Signal Processing 93 (2013) 2339–2351

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A novel spatio-temporal registration framework for video copy localization based on multimodal features

R. Roopalakshmi*, G. Ram Mohana Reddy

Information Technology Department, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India

ARTICLE INFO

Article history:
Received 13 December 2011
Received in revised form
31 May 2012
Accepted 4 June 2012
Available online 19 June 2012

Keywords:
Temporal registration
Geometric alignment
Pirate video
CBCD
SURF
Spectral centroid
Dynamic time warping

ABSTRACT

Fighting movie piracy requires copy detection followed by the accurate frame alignments of master and copy videos, in order to estimate distortion model and capture location in a theater. Existing research on pirate video registration utilizes only visual features for aligning pirate and master videos, while no effort is made to employ acoustic features. Further, most studies in illegal video registration concentrate on the alignment of watermarked videos, while few attempts are made to address the alignment of non-watermarked sequences. We attempt to solve these issues, by proposing a novel spatio-temporal registration framework that utilizes content-based multimodal features for frame alignments. The proposed scheme includes three stages: first, a video sequence is compactly represented using Speeded Up Robust Features (SURF) and audio spectral signatures; second, sliding window based dynamic time warping (DTW) is employed to compute temporal frame alignments; third, robust SURF descriptors are utilized to generate accurate geometric frame alignments. The results of experiments on three different datasets demonstrate the robustness and efficiency of the proposed method against various video transformations.

1. Introduction

We first define two terms, namely "master" and "pirate" video sequences. A master video corresponds to a reference/database video; while a pirate video is derived from the master sequence by applying different video and editing transformations such as camcording, caption insertion and frame rate changes. In this paper, the term "registration" defines a way of mapping master and pirate video contents with an objective to compute frame-to-frame alignments. In order to facilitate the discussion in this paper, we use the three terms, "pirate sequence", "copy clip" and "query video" interchangeably hereafter as we do not distinguish between these three terms.

The massive growth of media streaming activities have increased the amount of duplicate videos and caused a huge loss to movie industry. CMPDA-2011 report (Canadian Motion Picture Distributors Association) says that 133M (Million) pirated movies were watched in Canada in 2010 [1]. This report also indicates a loss of C$413M to Canadian economy due to Internet based digital piracy. Thus, rigorous forensic analysis frameworks and countermeasures are required for preventing illegal movie captures.

Fighting movie piracy requires copy detection as the first step, which aims to determine the best matching master video for a given query clip. There are two approaches for detecting illegal videos: digital watermarking and content-based video copy detection (CBCD) [2]. CBCD techniques utilize content-based features of the media to detect illegal videos [3]; hence, they are widely popular compared to digital watermarking [3,4].

* Corresponding author. Tel.: +91 80 23245586; fax: +91 80 28362393.
E-mail addresses:
roopanagendran2002@gmail.com (R. Roopalakshmi),
profgrmreddy@gmail.com (G. Ram Mohana Reddy).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.004

Existing CBCD methods do not address frame alignments of a pirate content with the master sequence, because their ultimate aim is to detect illegal videos by comparing the perceptual similarity between the two video sequences.

On the other hand, in case of camcorder capture in a theater, significant mismatches may exist between master and pirate contents [5]. These mismatches could be temporal, geometric or combination of both; hence, after copy detection frame alignment of two video contents is very much essential for a number of applications such as estimating distortion model, detecting forensic watermarks and identifying capture location in a theater [6].

This paper focuses on the spatio-temporal alignment of master and pirate video sequences by utilizing content-based multimodal features. More precisely, we handle the specific problem of locating a given pirate clip within a master video sequence and obtaining accurate frame-to-frame alignments of two video sequences.

1.1. Related work

The research of pirate video registration is brand new. Early research focuses on the visual features for the alignment of master and pirate video contents. Delannay et al. [7] proposed a temporal registration technique by matching restricted number of frames obtained from the two video sequences, where the frame rates are assumed constant. In case of high motion activity, this method extracts different sets of key frames from the master and pirate videos. Cheng [8] introduced an algorithm for temporally matching two video contents using dynamic programming. Although this method achieves good registration accuracy, it is severely affected by transformations such as noise addition.

Cheng and Isnardi [9] developed a spatial, temporal and histogram registration scheme for video sequences by incorporating contextual costs and applied this algorithm to digital forensic watermark detection. In [10], Cheng reviewed and compared three different video registration algorithms proposed for detecting forensic watermarks in digital cinema applications.

Chupeau et al. [11] employed color histograms to match two video contents using dynamic programming. Due to the global descriptive nature of color histograms, this method performs poor for region-based transformations. Baudry et al. [12] utilized both the global and local fingerprints for registering video sequences, but this method scores poor results for low motion frames and complex transformations such as letterbox insertion and subtitles.

Recently, Baudry et al. [13] designed a registration scheme for video copies using temporally adaptive fingerprints, which are computed based on hierarchical encoding of the wavelet coefficients. Although this method guarantees accurate alignments, the encoding of wavelet coefficients is expensive in terms of CPU and memory. Lee et al. [14] presented a scheme for matching two video sequences using dynamic programming. This method significantly reduces the probability of matching errors by defining an effective matching cost function. However, only few types of video modifications such as frame insertions, shuffle, removal and compression attacks are addressed in this study.

Delannay et al. [15] focused on the estimation of geometric distortions that occur due to the camera acquisition process in a theater. They presented a system for compensating these distortions using a modified block matching technique, in order to retrieve embedded watermark information in digital cinema applications. Chupeau et al. [16] introduced a registration framework for estimating the distortion model and performing accurate distortion compensations in video copies. This algorithm attempts to align the pirate video frames with the master content as a prerequisite to the recovery of embedded forensic watermarks.

A common point of existing registration methods is that they concentrate only on the visual features of videos [7–14]. But, if audio content is available, it constitutes a significant information source of a video. Further, in case of illegal camcorder captures audio data is less affected compared to its counterpart [17].

From another perspective, most registration schemes are focusing on the alignment of watermarked documents [7–11,15,16], while only few efforts are made to address the alignment of non-watermarked videos. In addition, it is to be noted that not all copyrighted content is watermarked [6]. To summarize, there are as yet no promising schemes for pirate video registration that employ visual and acoustic features in a unified framework, while this research field is ongoing.

1.2. Motivation and contributions

If audio is present, then it is possible to significantly improve the registration accuracy by jointly exploiting the visual and acoustic fingerprints; hence, a novel pirate video registration framework using content-based multimodal features is required, which is useful even in the absence of forensic watermarks.

We propose a novel spatio-temporal registration framework that exploits visual fingerprints extracted from SURF interest points [18] and audio signatures based on spectral centroid features [19]. First, we present a novel visual-profile extraction method, which is compact (1-D) compared to the existing multi-dimensional SURF fingerprint methods [21,22]. Roth et al. [21] utilized 16-D SURF descriptors, whereas Zhang et al. [22] used 64-D SURF signatures for their CBCD task. Second, we employ robust acoustic features for the temporal registration task which noticeably improves accuracy compared to the existing schemes[11,13].

To make registration efficient, we use a multimodal frame matching scheme to align visual and acoustic feature sequences, which considerably reduces false frame matches. Further, we present an algorithm for selecting a candidate segment of the master sequence using sliding window based dynamic time warping (DTW) technique [19], which substantially decreases the frame matching cost.

The rest of this article is organized as follows: We formulate the registration problem and detail the

proposed framework in Section 2. In Section 3, temporal alignment of frames including visual and acoustic profile extraction followed by sliding window based DTW is illustrated. Geometric alignment of frames is detailed in Section 4. Section 5 explains the multimodal frame matching scheme including frame matching using visual and audio signatures. In Section 6, we describe the extensive evaluation experiments on different datasets and we summarize our conclusions in Section 7.

