PAPER

An Improved Video Identification Scheme Based on Video Tomography

Qing-Ge JI[†], Zhi-Feng TAN[†], *Nonmembers*, Zhe-Ming LU[††a)], *Member*, *and* Yong ZHANG[†††], *Nonmember*

SUMMARY In recent years, with the popularization of video collection devices and the development of the Internet, it is easy to copy original digital videos and distribute illegal copies quickly through the Internet. It becomes a critical task to uphold copyright laws, and this problem will require a technical solution. Therefore, as a challenging problem, copy detection or video identification becomes increasingly important. The problem addressed here is to identify a given video clip in a given set of video sequences. In this paper, an extension to the video identification approach based on video tomography is presented. First, the feature extraction process is modified to enhance the reliability of the shot signature with its size unchanged. Then, a new similarity measurement between two shot signatures is proposed to address the problem generated by the original approach when facing the query shot with a short length. In addition, the query scope is extended from one shot only to one clip (several consecutive shots) by giving a new definition of similarity between two clips and describing a search algorithm which can save much of the computation cost. Experimental results show that the proposed approach is more suitable for identifying shots with short lengths than the original approach. The clip query approach performs well in the experiment and it also shows strong robustness to data loss.

key words: *video tomography, video signature, shot detection, video clip query*

1. Introduction

With the popularization of video collection devices and the perfection of the Internet's basic facilities, video information on the web has increased in a geometric progression. Video copy detection, which is also referred to as video identification, has become an important problem for copyright holders and media distributors due to the rapid development of many sorts of digital multimedia data operations including producing, processing and copying. The growing popularity of many kinds of video sharing web sites like YouTube, where huge amounts of video sequences are stored and spread, has intensified the requirement for controlling the copyright and video content. Aside from copy detection, such challenge also gives birth to new research areas such as multimedia indexing and multimedia content retrieval where MPEG has issued a call for proposals on video signature standardizing [1].

The main ways of solving this problem can be broadly classified into two categories: digital watermark based and content based. Digital watermark based methods determine the source of a video dependent on extracting the embedded watermark in the video. Doer et al. first proposed a watermark based solution for identification and tamper detection [2]. Although digital watermark based methods are useful for identifying video sources, they are not designed to discriminate unique clips from the same video. The robustness of the embedded watermark in the video and the existence of numbers of video sequences without watermarks are two major drawbacks to digital watermarking. Such difficulties are being addressed in an emerging research area called blind detection [3], [4]. Blind detection based methods are also used for tampering detection and source identification like digital watermarking based methods but with the characteristics inherent to the video and capture devices. Sevinc et al. proposed an approach to detect duplicate and modified copies of a video by extracting the noise generated by the imaging sensors that serve as device peculiarity [5]. However, neither digital watermarking nor blind detection is suitable for video copy detection or identification.

Content based approaches, on the other hand, which exploit the content of the video only to generate a unique signature based on video features without requiring any embedded watermark or device information, have received more and more interest lately. Two survey papers on content based identification systems were presented by Fang et al. [6] and Law-To et al. [7]. In Reference [8], the bag-of-words model formally used in text retrieval was applied to copy detection. This approach introduces the SIFT descriptors, which are robust to transformations such as brightness variations, as words to create a SIFT histogram for later matching. Yan et al. adopted a composite of the fingerprints extracted from individual frames in similarity calculation for copy detection in streaming video sequences [9]. The clustering and key frame analysis techniques are used in [10] where key frames for each cluster of the query are extracted and then a key frame based search for similarity regions in the target videos is performed. Key frame analysis is also used in [11] where key frames are extracted to match against a database and then the local spatial-temporal features are adopted to match the video sequences.

As can be inferred from the above, the video signatures adopted by many content based identification methods are created from individual frame content. This requires much computation cost, especially in long video sequences as the

Manuscript received July 2, 2013.

Manuscript revised November 8, 2013.

[†] The authors are with School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China.

[††] The author is with School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310027, China.

[†††] The author is with ATR National Defense Technology Key Laboratory, College of Information Engineering, Shenzhen University, Shenzhen, 518060, China.

a) E-mail: zheminglu@zju.edu.cn (Corresponding author)

DOI: 10.1587/transinf.E97.D.919

feature extraction and comparison operations on a frame basis are needed. Recently, the approach based on video tomography [12] proposes an alternative solution to the problem and it is the idea that this paper focuses on. The rest of this paper is organized as follows. In Sect. 2, more details about video tomography are introduced. In Sect. 3, the proposed approach based on the improved video tomography technique is proposed. In Sect. 4, experimental results and comparisons with the original video tomography technique are given. In Sect. 5, conclusions are drawn and the future work is suggested.

2. Related Works

Since the proposed approach is based on video tomography, this section introduces the concept of video tomography together with its original usage in video identification.

