I}(\mathrm{d_{xi}, d_{yi}}) \tag{3}$$

where K denotes the number of $\delta(\mathrm{a_{xi}, a_{yi}})$ marked 1.

First, considering the case of text insertion, the regions where texts are inserted would be marked salient, while the corresponding regions of the original frame have great possibility that marked un-salient. The reason lies in that texts are always inserted into the un-conspicuity regions in real life to prevent the degradation of user perception. Then considering the case of other distortions, the saliency map of the query would be similar with the original one, and most of DCT information would be reserved and compared. Lastly, considering two videos from different sources, the saliency map would be statistically independent, and the saliency-consistent regions may be selected randomly. But because two video contents are greatly different, the sufficient discrimination would still be expected optimistically.

Furthermore, taking the discrimination and robustness of saliency into account, the saliency-based hamming distance is also computed and appended to the frame similarity:

$$s(x,y) = \frac{M}{K}\sum_{i=1}^{M} \delta\left(a_{xi},a_{yi}\right) * I\left(d_{xi},d_{yi}\right) + \mu \sum_{i=1}^{M} \delta(a_{xi},a_{yi}) \qquad (4)$$

where μ is a weight that governs the relative proportion of additional saliency-based distance compared to the simple DCT-based fingerprint distance on saliency-consistent regions. The distance in equation 4 is called the weighted linear distance below.

Both DCT-based fingerprint matching on saliency-consistent regions (we call it DCT_SC for short) and the linear weighted distances are tested.

3 Experiments

3.1 Database and Evaluation Measurement

Experiments are carried out on the MPEG-7 video signature database [11]. 545 videos are selected from a total of 1900 3-minutes-long original videos. 6 query scenarios (2s, 5s, 10s; partial 2s, 5s, 10s) are involved and 9 kinds of transformations are evaluated as shown in Table 1. A total number of 19620 independent query videos and 71940 robust query videos are looked up through all original videos database to evaluate the performance of the proposed algorithm.

Table 1. Modifications and Levels (9 modifications, 22 categories) for robustness test [11]

level Modifications	coding	heavy	medium	light
Text/logo overlay	MPEG-2	30%	20%	10%
Sever compression(at CIF resolution)	AVC	64kbps	256kbps	512kbps
Resolution reduction	MPEG-2	-	QCIF	CIF
Frame-rate reduction	AVC	4fps	5fps	15fps
Capturing on camera	MPEG-2	10%	5%	0%
Analog VCR recording & recapturing	MPEG-2	3 times	2 times	1 times
Color to monochrome conversion	MPEG-2	-	-	I=0.299×R + 0.587×G + 0.114×B
Brightness change	MPEG-2	+36	+18	+9
Interlaced/Progressive conversion	MPEG-2	-	-	P→I→P->I→P

False negative (R_{fn}) and false positive (R_{fp}) rates are used to evaluate the discriminability and robustness of the algorithm:

$$R_{fn} = \frac{N_{fn}}{N_{ep}} * 100 \quad (5)$$

and

$$R_{fp} = \frac{N_{fp}}{N_{en}} * 100 \quad (6)$$

whereN_{fn} is number of false negatives and N_{ep} is number of expected positive;N_{fp} is number of false positives and N_{en} is number of expected negatives.

3.2 Experimental Results

To evaluate the performance of the proposed algorithm, the CGO (centroid of gradient orientations) fingerprint in [5] is implemented as contrast. We focus on the most severe scenario in which the shortest video clips (2s) are queried and attacks of the heaviest level for all 9 distortions are under consideration.

Table 2 presents the performance of the proposed DCT_SC distance and linear weighted distance. False negative rate and false positive rate against the threshold are plotted. The balance point, where the false negative rate equals to the false positive rate, is served as the measurement of the fingerprint.

On one hand, in case of the distortion of pattern insertion, the DCT_SC and linear weighted distances proposed achieve a false rate of 0.70% and 0.36% respectively, while the CGO fingerprint just achieve a false rate of 9%. On the other hand, the proposed method also keeps a lower false rate of approximately 0.20% in case of other 8 distortions. It is obviously that DCT_SC shows a better performance than CGO for all 9 distortions. Although DCT_SC and the linear weighted model both outperform the CGO fingerprint, the effect of additional saliency similarity is not so significant. More details of the result are shown in Figure 4.

Table 2. Comparision of two proposed distance and CGO

methods Modifications	CGO[5]	DCT_SC	The weighted linear distance
Text/logo overlay	7.2%	0.70%	0.36%
Sever compression	12.0%	0.88%	0.62 %
Resolution reduction	3.1%	0.09%	0.006%
Frame-rate reduction	1.8%	0.06%	0.002%
Capturing on camera	12.1%	0.13%	0.07%
Analog VCR recording & recapturing	7.6%	0.31%	0.12%
Color to monochrome conversion	1.7%	0.09%	0.007%
Brightness change	1.4%	0.04%	0.008%
Interlaced/ Progressive conversion	3.1%	0.13%	0.018%