2. Proposed framework

2.1. Problem formulation

The proposed spatio-temporal registration framework is formulated as follows: let $PS=\{p_i|i=1,2,\ldots,n_p\}$ be a pirate sequence with n_p frames, where p_i is i-th copy frame; and let $MS=\{m_j|j=1,2,\ldots,n_m\}$ be a master sequence with n_m frames, where m_j is j-th master frame and $n_m \gg n_p$. Here PS is derived from MS after applying transformations such as camcording, noise, caption insertion and so on. We select a subsequence of MS denoted as a candidate segment $CS=\{m_j,m_{j+1},\ldots,m_{j+n_c-1}\}$ with n_c frames, using a sliding window scheme. Our goal is to spatio-temporally match the candidate and pirate sequences and as a result accurate frame-to-frame alignments of CS and PS can be obtained.

2.2. Proposed methodology

We propose a novel spatio-temporal registration framework shown in Fig. 1, which consists of two stages. In the first stage, when a copy clip is given, we scan the master sequence with a sliding window of length equal to the copy clip. In this stage, the similarity between the pirate clip and the windowed sequence is measured based on their temporal signatures derived from SURF interest points and spectral centroid features. The windowed sequence with minimum distance score is selected and denoted as a candidate segment. More precisely, the algorithm used to select the candidate segment of the master sequence is detailed in Section 3.4. After this point, visual–audio fingerprints of two video contents are matched separately using DTW technique and the matching results are fused, in order to obtain temporal frame-to-frame alignments.

In the second stage, from the temporally aligned pirate and candidate frames, we select a set of highly similar frames denoted as *principal frames* of two video sequences. More specifically, *principal frames* are extracted using the algorithm explained in Section 5. The resultant *principal frames* are mapped using their SURF descriptors by means of enough control points, in order to achieve accurate spatial frame alignments.

3. Temporal alignment of frames

3.1. 1-D visual profile extraction

In the proposed framework, we employ SURF key points-based signatures to extract the visual profile of video contents. SURF is a scale and rotation invariant descriptor [18]; hence, it is widely used in the CBCD literature to detect pirate video clips [21–23]. The problems encountered during visual profile extraction and the proposed solutions are detailed below.

Problem. SURF descriptor associates each key point with a high dimensional feature vector typically 64 integers per key point. Each frame might contain multiple SURF key points; hence, there would be too much of information to process. Moreover, direct comparison of SURF feature descriptors across all frames would be computationally expensive.

From another perspective, existing multi-dimensional SURF fingerprints consider only spatial content of frames [21–23]. But, to generate a robust visual profile of a video, both the spatial and temporal information of frames need to be considered.

In order to solve the issues, in the proposed framework a video sequence is compactly represented using 1-D SURF signatures derived from SURF interest points of frames, which efficiently characterize the spatio-temporal content of frames. More precisely, we segment a video frame into

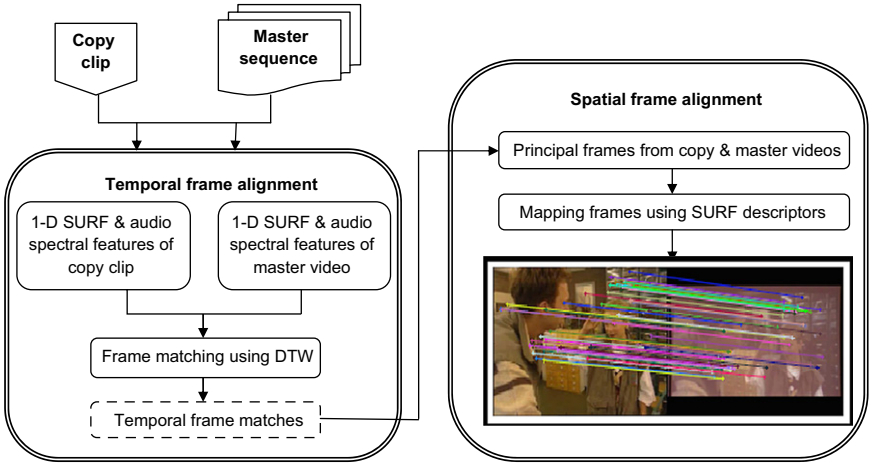

Fig. 1. Overview of proposed registration framework.

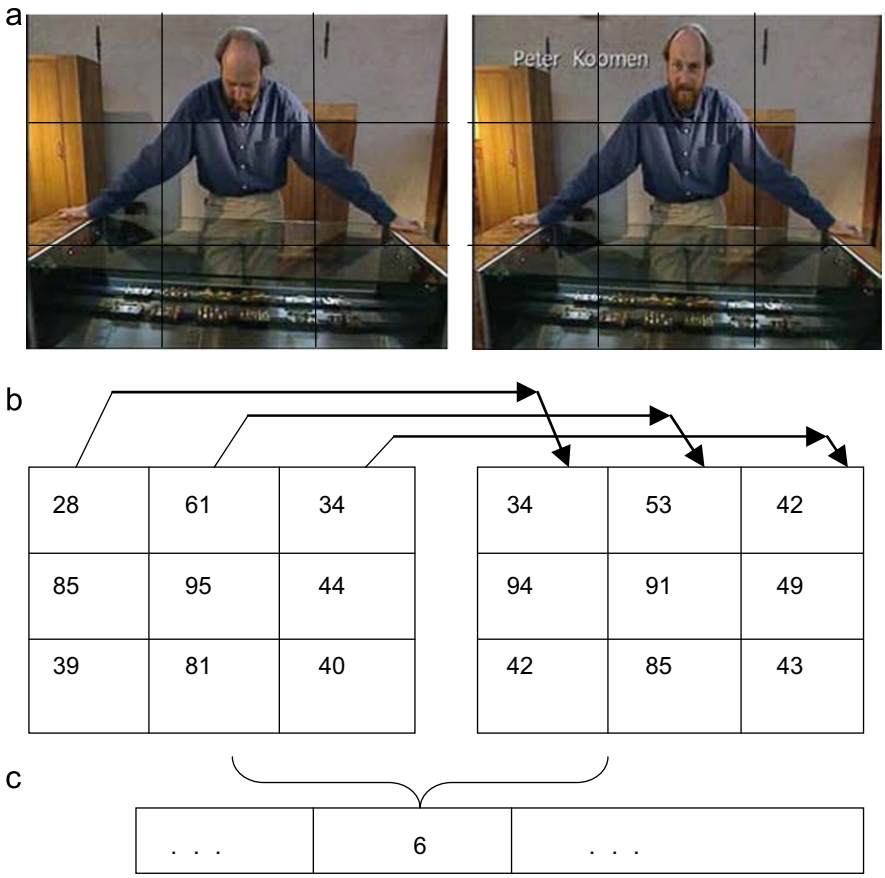

Fig. 2. 1-D SURF signature extraction: (a) video frames partitioned into 3 × 3 regions; (b) region-wise count of SURF key points; (c) computing 1-D signature with time series.

$k \times k$ regions and compute 1-D SURF signatures as the mean of differences between region-wise count of SURF interest points of consecutive frames. Fig. 2 illustrates computation of 1-D SURF signatures from sample frames on a 3×3 partition.

Problem. The segmentation of a frame into $k \times k$ regions plays a significant role in determining the registration accuracy and the computation speed. Smaller values of k increase the computational load, while larger values of k may decrease robustness of the proposed system.

In order to solve this discrepancy, experiments are conducted and registration performances are compared for different values of k ranging from 2 to 8. More specifically, we experiment on a dataset including 112 pirate clips and 198 master videos, where the copy clips vary between 18 and 35 s. Fig. 3 indicates the average registration results obtained for different k values and concludes that maximum accuracy (91.8%) is achieved at $k=3$. Thus, we set the value of k as 3 in the subsequent experiments, which yields the best balance of robustness and effectiveness.