2.1 Tomography Video Signatures

Video tomography was first presented as a way to extract camera information such as lens zoom, camera pan and camera tilt information for camera work identification in movies by Akutsu and Tonomura [13]. Since then such method has been explored for summarization and camera work detection in movies [14]. In video tomography, a fixed line is extracted from each frame of the Y component of a video shot. These lines are sequentially arranged to generate a tomography image. Figure 1 illustrates the process of creating a tomography image. And the method presented in [15] is described below to show how the tomography technique is applied to video identification.

First, the video is scaled to the resolution 360 × 240 (Here, some frame size normalization methods for this step can be found in [16]). Second, for every shot, six different patterns shown in Fig. 2 are selected for tomography generation, where two upper diagonals labeled ‘1’ and ‘2’, two lower diagonals labeled ‘3’ and ‘4’ and two regular diagonals labeled ‘5’ and ‘6’. Then, each tomography image is processed through the Canny edge detector to get the edges for revealing the patterns with high spatial-temporal correlations. Subsequently, both diagonals of each set are superimposed to create a composite edge image by using the OR operation.

The number of level changes (i.e. the black to white transitions) on these three composite edge images is counted on 8 specific horizontal and 8 specific vertical lines, which are equally spaced along the image. Figure 3 gives an example of this step. The 16 counts on each of the three composites are produced and they are combined to form a 48 byte signature for a shot regardless of the number of frames in it.

After signatures have been generated for all shots through the process above, the similarity of two shots can be measured by comparing signatures based on the Euclidean distance between two points in the 48-dimensional space:

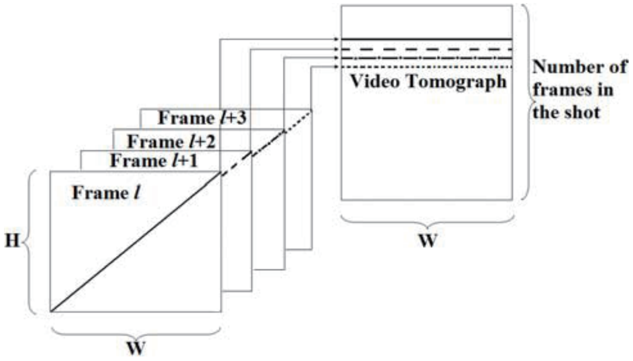


Fig. 1 Video tomography performed on a video shot with frame size W × H.

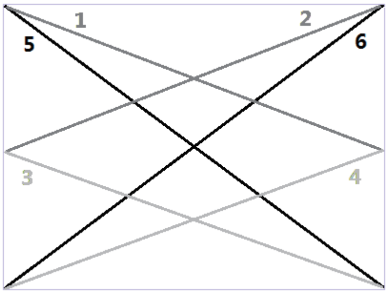


Fig. 2 Six different tomography line patterns.

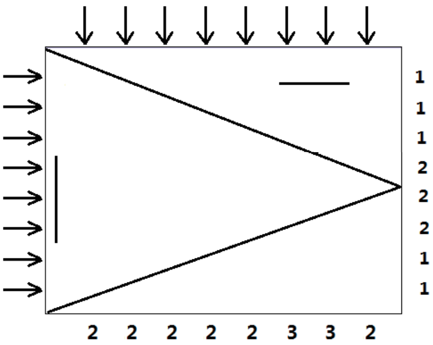


Fig. 3 Level changes measured at 8 evenly distributed horizontal and vertical lines.

$$D = \sqrt{\sum_{i=1}^{48} (a_i - b_i)^2} \tag{1}$$

where a_i and b_i denote the i-th components in shot signatures of A and B, respectively. Here, A denotes the query shot and B stands for the shot to be compared with.

By adopting the same technique, we can also generate the frame tomography signatures. For each pattern line of a frame, it is divided into 4 segments and the number of edges is counted among them, resulting in a 24-dimensional signature as shown in Fig. 4. Besides, no extra video operation is required for frame signature calculation as it can be performed on the same tomography image used to generate the shot signature. The Euclidean distance can also serve as the similarity measurement of frame signatures.

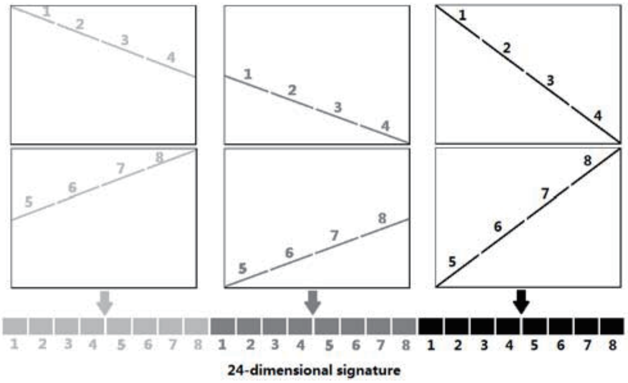


Fig. 4 Frame tomography signature.

2.2 Query Matching Process

With all the required shot signatures and frame signatures in hand, the search for the best matching video clip can be performed in the following two stages:

Stage 1: Closest shot identification. In this stage, we first use shot signature of the query to compare with all the shot signatures in the database, and then we choose a set of shots with 30 smallest similarities.