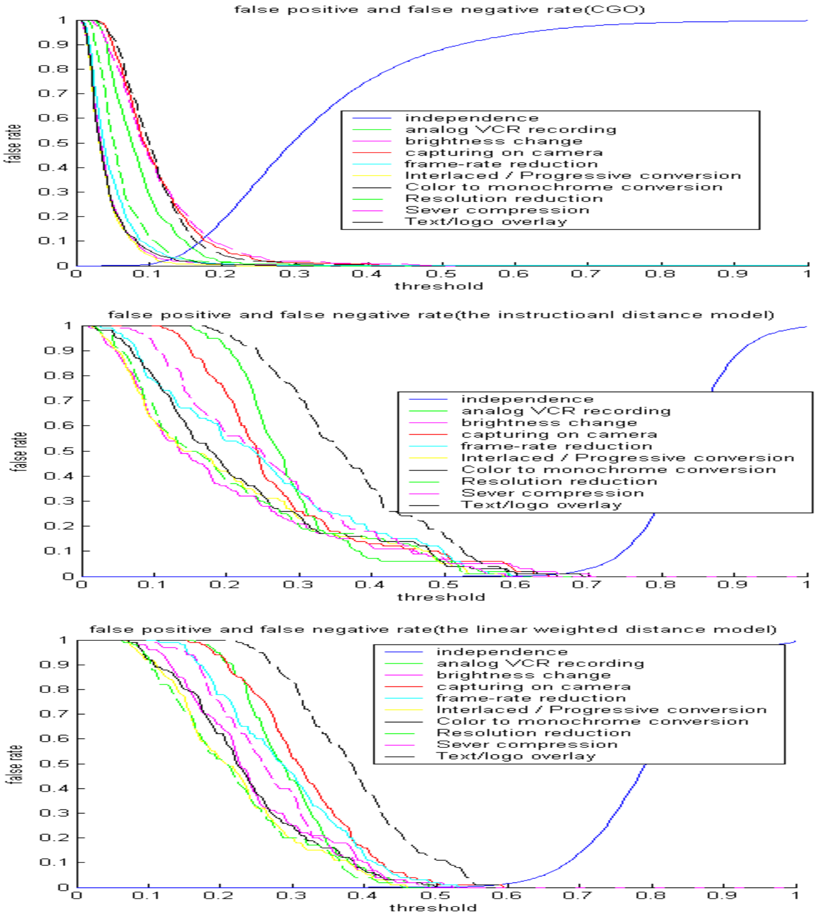


Fig. 4. Performances of 3 video fingerprint matching: the false negative rate against the threshold is plotted in blue; the false positive rates of 9 distortions are also plotted. First sub-image: CGO; second sub-image: DCT_SC; third sub-image: the weighted linear distance.

4 Conclusion and Future Work

In this paper, a DCT-based fingerprinting on saliency-consistent regions is proposed to resist against the distortion of text insertion in video copy detection. The regions between the query and original frames, where saliency quality are the same, are selected as saliency-consistent regions. A DCT-base hamming distance is calculated on those regions to measure the frame similarity. Experiments on MPEG-7 video signature database show that our proposed method achieves a better performance than CGO.

It is found that some text insertions are not in salient regions because the original content contains more conspicuous object. Considering that the visual saliency system implemented is a generalized framework for rapid target detection, how to bias saliency to texts' particular low-level features will be a further question we may be interest in.

Acknowledgement

The work is supported by grants from Chinese NSF under contract No. 60605020 and No. 90820003, National Hi-Tech R&D Program (863) of China under contract 2006AA010105, and National Basic Research Program of China under contract No. 2009CB320906.

References

1. Chen, L., Stentiford, F.W.M.: Video Sequence Matching based on Temporal Ordinal Measurement. Pattern Recognition Letters 29, 1824–1831 (2008)
2. Coskun, B., Sankur, B., Memon, N.: Spatio-Temporal Transform Based Video Hashing. IEEE Transactions on Multimedia 8, 1190–1208 (2006)
3. Mohan, R.: Video Sequence Matching. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 6, pp. 3697–3700 (1998)
4. Oostveen, J., Kalker, T., Haitsma, J.: Feature Extraction and a Database Strategy for Video Fingerprinting. In: Chang, S.-K., Chen, Z., Lee, S.-Y. (eds.) VISUAL 2002. LNCS, vol. 2314, pp. 117–128. Springer, Heidelberg (2002)
5. Lee, S., Yoo, C.D.: Robust Video Fingerprinting for Content-based Video Identification. IEEE Trans. Circuits Syst. Video Technol. 18, 983–988 (2008)
6. Sarkar, A., Ghosh, P., Moxley, E., Manjunath, B.S.: Video Fingerprinting: Features for Duplicate and Similar Video Detection and Query-based Video Retrieval. In: Proc. SPIE-Multimedia Content Access: Algorithms and Systems, vol. 6820 (2008)
7. Law-To, J., Buisson, O., Gouet-Brunet, V., Boujemaa, N.: Robust Voting Algorithm based on Labels of Behavior for Video Copy Detection. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, Santa Barbara (2006)
8. Iwamoto, K., Kasutani, E., Yamada, A.: Image Signature Robust to Caption Superimposition for Video Sequence Identification. In: International Conference on Image Processing, pp. 3185–3188 (2006)
9. Kim, C.: Content-based Image Copy Detection. Signal Processing: Image Communication 18, 169–184 (2003)
10. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-based Visual Attention for Rapid Scene Analysis. IEEE Patt. Anal. Mach. Intell., 1254–1259 (1998)
11. Bober, M., Brasnett, P., Iwamoto, K.: Description of Core Experiment for MPEG-7 Visual Descriptors, `http://www.chiariglione.org/mpeg/working_documents/mpeg-07/visual/visual_ce.zip`