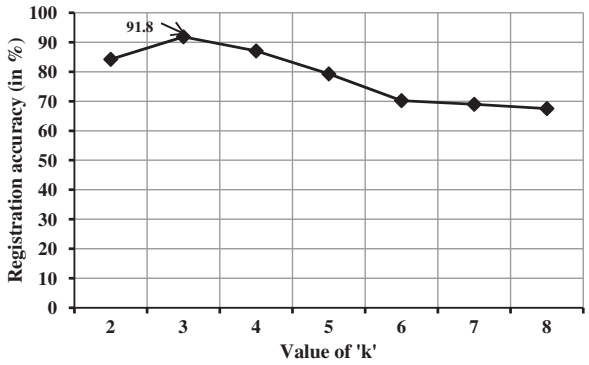

Fig. 3. "k" versus registration accuracy.

3.2. 1-D acoustic profile extraction

In the literature of sound synthesis, spectral centroid is proved to be an important timbrel descriptor, which specifies the center of gravity of the signal spectrum [19,20]. Specifically, centroid is a highly robust spectral feature that describes brightness of a sound signal [24]; hence, it is popularly used in speech recognition applications [25]. On the other hand, the most important perceptual audio features exist in the frequency domain. Due to these reasons, we utilize 1-D spectral centroid signatures to describe the acoustic profile of video contents, which is computed as follows.

First an audio signal is down sampled to 22 050 Hz, in order to reduce the size of data to be processed. The magnitude spectrum of the audio signal behaves almost stationary for 10–30 ms of window length; hence, the down sampled audio signal is segmented into 11.60 ms windows using Hamming window function with an overlap factor of 80% [26]. From the power spectrum of the audio signal, the Spectral Centroid descriptor *SC* is computed using frequency distribution values as follows:

$$SC = \sum_{k=1}^{N} k \times x^d[k] \Big/ \sum_{k=1}^{N} x^d[k] \tag{1}$$

where $x^d[k]$ represents the magnitude of k-th frequency bin of d-th frame and N is the frame length. As compared with [26], we use absolute values of spectral centroid features for the proposed framework. In addition, we apply normalization to the resultant features in order to improve the robustness of audio signatures considered in this framework.

3.3. Introduction to dynamic time warping

Dynamic time warping (DTW) is extremely efficient in synchronizing two time-dependent sequences, because it minimizes shifting effects in time by allowing elastic transformation of sequences [19,27]. Therefore, DTW is extensively explored in a wide range of applications such as speech recognition [28], sequence alignment and information retrieval [27].

Given two time-dependent feature sequences $X = \{x_i | 1 \le i \le N\}$ of length N and $Y = \{y_j | 1 \le j \le M\}$ of length M. A local cost measure C indicating the distance between x_i and y_j is formulated as

$$C(x_i, y_j) = Dist(x_i, y_j) \tag{2}$$

where *Dist* denotes Manhattan/Euclidean distance metric in the proposed registration framework. In order to find an alignment of X and Y, we need to compute a warping path $W = \{w_1, w_2, \ldots, w_L\}$ with $w_l = (x_l, y_l) \in [1:N] \times [1:M]$ for $l \in [1:L]$. The accumulated Path Cost *PC* associated with W of sequences X and Y is defined as

$$PC_W(X,Y) = \sum_{l=1}^{L} C(x_{i_l}, y_{j_l}) \tag{3}$$

The goal of DTW is to find an optimal warping path of sequences X and Y having minimal path cost among all possible warp paths [27], which is denoted as

$$DTW(X,Y) = W_{op} = \min\{(PC_W(X,Y)) | W \in P^{N \times M}\} \tag{4}$$

where W_{op} is the optimal warping path and $P^{N \times M}$ represents the set of all possible warping paths. The optimal warping path $W_{op} = \{wp_1, wp_2, \ldots, wp_L\}$ with $wp_l = (x_l, y_l) \in [1:N] \times [1:M]$ for $l \in [1:L]$. The accumulated path cost of $DTW(X,Y)$ is denoted as

$$PC_{dtw}(X,Y) = \sum_{l=1}^{L} C(x_{n_l}, y_{m_l}) \tag{5}$$

Let $D(N,M)$ be the global cost matrix of size $N \times M$. DTW algorithm determines the warping path W_{op} based on dynamic programming [27] in three steps as follows:

a: **Initialization**:

$D(1,1) = 0$;

First column: $D(i,1) = \sum_{k=1}^{i} C(x_k, y_1), i \in [1:N]$;

First row: $D(1,j) = \sum_{k=1}^{j} C(x_1, y_k), j \in [1:M]$.

b: **Recursion**:

All other elements of $D(i,j)$ are recursively computed as

$$D(i,j) = \min\{(D(i-1,j-1), D(i-1,j), D(i,j-1)) + C(x_i, y_j)\} \tag{6}$$

where $i \in [1:N]$ and $j \in [1:M]$.

c: **Termination**:

Once the entire D matrix is computed, backtracking is done to determine the optimal alignments starting from $W_{op}=(M,N)$ to $W_{op}=(1,1)$.

In this study, the optimal warping path W_{op} specifying the alignment of sequences X and Y satisfies the following conditions:

(a) **Endpoint constraints**:

For the warping path W_{op}, starting point is $wp_1 = (1,1)$ and ending point is $wp_L = (N,M)$.

(b) **Monotonicity conditions**:

In order to preserve temporal continuity, the warping function is monotonically increasing as given by $x_1 \le x_2 \le \cdots \le x_L$ and $y_1 \le y_2 \le \cdots \le y_L$.

(c) **Local continuity constraints**:

This criteria constraints slope of the warping path by limiting long jumps in the alignment of X and Y sequences. Generally, the possibility of huge changes in the feature sequences of consecutive frames is very low and thus we considered the step size condition formulated as $wp_{l+1} - wp_l \in \{(1,0),(0,1),(1,1)\}$ for $l \in [1:L-1]$.

3.4. Sliding window based DTW

The computational complexity of DTW algorithm to match two sequences of size M and N is $O(MN)$; hence, if sequence size increases, the performance of the algorithm degrades. In order to overcome this problem, we computed frame matches between the copy clip and the candidate segment instead of the entire master sequence. Algorithm 1 explains the steps used to select a candidate segment of the master sequence.

Algorithm 1. Selection of a candidate segment

1: Divide the master sequence into overlapping segments of length equal to the query clip.
2: Extract 1-D visual and audio profiles for each segment using the procedures explained in Sections 3.1 and 3.2.
3: Let a master sequence *MS* be
$$MS \in \{S_i | \ 1 \le i \le m\} \tag{7}$$
where S_i is the i-th segment and m is total segments of *MS*. Here, each segment S_i of *MS* can be represented as
$$S_i \in \{(V_i^k \cup A_i^r) | 1 \le k \le n, \ 1 \le r \le p\} \tag{8}$$
where V_i^k is k-th feature vector of visual fingerprint of S_i and n indicates total feature vectors. Here, A_i^r is r-th vector of audio fingerprint of S_i and p represents number of feature vectors.
4: Let a pirate sequence *PS* is compactly represented as
$$PS \in \{(QV^k \cup QA^r) | 1 \le k \le n_q, \ 1 \le r \le p_q\} \tag{9}$$
where QV^k is the k-th feature vector of visual fingerprint of *PS* and n_q is total vectors. Here, QA^r is r-th vector of audio fingerprint of *PS* and p_q indicates total feature vectors.

5: Compute the segment similarity Seg_{sim} between S_k of MS and PS using DTW as follows:
$$Seg_{sim}(S_k,PS) = PC_{dtw}(V_k,QV) + PC_{dtw}(A_k,QA) \quad (10)$$
where PC_{dtw} represents the accumulated path cost of optimally warped visual sequences (i.e., V_k and QV) and audio feature sequences (i.e., A_k and QA), respectively.
6: Select S_i having lowest Seg_{sim} value (i.e., distance score) as a candidate segment of the master sequence for further comparison.

4. Multimodal frame matching

In this scheme, the visual-acoustic fingerprints of two video sequences are matched separately and the resultant matches are fused in order to get final temporal alignments. The multimodal frame matching scheme is implemented as follows.