Stage 2: Closest frame identification. In order to identify the precise location of the query, a frame evaluation is required. Such evaluation is finished by calculating and comparing the frame signatures for the video clip and the query. The distance between two frame signatures is averaged over all candidate shots selected through the previous stage. And the one which has the lowest average value is the final matched video clip.

2.3 Shot Detection

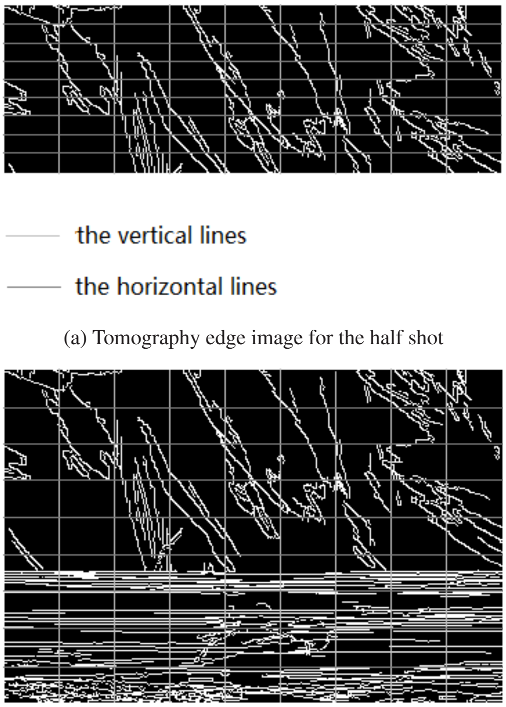
It should be noted that, before shot signature generation, the shot detection is a necessary step. The crater distance method based on tomography is naturally applied. By taking three consecutive frames each time, the Euclidean distance between frame signatures is calculated and a depression pattern (a high-low-high distance value) is searched. And a shot is declared if such pattern complies with a threshold value.

3. The Proposed Approach

To achieve the improvement in accuracy, we propose improved feature extraction and signature comparison processes in this section. Furthermore, we extend the query scope for more applications. The proposed approach can be described in detail as follows.

3.1 Modified Feature Extraction

Since the tomography signature does not reflect the duration of shot, the similarity between a shot and a portion of it will decrease dramatically as the length of the portion decreases.



(a) Tomography edge image for the half shot

(b) Tomography edge image for the whole shot

Fig. 5 Comparison of the feature extraction result between the whole shot and its half portion.

One important reason for this result is the way of feature extraction described in Sect. 2.1. An example is shown in Fig. 5 to compare the tomography images obtained from the half shot and the whole shot. We can see that each of the vertical line in Fig. 5 (a) is a part of the corresponding line in Fig. 5 (b). This leads to a small Euclidean distance, therefore, correctly represents the high correlation between a shot and its portion. However, it is not the same case when considering the horizontal lines. What's worse, the correlation between the horizontal line in Fig. 5 (a) and the corresponding line in Fig. 5 (b) will be even lower as the length of the portion decreases.

With the consideration of this situation, we modify the feature extraction process by canceling 8 horizontal lines and adding another 8 vertical lines. Now every edge image is counted on 16 specific vertical lines which are evenly distributed along the image. The shot signature is kept 48 byte unchanged but the reliability increases. It is not so meaningful to show the superiority of our scheme to the original one solely as it will work much better with the modified signature comparison process described below. So only the performance of the combination of these two approaches will be shown in Sect. 4.

3.2 Modified Signature Comparison

Feature extraction is affected by not only the length of the query shot but also the shot signature comparison approach. As the length decreases, each of the number of edges will decrease accordingly, and the distance will increase as a result. To achieve a more reliable signature comparison, two restrictions are introduced and a modified signature comparison approach is proposed in this sub-section.

Let A denote the query shot, B denote the shot to be compared with, M denote the best matched shot, and L_A and L_B denote the lengths of A and B, respectively. The proposed two restrictions are given as below.

Restriction 1: Since A is a portion of M (including M), the length of A should not be larger than that of the shot B to be compared with, i.e., $L_A \leq L_B$.

Restriction 2: The tomography image of A is also a portion of that of M, so each of the number of edges counted on vertical lines should not be larger than the corresponding one compared with.

With the consideration about different qualities of the video, however, there may be slight change in the tomography image even the contents in both shots are the same. Thus we enlarge the upper bound of the number of edges to decrease the probability of discarding the best matched shot because of Restriction 2. Explicitly, we change the upper bound of a_i from b_i to $U_i = b_i + \text{round}(2 \times (L_A/L_B)^6)$. Thus, if $L_A = L_B$, we increase the upper bound of a_i by 2 (this value is determined experimentally which is according to the influence of the common distortions to the video). Otherwise, U_i will quickly drop to b_i as L_A/L_B decreases due to the power function such that more unrelated shots can be excluded. In a word, Restriction 2 can be expressed simply as $a_i \leq U_i$.