4.1. Frame matching using visual signatures

Let CS be a candidate segment of the master sequence with n_c frames and PS be a pirate sequence with n_p frames. Let VF is the visual fingerprint of CS such that, $CS \in \{VF_i \mid 1 \le i \le n_{vf}\}$ with n_{vf} signatures. Consider QVF is the visual fingerprint of PS such that $PS \in \{QVF_j \mid 1 \le j \le n_{qvf}\}$ with n_{qvf} signatures. In this study, we assume that the length of candidate sequence is equal to size of the query clip; hence, $n_{vf} \simeq n_{qvf}$. The cost measure C_{vis} denoting the dissimilarity between two visual signatures is computed using comparative Manhattan distance metric as follows:

$$C_{vis}(CS_k,PS_k) = \frac{|(VF_k - QVF_k)|}{|(VF_k)| + |(QVF_k)|}, \quad 1 \le k \le n_{vf} \quad (11)$$

After this, the optimal alignments between the visual fingerprints of two video sequences are computed using DTW algorithm described in Section 3.3. The resultant frame matches FM_{vis} based on visual signatures is formulated as

$$FM_{vis} = \{\{cv_i,pv_j\} \mid 1 \le i \le n_c,\ 1 \le j \le n_p\} \quad (12)$$

where cv and pv indicate the matching frames of candidate and pirate video sequences, respectively.

Fig. 4 shows the frame alignments of copy and candidate feature sequences in terms of global cost matrix D and optimally warped path. The dark strips in matrix D indicate high similarity between the two video contents.

4.2. Frame matching using acoustic signatures

Let SF is the spectral centroid-based audio fingerprint of CS such that $CS \in \{SF_m, \mid 1 \le m \le n_{sf}\}$ with n_{sf} signatures. Let QSF is the audio fingerprint of PS such that $PS = \{QSF_m \mid 1 \le m \le n_{qsf}\}$ with n_{qsf} signatures. In this study, $n_{sf} \simeq n_{qsf}$. The cost measure C_{aud} denoting the difference between two audio signatures is computed using squared Euclidean distance as follows:

$$C_{aud}(CS_k,PS_k) = |(SF_k - QSF_k)^2|, \quad 1 \le k \le n_{sf} \quad (13)$$

After this, the optimal warping path specifying the frame alignments of SF and QSF signatures is computed using DTW algorithm described in Section 3.3. The resultant frame matches FM_{aud} based on audio spectral signatures is formulated as

$$FM_{aud} = \{\{cs_i,ps_j\} \mid 1 \le i \le n_c,\ 1 \le j \le n_p\} \quad (14)$$

where cs and ps indicate the matching frames of candidate and pirate sequences, respectively.

4.3. Decision fusion

Frames mapped by both the visual and audio signatures are considered as final frame matches of two video

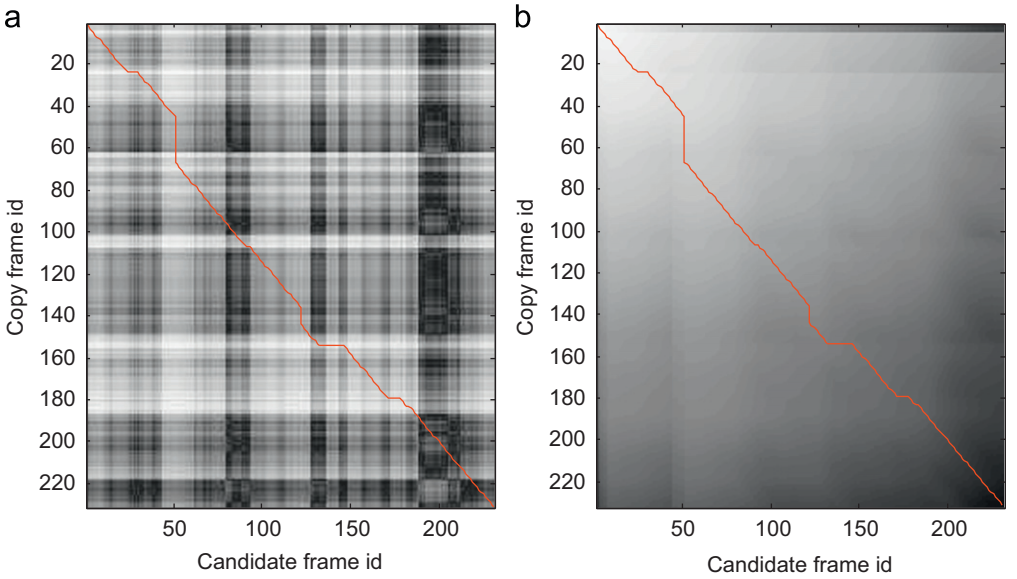

Fig. 4. Frame alignments of copy and candidate feature sequences: (a) global cost matrix D, darker regions indicate high similarity; (b) optimally warped path.

contents, which is given by

$$FM_{final} = \{\{FM_{vis}\} \cap \{FM_{aud}\}\} \quad (15)$$

where FM_{final} provides frame-to-frame alignments of *CS* and *PS* sequences, respectively. The advantage of proposed multimodal frame matching scheme is, it significantly reduces false frame matches, because only frames with similar visual and audio signatures are mapped. In addition, the proposed matching scheme noticeably improves registration accuracy which is evident in Section 6.

5. Geometric alignment of frames

Performing the geometric alignment across all temporally aligned frames of two video contents is not feasible due to computational load. Further, all video frames may not provide necessary key points to enable accurate geometric registration.

In order to solve this problem, we employ a small set of highly similar frames, denoted as *principal frames* for implementing the geometric registration task. The SURF descriptors and DTW optimal paths computed for temporal registration provide significant guidelines for selecting *principal frames*. More specifically, *principal frames* are extracted from temporally aligned candidate and pirate feature sequences using Algorithm 2, which is detailed as follows.

Algorithm 2. Principal frames extraction

1: Let the optimal warping path W_{op} specifies the alignment of two feature sequences *VF* and *QVF*, such that $W_{op} = \{wp_1, wp_2, \ldots, wp_L\}$ with $wp_l = (x_l, y_l) \in [1:N] \times [1:M]$ for $l \in [1:L]$. Here, *VF* and *QVF* represent the visual fingerprints of candidate and pirate sequences, respectively.

2: Consider a cost vector W^c_{op} representing the feature distances in terms of cost in each entry of optimal path W_{op} as follows:
$$W^c_{op} = \{wp^c_1, wp^c_2, \ldots, wp^c_L\}, \quad (16)$$
where wp^c_1 indicates the cost given in entry wp_1 and so on.

3: Sort the W^c_{op} vector to generate the sorted list of costs represented in DTW path.

4: Lower cost values in the W^c_{op} vector indicate highly similar frames; hence, select frame pairs corresponding to lower cost values in W^c_{op} as *principal frames*.

The resultant *principal frames* are characterized by a list of interest points and their associated SURF descriptors. Two control points are matched, if the squared Euclidean distance between their feature vectors is minimum. On the other hand, blind comparison of all feature vectors of two frames is computationally expensive and may lead to false correspondences. In order to solve this discrepancy, feature vectors with minimum feature distances are computed and mapped in terms of their descriptors to provide accurate pixel correspondences of frames.

6. Experimental setup and results

The proposed framework is evaluated on three different datasets, namely TRECVID sound & vision data [29], CC_WEB_VIDEO dataset [30] and a set of real data consisting of camcorded copies of master video files.

6.1. Master video database and query dataset construction

6.1.1. TRECVID dataset

TRECVID sound & vision data [29] is a benchmark dataset, which covers a wide variety of contents including science news, reports, documentaries and educational programming. Our TRECVID master database comprises approximately 110 h of sound & vision data used in TRECVID-2009 copy detection task, plus another 80 h of sound & vision data used in TRECVID-2008 copy detection task. We transformed entire video data into the following uniform format: 352×288 pixels and 15 fps (frames/s). It is not necessary to utilize every frame in a video sequence for registration; hence, when a copy clip is given with a different frame rate, it is resampled to 15 fps, in order to synchronize it with the master sequence. For example, a 5-s copy clip with 60 fps becomes a 240-frame sequence after performing the resampling procedure.

In case of piracy, normally users capture videos by using camcorders and distribute them with some modifications [30]. Thus, most of the pirate videos suffer from distortions such as camcording, photometric variations (lighting changes), editing operations (pattern insertions), frame rate changes, format changes (mp3 format), cropping, rotation attacks and so on; hence, in this context we considered 15 types of transformations listed in Table 1 to generate the query dataset. From the TRECVID master database, 50 video clips are randomly selected and Table 1 transformations are applied to produce the query clips. The resulting 750 (50∗15) video sequences of duration 20–35 s are used as query clips for the proposed temporal registration task.

6.1.2. CC_WEB_VIDEO dataset

CC_WEB_VIDEO dataset [30] includes video collections from video sharing websites and search engines such as YouTube, Google Video and Yahoo ! Video. Our CC_WEB_VIDEO master database includes 24 most viewed and top favorite videos provided by CC_WEB_VIDEO collection [30]. The representative snapshots of all 24 master videos are shown in Fig. 5. From the CC_WEB_VIDEO collection, we retrieved duplicate and near-duplicate videos ranging from 15 to 25 for each of the master video. In total, our CC_WEB_VIDEO query dataset includes approximately 600 video files with two different classes of distortions namely formatting and content distortions.

Formatting distortions include changes in frame rate, bit rate, encoding format and frame resolution. Photometric variations (lighting changes), editing variations (e.g., logo insertions) and content modifications such as addition of unrelated frames with different content are categorized into content distortions type.

6.1.3. Camcorded copies

To assess the performances of our algorithm against camcorder captured videos, we worked on a dataset of 30 master videos and their camcorded versions. We generated 75 camcorded copies of master videos ranging from 1.55 mn to 15 mn. The quality of camcorded copies varies from clean copies to heavily modified ones with a large amount of lighting, cropping and compression distortions.

Table 1
List of transformations used in the proposed registration framework.

#	Category	Description
T1	Zoom in	Zoom in to the frame by 19%
T2	Slow motion	Halve the video speed
T3	Fast forward	Double the video speed
T4	Pattern insertion	Insert text pattern into selected frames
T5	Moving caption	Insert moving titles into entire video
T6	Rotation	Rotating by 10–12°
T7	Random noise	Add 10% gaussian noise
T8	Blurring	Blur by 13%
T9	Brightness change	Increase brightness by 10%
T10	Cropping	Crop top and bottom regions by 20% each
T11	Picture-in-picture	Insert smaller resolution picture into selected frames
T12	3 combined	Cropping by 15%, 10% of noise and moving caption
T13	5 combined	14% noise, 11% blurring, 14% brightness, cropping and pattern insertion
T14	Mp3 compression	Change audio file format
T15	Single band compression	Compress only specific frequency band

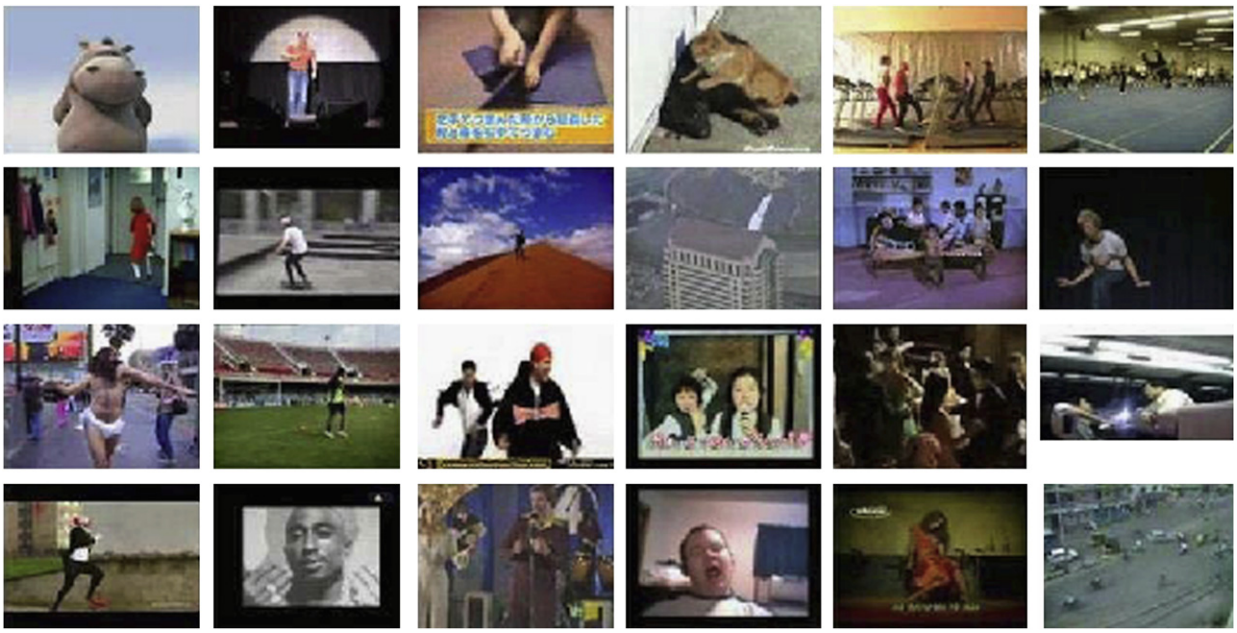

Fig. 5. Snapshots of 24 master videos of CC_WEB_VIDEO collection [30].

6.2. Overview of evaluated methods

We implemented the following six methods for evaluating performance:

(1) The SURF signatures based matching (abbreviated as SURF).
(2) The spectral centroid features based matching (SC).
(3) SURF and spectral features without sliding window (SURF+SC).
(4) SURF and spectral signatures with sliding window (ALL).
(5) Chupeau et al.'s method [11] (CHE).
(6) Baudry et al.'s method [13] (BA).

Our methods [methods (1)–(4)] evaluated different combinations of the proposed techniques. Methods (1) and (2) used different video signatures (namely SURF and spectral centroids) to implement the temporal registration of two video contents. We implemented methods (3) and (4), to see the effect of sliding window scheme for the proposed registration task.

In method (1), 1-D visual signatures of the pirate clip are matched with that of the entire master sequence (i.e., query clip is matched with all segments of the master sequence). In method (2), 1-D spectral signatures of the query clip are mapped with the acoustic profile of the complete master sequence. Method (3) utilizes both 1-D SURF and spectral centroid signatures for temporally registering the video contents. In this method, visual–audio fingerprints of the query clip are separately aligned with that of the entire master sequence.

In method (4), we employed a sliding window mechanism to align multimodal signatures of the copy clip with the corresponding features of the candidate

segment, instead of entire master sequence. The candidate segment of the master sequence is selected using Algorithm 1 explained in Section 3.4.

Chupeau et al. [11] utilized color histograms for calculating frame-to-frame correspondences between pirate and master contents. It is implemented as follows: color histograms of size 512 bins are extracted from consecutive video frames. A sequence of distances (Euclidean distance) between color histograms of successive frames are utilized as temporal fingerprints of videos and dynamic programming is applied to achieve temporal registration of frames.

Baudry et al.'s method [13] is one of the latest methods that uses fingerprints based on wavelet coefficients for temporally registering query and master video sequences. In this method, first the difference between successive frames is computed and transformed into wavelet coefficients. Then the resultant coefficients are hierarchically encoded and temporal frame alignments are computed using dynamic programming.

6.3. Temporal registration results

In the following subsections, we show and discuss the registration performances of six compared methods tested on different datasets against different types of video transformations.

6.3.1. Registration results for TRECVID dataset

Table 2 shows the temporal registration performances of six compared methods in terms of percentage of perfectly matched frames (MF) for T1–T7 types. The bold font indicates the highest MF scores in the table.