In our approach, if any restriction is violated during the process of signature comparison, the comparison will be stopped immediately. In this case, a predefined large value N_R is set as the distance between the two shots under comparison, which denotes the low correlation between them. On the contrary, if all the restrictions are satisfied, the similarity between the two shots under comparison is calculated through the following improved approach.

In our scheme, instead of directly calculating the difference between a_i and b_i, a value e_i estimated from a_i is constructed to perform the same comparison with b_i. The construction of e_i is done with the consideration of the information containing in both A and B as follows:

$$e_i = a_i + [w_i \cdot da_i + (1 - w_i) \cdot db_i] \cdot (L_B - L_A) \tag{2}$$

where w_i denotes the weight of da_i, and da_i and db_i denote the densities of the number of edges distributed along the i-th line in A and B respectively, i.e.,

$$da_i = \frac{a_i}{L_A},\ db_i = \frac{b_i}{L_B} \tag{3}$$

There are several choices for w_i. For example, $w_i = 1$ or $w_i = 0$ is not a good choice as neither of them considers the information from both shots, while $w_i = 1/2$ may be a proper one. In our method, a dynamic weight is adopted as follows

$$w_i = 1 - \frac{L_A}{L_B} \tag{4}$$

To show the superiority of our weight over the constant weight 1/2, we give some detailed explanations as follows. Assume $L_B/L_A = k$ $(k \geq 1)$, if $w_i = 1/2$ is chosen, then we can obtain

$$\begin{aligned}(e_i - b_i)^2 &= \left[a_i + \left(\frac{da_i}{2} + \frac{db_i}{2}\right) \cdot (L_B - L_A) - b_i\right]^2 \\ &= \left[\frac{k+1}{2} \cdot L_A \cdot (db_i - da_i)\right]^2\end{aligned} \tag{5}$$

Similarly, if our weight in Eq. (4) is chosen, then we can obtain

$$\begin{aligned}&(e_i - b_i)^2 \\ &= \left\{a_i + \left[\left(1 - \frac{L_A}{L_B}\right) \cdot da_i + \frac{L_A}{L_B} \cdot db_i\right] \cdot (L_B - L_A) - b_i\right\}^2 \\ &= \left[\frac{k^2 - k + 1}{k} \cdot L_A \cdot (db_i - da_i)\right]^2\end{aligned} \tag{6}$$

Subtracting Eq. (5) from Eq. (6), we can obtain the difference:

$$\left[\frac{k(3k-1)+2}{4k^2}\right](k-1)(k-2)L_A^2(db_i - da_i)^2 \tag{7}$$

In order to let the difference be not less than 0, we can easily obtain the requirement $k \geq 2$, which means that the distance value is smaller by using our weight given in Eq. (4) if L_A is larger than half of L_B, and larger in the opposite case. It is just the property we need. When L_A/L_B is large, the similarity calculated between shot signatures is more reliable so we hope the distance is smaller to reveal such correlation. On the contrary, when L_A/L_B is small, we hope the distance become larger to reveal the unreliable of the shot signature and a more exact similarity generation process can be done.

Based on the above discussion, the improved distance between two shot signatures is defined as follows

$$D_{\text{new}} = \frac{\sqrt{\sum_{i=1}^{48} (e_i - b_i)^2}}{R_S} \tag{8}$$

Where R_S is a constant which is used to make the similarity value calculated between shot signatures be close to the value calculated between frames. In fact, Eq. (6) also provides a faster way for calculating D_{new} as follows

$$\begin{aligned}\sum_{i=1}^{48} (e_i - b_i)^2 &= \sum_{i=1}^{48} \left[\frac{k^2 - k + 1}{k} \cdot L_A \cdot (db_i - da_i)\right]^2 \\ &= \left(\frac{k^2 - k + 1}{k}\right)^2 \cdot L_A^2 \cdot \left(\frac{\sum_{i=1}^{48} a_i^2}{L_A^2} + \frac{\sum_{i=1}^{48} b_i^2}{L_B^2} - 2 \cdot \frac{\sum_{i=1}^{48} a_i b_i}{L_A \cdot L_B}\right)\end{aligned} \tag{9}$$

Based on Eq. (8) together with Eq. (9), after all shots in the database have been searched, a list of shots with N_S

smallest distances is stored. If the least similarity in the list is smaller than the predefined threshold value TH_S, then a reliable similarity has been found. Otherwise, the shot signature is considered to be unable to provide enough information for generating reliable similarity due to its short length. Then the closest frame identification process described in Sect. 2.2 is applied to every shot in the list in the ascending order of D_{new}. For each shot, if the similarity calculated between frames is smaller than the one calculated using Eq. (8), the shot similarity is replaced with the former one. Once a similarity calculated between frames is smaller than TH_S, the frame-based comparison is stopped for the best matched shot is thought to be found.

3.3 Extension of Query Scope

The approach described in Sect. 2 is applied to the "shot to whole" scenarios. In this sub-section, we would like to extend the scope to "clip to whole" setting, that is, matching a sequence of consecutive shots to a video.