The performance of spectral centroid-based methods (methods (2), (3) and (4)) is superior compared to other methods for all seven types. This is because, applying transformations on the visual content would not affect acoustic features substantially. Method (4) slightly improves the registration accuracy (by 1%) compared to that of method (3), because of the incorporation of sliding window scheme, which reduces false positives. Though 1-D SURF and spectral centroid signatures have their own constraints, they balance each other very well; hence, their integrated usage in a sliding window manner noticeably improves the registration accuracy. The improved results of method (4) shown in Table 2 prove this view point.

On the other hand, Chupeau et al.'s method [11] yields poor results for moving caption and pattern insertion types in terms of low MF rates. This is because, inserting patterns or adding captions noticeably changes color histogram properties. The MF rate of Baudry et al.'s method [13] declines sharply for random noise type. The reason is, adding random noise might alter the wavelet coefficients substantially, which leads to false fingerprints.

Table 3 lists the temporal registration accuracy of six compared methods for T8–T15 types in terms of MF rates. Method (4) generally performs well for all eight types and improves the MF rates (up to 15%) compared to the reference methods. Method (4) slightly enhances the registration accuracy (by 1%) compared to method (3). The reason for this improvement is, when the sliding window scheme is utilized, query features are matched only with that of candidate segment and thus false positive rate is reduced.

The MF rate of Chupeau et al.'s method [11] is severely decreased for cropping and picture-in-picture types. This is because, cropping introduces black borders on top and bottom regions that might generate very different signatures for master and query clips. In case of picture-in-picture type, insertion of picture produces different signature pattern for the query video compared to the original file.

On the other hand, Baudry et al.'s method [13] yields poor MF rates for picture-in-picture and five combined types. In picture-in-picture type, there exists a discrepancy between the wavelet coefficients extracted from master and query videos, because of the insertion of a picture. This discrepancy leads to mismatches and thus reduces the accuracy of method (6). In case of five combined type, the wavelet coefficients vary widely after applying noise, cropping and pattern insertions and hence a lot of mismatches are retrieved.

The accuracy of method (2) is sharply reduced for mp3 and single band compression types. Audio spectral features are much affected by these two types and hence MF rates decline sharply. Yet our methods using SURF features (methods (1), (3) and (4)) are less affected by these two types.

Table 2
Registration results for T1–T7 types. **MF**: % of perfectly matched frames.

Attacks	SURF (1)	SC (2)	SURF+SC (3)	All (4)	CHE (5)	BA (6)
	MF	MF	MF	MF	MF	MF
Zoom in	71.9	92.7	93.2	**93.2**	55.8	69.8
Slow motion	78.1	79.5	89.9	**90.4**	60.0	68.8
Fast forward	84.7	85.6	91.0	**91.0**	59.8	61.2
Pattern insertion	81.7	92.5	93.7	**94.2**	54.8	54.0
Moving caption	88.0	93.8	93.8	**93.8**	50.7	62.7
Rotation	84.6	92.8	95.2	**95.2**	68.9	59.8
Random noise	89.7	92.4	94.7	**94.8**	64.2	51.2

Table 3
Registration results for T8–T15 types. **MF**: % of perfectly matched frames.

Attacks	SURF (1)	SC (2)	SURF+SC (3)	ALL (4)	CHE (5)	BA (6)
	MF	MF	MF	MF	MF	MF
Blurring	82.7	91.5	93.5	**94.2**	53.8	55.8
Brightness	90.0	95.8	95.9	**95.9**	62.0	59.7
Cropping	80.0	90.0	92.4	**92.4**	44.8	50.5
Picture-in-picture	75.4	92.3	92.9	**93.0**	39.7	42.0
3 combined	88.7	92.6	94.2	**94.4**	50.9	53.6
5 combined	89.1	92.7	92.8	**93.1**	53.9	44.8
Mp3	90.8	78.9	90.6	**90.6**	86.6	89.0
Single band	93.7	75.6	94.6	**94.7**	85.7	87.0

Although the SURF and spectral features have their own advantages and limitations, they complement each other by their different characteristics; hence, the combination of local and spectral features not only improves the registration accuracy, but also widens the coverage to more number of transformations. The promising results of method (4) provide good evidence for supporting this viewpoint.

Fig. 6(a) shows the registration results of six compared methods for T1–T7 types, in terms of Average Distance between true and estimated frame indexes (AD). The curves indicate the better performance of spectral signature based methods (methods (2), (3) and (4)), compared to other methods because their AD rates are always less than 1. It is clear that method (4) yields lowest AD rates and significantly improves accuracy compared to other methods. The combined utilization of robust visual and acoustic features in a sliding window manner is the exact reason for the enhanced performance of method (4).

Fig. 6(b) indicates the registration results of six compared methods for T8–T15 types in terms of their AD rates. We observe that the curves show the superior performance of method (4), compared to other methods because its AD rates are always less than 1. For T12 and T13 types, only visual features based methods (methods (1), (5) and (6)) indicate poor results in terms of higher AD rates. However, spectral signature based methods (methods (2), (3) and (4)) are less affected by this category.

6.3.2. Computational cost comparison

Table 4 shows the total time costs of methods (1)–(6), which includes signature extraction and frame matching costs. The program is executed in MATLAB and run on a PC with 2.8 GHz CPU and 3 GB RAM. They are measured by implementing frame alignment of a 298 s query clip with a 3041 s master sequence.

The signature extraction cost of method (4) is higher (up to 47%), compared to two reference methods. Interestingly, the frame matching cost of method (4) is noticeably reduced (up to 94%) compared to methods (5) and (6). This is because, in method (4) query clip signatures are aligned only with the corresponding candidate segment features instead of the entire master sequence. Thus, in method (4), the usage of sliding window scheme significantly reduces the total time cost (up to 32%) and yields lowest computational cost.

6.3.3. Registration results for CC_WEB_VIDEO dataset

Table 5 lists the registration results of five compared methods for first 12 master videos of CC_WEB_VIDEO dataset in terms of percentage of Incorrectly matched Frames (IF). The bold font indicates the lowest IF scores in the table.

In case of the first master video, the visual content is affected by distortions such as encoding format change and logo insertions; hence, only visual feature based methods (methods (1), (4) and (5)) yield higher IF rates. For the second master video, few unrelated frames are added with same acoustic information; hence, method (2) leads to lot of false matches. However, characteristics of SURF and spectral features complement each other and hence method (3) improves accuracy and yields lowest IF rate for the second video.

In case of fourth master video, acoustic information is removed and captions are inserted to create the query videos; hence, method (2) leads to null matches. However, method (3) scores lowest IF rates, because of the robust nature of SURF-based visual signatures. For the fifth video, Chupeau et al. [11] and Baudry et al. [13] methods score poorly in terms of higher IF rates. The reason is, visual descriptors might be affected substantially due to the application of photometric and formatting variations such as color, lighting, frame rate and resolution changes. For the ninth video, Baudry et al.'s method [13] gives highest IF rate compared to other methods. This is because, editing and encoding format

Table 4
Computational cost comparison.