Assume the number of shots in the query clip C is n ($n \geq 1$), G denotes the clip to be compared with, and LC_i and LG_i denote the lengths of the i-th shot in C and G, respectively. Then the similarity between clips C and G is defined as:

$$S_C = \sum_{i=1}^{n} \frac{LC_i}{LG_i} \cdot S_i \tag{10}$$

where

$$S_i = e^{-D_i} \tag{11}$$

Here, D_i is the distance between the i-th shot of C and the i-th shot of G calculated by the approach described in Sect. 3.2. Thus, S_C is the sum of all the shot similarities in the clip while the ratio LC_i/LG_i serves as the weight. Obviously, the smaller the ratio is, the more unreliable the similarity is and thus the less contributions it should make to the sum of the whole similarity.

According to the whole similarity defined above, a clip searching strategy is proposed to save computation costs without losing much accuracy than the exhaustive search.

Step 1: For each shot in C, we calculate the similarity with all the "appropriate" shots in the database (e.g., assume there are seven shots in C, if the current query shot in C is the third shot as well as the last shot but four, then the first two shots and the last four shots in the video to be compared should be excluded).

Step 2: Every shot in C owns a list of N_S candidate shots according to Sect. 3.2. Based on these n lists, all candidate sequences of consecutive shots can be constructed by the following substeps:

Step 2.1: Find out the shot with the least similarity in all the$n$ lists. Assume we find Shot q of Video p in the database having the least similarity with Shot j in C, then the shot to be searched for each shot i ($1 \leq i \leq n$) in C is Shot $q + (i - j)$ of Video p.

Step 2.2: For each shot i in C, if we can find Shot $q + (i - j)$ of Video p in its list, we record the corresponding similarity value as D_i and delete this item from the list. Otherwise, we set $D_i = N_R$. Thus, we can construct a sequence of consecutive shots as well as obtaining the corresponding sequence of D_i values.

Step 2.3: Based on the obtained D_i values, Eqs. (10) and (11) are used to calculate the whole similarity between the obtained consecutive shot sequence and the query clip C.

Step 2.4: If all the n lists are empty or all the remaining similarity values in all lists are equal to N_R, go to step 3. Otherwise, go to step 2.1 to construct the next sequence of consecutive shots.

Step 3: Among the obtained sequences, the best matching clip is the one with the largest whole similarity S_C. In our approach, if S_C is smaller than a predefined threshold TH_Q, the query clip is not thought to exist in the database.

Before showing how the matrix $\mathbf{X}$ is factorized, we first show how it is constructed. Each $\mathbf{x}_i$ is the feature vector extracted from the i-th image, and it is generated as follows. First, the i-image is represented using the HSV color model. Then the V component of the image is extracted for histogram calculation with 32 bins. Finally, the calculation result is stored to form a 32-dimensional feature vector $\mathbf{x}_i$. The reason that we choose such feature extraction scheme in our approach lies in two aspects. First, we adopt the histogram information other than the image pixels since the former is more robust to some content preserved distortions such as translation and rotation. Second, as the LNMF procedure needs lots of computation time, only the most significant component of the image is selected for feature extraction so as to save computation cost.

4. Experimental Results

To evaluate the performance of the proposed approach, a similar scenario to the one presented in [15] is used. Reference [15] adopted a database including 1883 three minute clips. Each clip is divided into six 30 second segments and three query videos are created using the first 2, 5, and 10 seconds of each 30 second segment. Three additional queries are created by inserting these 2, 5, and 10 second segments in a 30 second video that is not in the database. In this paper, our database consists of 100 video sequences with durations varying from 3 to 10 minutes including different contents such as news, sports, cartoons, commercials, and movies. Each video sequence is divided into several segments of length 30 seconds and each segment is considered as a shot artificially. In our experiment, the parameters described in Sect. 3 are given in Table 1. All the values of the parameters are determined experimentally. For N_R, we just define a value which is larger than the possible maximum distance between shots. For N_S, we should consider the tradeoff between complexity and reliability, thus we select N_S =5. For R_S, we should select a suitable value close to 48, thus we adopt R_S = 50.0. TH_S and TH_Q should be small, and we

Table 1 The parameters used in the experiment.

Parameter	Value
N_R	10000000.0
R_S	50.0
N_S	5
TH_S	1.0
TH_Q	0.2

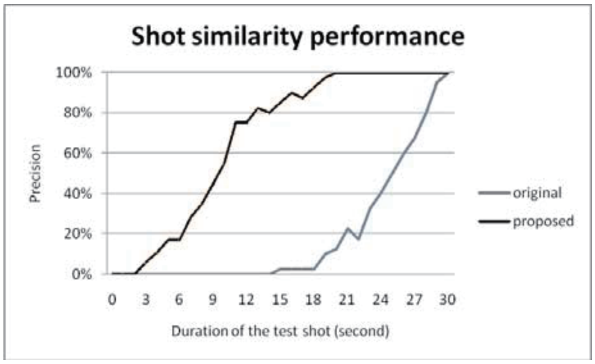


Fig. 6 Performance comparison between the proposed and original shot similarity approaches.

select them by averaging over 100 test queries' D_i and S_C values, and then dividing them by 5.

4.1 Performance of the Improved Shot Similarity

To test the accuracy of the modified shot similarity, 50 query shots with the same duration are randomly chosen from the database. Figure 6 shows the result generated by the proposed shot comparison approach and the original approach as the duration of the query shots changes. From this result, we can conclude that the accuracy of the original approach is affected dramatically by the duration, while the proposed approach provides a good solution to this problem.

4.2 Performance of the Improved Shot Query Approach

To verify if the improved shot query approach based on the modified shot similarity is able to outperform the original approach, 50 query shots with different durations are chosen from the database for shot query testing. Table 2 shows the comparison of query results between the original and proposed approaches. Figure 7 shows the first query example by our approach in Table 2 with five returned shots ($N_S = 5$), each shot with three representative frames. Figure 8 compares the search time required by the ten queries listed in Table 2.

As shown in Table 2, the original approach performs well when the length is long enough with the help of Stage 2 described in Sect. 2.2. But the false results increase dramatically as the length becomes short. Even 30 candidate shots are reserved to Stage 2, the correct one may be excluded out of the 30 candidates due to the unreliability of the original shot signature comparison. Whereas, the proposed approach shows better performance than the original approach in the latter case.

As shown in Fig. 8, the search time required by the original approach is about 2 to 4 times that required by the

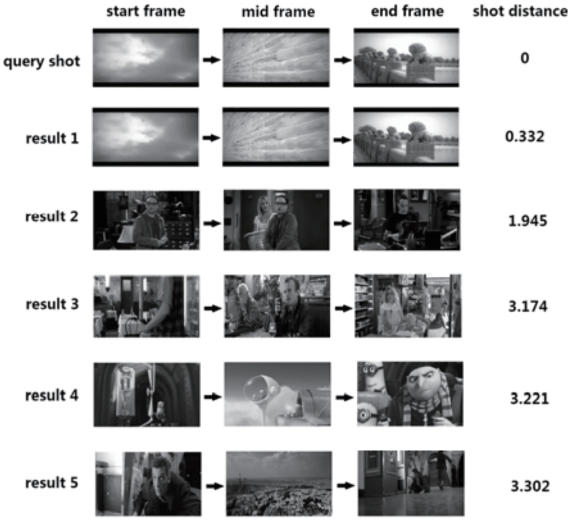


Fig. 7 The first query example by our approach in Table 2.

Table 2 Comparison of query results between the original and proposed approaches.

Query shots				Original approach			Proposed approach		
Query number	Original Video	Original Shot	Length (second)	Video found	Shot found	match	Video found	Shot found	match
1	3	2	30	3	2	1	3	2	1
2	5	3	28	5	3	1	5	3	1
3	13	4	26	13	4	1	13	4	1
4	18	1	24	18	1	1	18	1	1
5	20	3	22	20	3	1	20	3	1
6	25	5	20	25	5	1	25	5	1
7	31	4	18	41	17	0	31	4	1
8	34	2	16	41	17	1	34	2	1
9	40	4	14	46	1	0	40	4	1
10	45	1	12	7	4	0	43	2	0

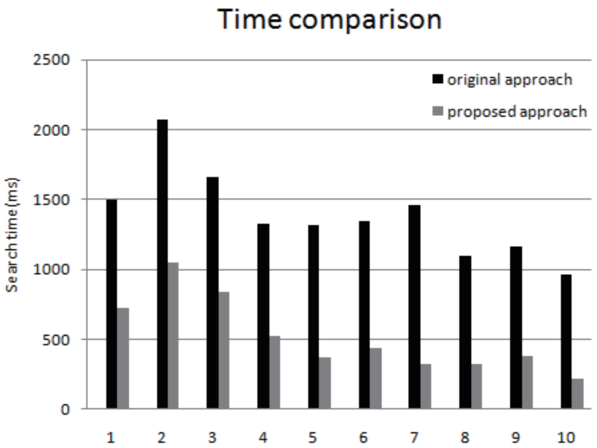


Fig. 8 Comparison of the search time required by the 10 queries shown in Table 2.

Table 3 The precision and the average S_C values for each test set.

Test sets	2 shots	3 shots	4 shots	5 shots	6 shots
Precision	100%	100%	100%	100%	100%
Average S_C	0.95	1.82	2.75	3.73	4.71

proposed approach. The main reason is that the process of frame signature comparison, which contributes much of the computation time, is done 30 times for every query by the original approach. For the proposed approach, at most N_S ($N_S = 5$) times of the same process is done and some methods described in Sect. 3.2 also help in saving computation costs.