Process	SURF (1)	SC (2)	SURF+SC (3)	ALL (4)	CHE (5)	BA (6)
Signature extraction	68.1	21.1	90.1	90.1	47.4	59.8
Frame matching	108.0	87.4	95.0	4.1	66.8	74.1
Total cost	176.1	108.5	185.1	94.2	114.2	133.9

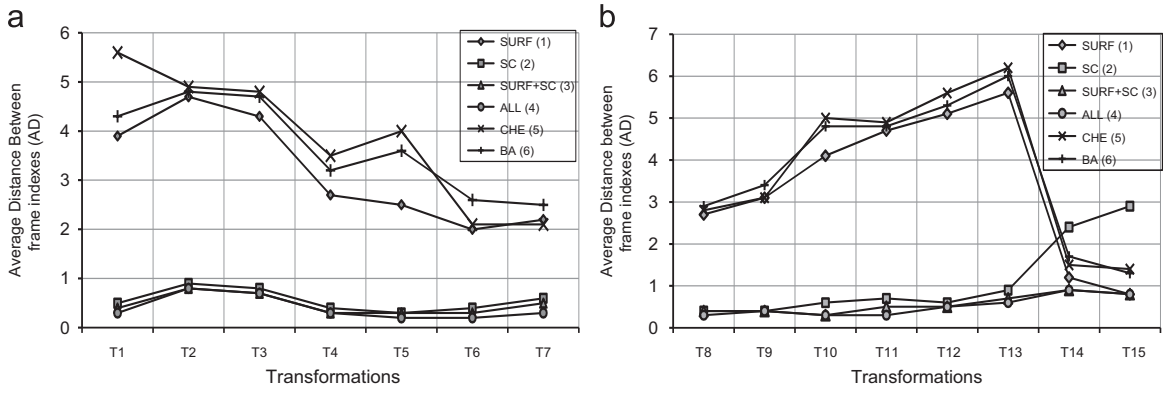

Fig. 6. Comparison of AD curves for different transformations: (a) T1–T7; (b) T8–T15.

Table 5
Registration results for 1–12 master videos. **IF**: % of incorrectly matched frames.

#	Video name	SURF (1)	SC (2)	SURF+SC (3)	CHE (5)	BA (6)
		IF	IF	IF	IF	IF
(1)	The lion sleeps tonight	15.7	13.9	**10.4**	28.1	26.4
(2)	Evolution of dance	24.4	42.3	**23.6**	29.5	30.4
(3)	Fold shirt	27.5	14.6	**12.3**	38.4	40.9
(4)	Cat massage	20.2	–	**20.2**	39.3	25.0
(5)	Ok go here it goes again	40.5	34.0	**23.1**	55.6	53.9
(6)	Urban ninja	38.2	46.6	**35.5**	38.2	39.4
(7)	Real life Simpsons	41.4	52.6	**39.1**	43.3	42.2
(8)	Free hugs	39.1	25.6	**20.6**	42.0	40.2
(9)	Where the hell is Matt	21.6	13.3	**11.2**	29.1	34.2
(10)	U2 and green day	12.6	14.9	**10.4**	28.1	21.6
(11)	Little superstar	41.3	38.5	**33.6**	44.1	42.0
(12)	Napoleon dynamite dance	31.5	46.1	**27.5**	37.1	33.2

Table 6
Registration results for 13–24 master videos. **IF**: % of incorrectly matched frames.

#	Video name	SURF (1)	SC (2)	SURF+SC (3)	CHE (5)	BA (6)
		IF	IF	IF	IF	IF
(13)	I will survive Jesus	9.2	10.5	**7.2**	19.2	15.4
(14)	Ronaldinho ping pong	12.5	11.1	**11.0**	20.9	19.5
(15)	White and Nerdy	15.6	10.2	**10.0**	25.6	21.8
(16)	Korean karaoke	26.0	35.5	**24.1**	32.5	31.0
(17)	Panic at the disco...	18.3	17.3	**15.0**	22.6	25.5
(18)	Bus uncle	27.2	42.6	**25.7**	28.5	32.6
(19)	Sony Bravia	20.7	40.2	**30.1**	31.5	33.3
(20)	Changes Tupac	35.1	20.6	**18.5**	40.7	38.2
(21)	Afternoon delight	12.5	11.2	**11.1**	19.6	20.5
(22)	Numa Gary	14.5	40.2	**14.1**	18.6	16.2
(23)	Shakira hips don't lie	40.3	36.2	**33.5**	41.3	42.5
(24)	India driving	49.3	25.6	**23.1**	52.8	50.9

changes widely vary wavelet coefficients and lead to lot of false positives.

Table 6 lists the registration accuracy of five compared methods for 13–24 master videos of CC_WEB_VIDEO dataset in terms of IF rates. Among all the methods, method (3) yields more accurate results due to the combined usage of visual and acoustic features against various types of formatting and editing attacks.

Chupeau et al.'s method [11] performs well for 22-nd and 13-th master videos but not as well for the 23-rd and 24-th master videos. This is because, color histograms are robust against lighting changes that are applied to the former videos, while the latter videos suffer from combined lighting and editing attacks.

On the other hand, Baudry et al.'s method [13] yields less accurate results for 24-th video in terms of higher IF rate. The reason is, 24-th video is modified by editing differences such as overlay text and addition of borders around frames, which in turn noticeably vary wavelet coefficients.

6.3.4. *Registration results for camcorded videos*

In the subsequent experiments, for comparison purpose we evaluated the following three methods:

(1) Chupeau et al.'s method [11] (CHE);
(2) Baudry et al.'s method [13] (BA);
(3) SURF+SC+sliding window for matching (Proposed).

Fig. 7(a) lists the registration results of three compared methods in terms of MF and AD rates. The proposed method gives extremely good results and improves the MF rates (up to 44%), compared to two reference methods. Although SURF and audio features have their own limitations, they balance each other; hence, the integrated utilization of visual and acoustic features significantly improves the registration accuracy. The promising results of method (3) against heavily modified camcorded copies of master videos support this view point.

Among all the three methods, the proposed method yields more accurate results, because the AD rate is lesser than the reference methods. Chupeau et al. [11] and Baudry et al. [13] methods score poor MF rates compared to the proposed method. The reason is, heavy cropping

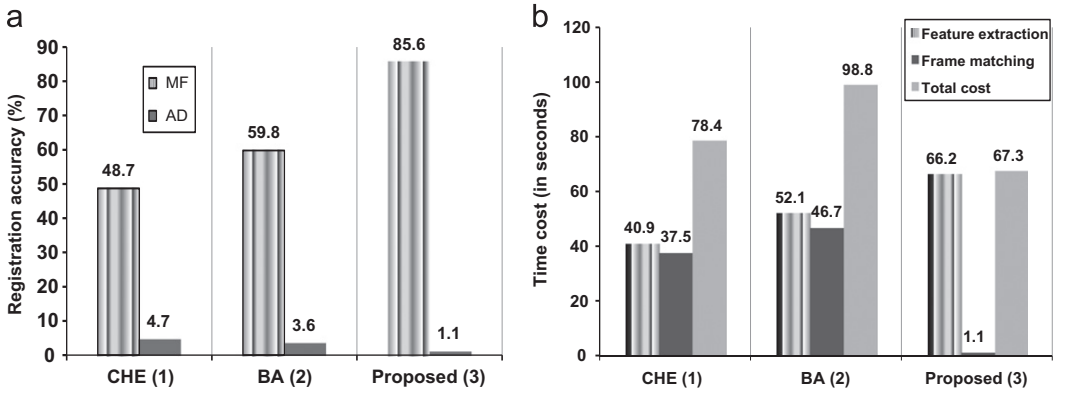

Fig. 7. Comparison of accuracy and time cost: (a) MF & AD rates; (b) total time cost.

and compression distortions might substantially alter visual descriptors such as color histograms and wavelets coefficients-based signatures.

Fig. 7(b) shows the total time costs of methods (1)–(3), which includes feature extraction and frame matching costs. They are measured by implementing the frame-to-frame mapping of a 215 s query sequence with 2493 s master sequence. Although the feature extraction cost of proposed method is higher, its frame matching cost is lower (by 97.5%) compared to the reference methods. This is because, query clip features are aligned only with the candidate segment instead of the entire master sequence. Thus, in the proposed method, usage of sliding window scheme noticeably reduces the total time cost up to 46.8% and provides lowest computational cost.

6.4. Geometric registration results

Table 7 shows the geometric registration results of the proposed method for different video transformations in terms of mean and maximum pixel distances. Although the query video (i.e., camcorded version of the master video) is modified by heavy cropping, lighting and compression attacks; still the proposed method provides more accurate results in terms of low pixel distances. The spatial registration performance of the proposed method is very efficient, because the mean pixel distance is always less than one. The robust nature of powerful SURF descriptors is the exact reason for this enhanced performance of the proposed method.