4.3 Performance of the Proposed Clip Query Approach

To test the accuracy of the proposed clip query approach, 5 different sets, each of which containing 50 query clips with the same number of shots (a clip with only one shot is excluded as the case has been tested in Sect. 4.2) are chosen from the database for testing. Besides, some query clips, which are generated by the video sequences out of the database, are added to test the identification ability.

As shown in Table 3, the clip query result is very good. The reason is that the shots between the first one and the last one are complete which contribute much higher values to S_C than the incomplete shot, therefore, resulting in correct matching with high probability. To further test the robustness of the proposed clip query approach, the query clips whose each shot has some part of data missing are used for testing.

As shown in Fig. 9, the proposed approach shows strong robustness to data missing as it keeps high identification precision values even the missing rate reaches 60%. There are two reasons, one is that the new shot similarity is suitable for shots with short lengths, the other is the sensible definition of S_C. Though the average S_C value decreases as the missing rate increases, which may probably causes the mismatching in some shots, the whole clip can still be correctly matched as S_C is the sum of the similarities of all shots in it.

4.4 Comparison of Our Approach with Others

To show the superiority and the robustness of the proposed approach, we compare it with the method based on centroid of gradient orientations (CGO) by Lee and Yoo [17] and the method based on difference of block mean luminance (DBML) by Oostveen *et al.* [18]. All the parameters used in these two approaches are set as the same as what the authors give in their work. We first test the identification ability of them by choosing 50 query shots in the database with no distortion. The result is shown in Table 4.

As the query shots are not distorted, all of them can be found in the database. While all the recall rates are the same, the precision of the proposed method is the highest as at most N_S ($N_S = 5$) results will be returned for every query. And the search speed of the proposed method is much faster than others as CGO and DBML only have frame signatures but no shot signatures.

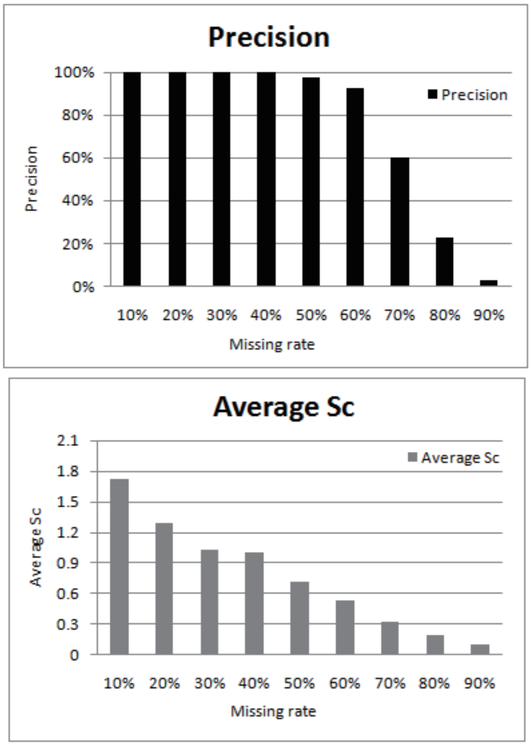


Fig. 9 The precision and the average SC values of queries with data missing in the case of 4 shots.

Table 4 Identification ability comparison of the three methods.

Algorithm	Recall rate	Average number of unrelated results	Search time (s)
CGO	100%	10.22	202.845
DBML	100%	5.56	217.395
Proposed	100%	3.52	19.78

Table 5 Robustness (recall rate) comparison of the three methods.

Methods	Median filtering	Blurring	Contrast enhancement	Rotation	Translation	Red+ 20%
CGO	98%	94%	94%	28%	90%	96%
DBML	100%	100%	98%	78%	42%	98%
Proposed	98%	90%	98%	82%	100%	100%

For robustness evaluations, a series of content-preserved distortions are implemented on test sequences, including 3 × 3 median filter, blur, contrast enhance, 10° rotation, 18-pixel horizontal translation and color variation (red +20%). The comparative test is performed using 50 query shots affected by each of the aforementioned distortions and the result is shown in Table 5.

As shown in Table 5, the proposed method show good robustness under various distortions. The recall rate reaches 95% on average and it is even completely unaffected by some types of them. And it also outperforms CGO and DBML in many cases especially the geometric distortion: rotation and translation. The robustness of CGO is low as the rotation will dramastically affect the gradient orientation. The robustness of DBML is significantly impaired as the translation will greatly change the value of the block mean luminance.

5. Conclusions

In this paper, an improved video identification method based on video tomography is presented. The proposed approach modifies the way of feature extraction and the process of similarity calculation with the consideration of some properties existing in video tomography. And the scope of query is also extended from one shot to several consecutive shots. The results show that the proposed similarity calculation approach is more suitable for shots with short lengths than the original approach, and the computation time is less than 50% of the original. The clip query performs very well and shows strong robustness to data missing. The algorithm comparison also shows its advantages in precision, search speed and robustness. For future development, more types of video transformations should be included to test the robustness of the approach and more types of features should be considered for resisting these transformations.

Acknowledgements

This work was partially supported by Zhejiang Provincial Natural Science Foundation of China (No.R1110006) and Special Fund on Strategic New Industry Development of Shenzhen under grant JCYJ20120817163934173, the National Science & Technology Pillar Program under grants No.2011BAH20B02 and 2011BAH20B03 and the Natural Science Foundation of SZU under grant 201215.

References

[1] MPEG Video Subgroup, "Updated call for proposals on video signature tools," MPEG2008/N10155, 2008.

[2] G. Doerr and J.L. Dugelay, "A guide tour of video watermarking," Signal Processing: Image Communication, vol.18, pp.263–282, 2003.

[3] T.T. Ng, S.F. Chang, C.Y. Lin, and Q. Sun, "Passive-blind image forensics," Multimedia Security Technologies for Digital Rights, pp.383–412, 2006.

[4] W. Luo, Z. Qu, F. Pan, and J. Huang, "A survey of passive technology for digital image forensics," Frontiers of Computer Science in China, vol.1, pp.166–179, 2007.

[5] B. Sevinc, T.S. Husrev, and M. Nasir, "Video copy detection based on source device characteristics: a complementary approach to content-based methods," ACM Multimedia Information Retrieval, pp.435–442, 2008.

[6] X. Fang, Q. Sun, and Q. Tian, "Content-based video identification: A survey," Proc. Information Technology: Research and Education, pp.50–54, 2003.

[7] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: A comparative study," Proc. 6th ACM International Conference on Image and Video Retrieval, pp.371–378, 2007.

[8] C.-Y. Chiu, C.-C. Yang, and C-.S. Chen, "Efficient and effective video copy detection based on spatiotemporal analysis," Proc. 9th IEEE International Symposium on Multimedia, pp.202–209, 2007.

[9] Y. Yan, B.C. Ooi, and A. Zhou, "Continuous content-based copy detection over streaming videos," Proc. 24th IEEE International Conference on Data Engineering, pp.853–862, 2008.

[10] N. Guil, J.M. Gonzalez-Linares, J.R. Cozar, and E.L. Zapata, "A clustering technique for video copy detection," Pattern Recognition and Image Analysis, LNCS. vol.4477, pp.451–458, 2007.

[11] G. Singh, M. Puri, J. Lubin, and H. Sawhney, "Content-based matching of videos using local spatio-temporal fingerprints," Proc. 8th Asian Conference on Computer Vision, pp.414–423, 2007.

[12] G. Leon, H. Kalva, and B. Furht, "Video identification using video tomography," Proc. IEEE International Conference on Multimedia and Expo, pp.1030–1033, 2009.

[13] A. Akutsu and Y. Tonomura, "Video tomography: An efficient method for camera work extraction and motion analysis," Proc. 2nd International Conference on Multimedia, ACM Multimedia 94, pp.349–356, 1994.

[14] A. Yoshitaka and Y. Deguchi, "Video summarization based on film grammar," Proc. 7th IEEE Workshop on Multimedia Signal Processing, pp.1–4, 2005.

[15] S. Possos, A. Garcia, M. Mendolla, J. Schwartz, and H. Kalva, "An analysis of independence of video signatures based on tomography," Proc. IEEE International Conference on Multimedia and Expo, pp.698–701, 2009.

[16] S. Possos and H. Kalva, "Accuracy and stability improvement of tomography video signatures," Proc. IEEE International Conference on Multimedia and Expo, pp.133–137, 2010.

[17] S. Lee and C.D. Yoo, "Video fingerprinting based on centroids of gradient orientations," Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.401–404, 2006.

[18] J. Oostveen, T. Kalker, and J. Haitsma, "Feature extraction and a database strategy for video fingerprinting," Proc. International Conference on Recent Advances in Visual Information Systems, pp.117–128, 2002.

Qing-Ge Ji received the Ph. D. degree in computer science and technology from Harbin Institute of Technology, Harbin, P. R. China, in 2002. He is currently working as an associate professor in Sun Yat-Sen University, Guangzhou, P. R. China. His research interests include video analysis and processing, computer graphics and game theory, etc.

Zhi-Feng Tan received the B.S. degree in software engineering from Sun Yat-sen University in 2009 and the M.S. in computer software and theory from Sun Yat-sen University in 2012. His research interests include video analysis and video processing.

Zhe-Ming Lu received the B.S. and M.S. degrees in Electrical Engineering from Harbin Institute of Technology (HIT), Harbin, P. R. China, in 1995 and 1997 respectively, and the Ph. D. degree in instrument science and technology from HIT, Harbin, P. R. China, in 2001. He is currently working as an full professor in Zhejiang University, Hangzhou, China. His research interests include multimedia single analysis and processing, information hiding and astronautics signal processing, etc.

Yong Zhang received the B.S., M.S. and ph. D. degrees in Computer Science from the PLA Science and Technology University, Nanjing, P. R. China, in 1997, 2001 and 2004 respectively. He is currently working as an associate professor in Shenzhen University, China. His research interests include information security and multimedia information processing, etc.