Table 7
Geometric registration results.

Attacks	Mean distance	Maximum distance
Zoom in	0.60	1.20
Pattern insertion	0.62	1.30
Moving caption	0.62	1.24
Rotation	0.85	1.62
Random noise	0.63	1.30
Blurring	0.62	1.18
Brightness change	0.59	1.17
Cropping	0.63	1.41
Picture-in-picture	0.85	1.67
3 combined	0.62	1.12
5 combined	0.64	1.29
Camcording	0.69	1.23

For illustration purpose, we considered temporally aligned master and query sequences, consisting of 984 and 375 frames, respectively. We selected 74 principal frames from the temporally aligned video segments using Algorithm 2 described in Section 5 and utilized them for the geometric alignment task. Fig. 8 shows the geometrical mapping of the sample candidate and query frames, in which extracted control points are highlighted with crosses. Here, query video is generated as the camcorded version of the master video.

Summary. The experiments conducted on different datasets demonstrate that the proposed method consistently outperforms the reference methods for different types of transformations. It achieves promising results with higher MF and AD rates, by integrating visual and acoustic features for the registration task. Frame matching in a sliding window manner is another good characteristic of the proposed method which proves that effective performance can be achieved with lowest computational cost, though the feature extraction cost is higher.

7. Conclusion

This paper proposes a novel spatio-temporal framework by utilizing local and spectral features for aligning master and pirate contents. It can be used for video forensic activities such as estimating distortion model and capture location in a theater. Further, the proposed framework can be utilized for sensor forensics, which attempts to identify the acquisition device that captured the video.

To the best of our knowledge, this is the first paper to discuss the frame alignment of master and pirate videos by exploiting content-based multimodal features. The experiments carried out on three different datasets demonstrate the promising results of the proposed method compared to the reference methods. Our future work will focus on how to enhance the robustness of proposed framework against attacks such as strong encoding, mix with speech and changing the background/foreground content.

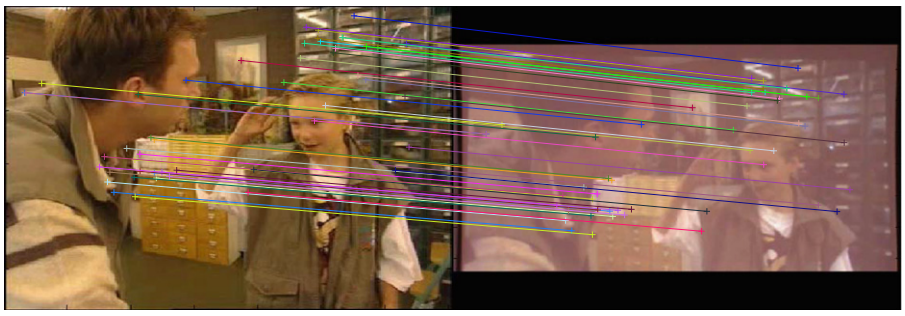

Fig. 8. Pairs of matched interest points of candidate (left) and query (right) video frames; here, query is camcorded copy of the master video.

References

[1] Economic consequences of movie piracy, CMPDA-February 2011 Report. 〈http://www.mpa-canada.org/press/IPSOS-OXFORD-ECONOMICS-Report_February-17-2011.pdf〉.

[2] S. Wei, Y. Zhao, C. Zhu, C. Xu, Z. Zhu, Frame fusion for video copy detection, IEEE Transactions on Circuits and Systems for Video Technology 21 (1) (2011).

[3] A. Sarkar, V. Singh, P. Ghosh, B.S. Manjunath, A. Singh, Efficient and robust detection of duplicate videos in a large database, IEEE Transactions on Circuits and Systems for Video Technology 20 (6) (2010).

[4] Jian Lu, Video fingerprinting for copy identification: from research to industry applications, in: Proceedings of SPIE, Media Forensics and Security XI, vol. 7254, 2009.

[5] B. Chupeau, A. Massoudi, F. Lefèbvre, In-theater piracy: finding where the pirate was, in: Proceedings of SPIE, Security, Forensics, Steganography & Watermarking of Multimedia Contents X, vol. 6819, USA, 2008.

[6] F. Lefèbvre, B. Chupeau, A. Massoudi, E. Diehl, Image and video fingerprinting: forensic applications, in: Proceedings of SPIE and IS & T, Journal of Electronic Imaging, vol. 7254, 2009.

[7] D. Delannay, C. de Roover, B. Macq, Temporal alignment of video sequences for watermarking, in: IS & T/SPIE 15th Annual Symposium on Electronic Imaging, vol. 5020, USA, 2003, pp. 481–492.

[8] H. Cheng, Temporal registration of video sequences, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, China, 2003, pp. 489–492.

[9] H. Cheng, M.A. Isnardi, Spatial, temporal and histogram video registration for digital watermark detection, in: Proceedings of International Conference on Image Processing (ICIP 2003), Spain, 2003, pp. 735–738.

[10] H. Cheng, A review of video registration methods for watermark detection in digital cinema applications, in: Proceedings of ISCAS 2004, 2004, pp. 704–707.

[11] B. Chupeau, L. Oisel, P. Jouet, Temporal video registration for watermark detection, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), vol. 2, France, 2006, pp. 157–160.

[12] S. Baudry, B. Chupeau, F. Lefèbvre, A framework for video forensics based on local and temporal fingerprints, in: Proceedings of IEEE International Conference on Image Processing (ICIP 2009), 2009, pp. 2889–2892.

[13] S. Baudry, B. Chupeau, F. Lefèbvre, Adaptive video fingerprints for accurate temporal registration, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), 2010, pp. 1786–1789.

[14] Y.Y. Lee, C. Kim, S. Lee, Video frame matching algorithm using dynamic programming, in: Proceedings of SPIE and IS & T Journal of Electronic Imaging, vol. 18(1), 2009.

[15] D. Delannay, F. Delaigle, H. Demarty, M. Barlaud, Compensation of geometrical deformations for watermark extraction in digital cinema applications, in: Proceedings of SPIE Electronic Imaging 2001, Security and Watermarking of Multimedia Content III, vol. 4314, 2001, pp. 149–157.

[16] B. Chupeau, A. Massoudi, F. Lefèbvre, Automatic estimation and compensation of geometric distortions in video copies, in: Proceedings of SPIE, Visual Communications and Image Processing, vol. 6508, USA, 2007.

[17] A. Saracoğlu, E. Esen, T.K. Ateş, B. Acar, Ü. Zubari, Ezgi C. Ozan, E. Özalp, A. Aydin Alatan, Tolga Çiloglu, Content based copy detection with coarse audio–visual fingerprints, in: VII International Workshop on Content-Based Multimedia Indexing (CBMI), 2009, pp. 213–218.

[18] H. Bay, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, Computer Vision and Image Understanding (2008) 346–359.

[19] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall Signal Processing Series, New Jersey, , 1993.

[20] Tae Hong Park, Introduction to Digital Signal Processing—Computer Musically Speaking, World Scientific Press, 2010.

[21] G. Roth, R. Laganière, P. Lambert, I. Lakhmiri, T. Janati, A simple but effective approach to video copy detection, in: Proceedings of Canadian Conference on Computer and Robot Vision, 2010.

[22] Z. Zhang, C. Cao, R. Zhang, J. Zou, Video copy detection based on speeded up robust features and locality sensitive hashing, in: Proceedings of 2010 IEEE International Conference on Automation and Logistics, Hong Kong and Macau, 2010.

[23] G. Yang, N. Chen, Q. Jiang, A robust hashing algorithm based on SURF for video copy detection, Elsevier Computers & Security 31 (2012) 33–39.

[24] Kris West, Novel Techniques for Audio Music Classification and Search, Doctoral Thesis, 2008.

[25] A. Eronen, A. Klapuri, Musical instrument recognition using cepstral coefficients and temporal features, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), vol. 2, 2000, pp. 1753–1756.

[26] R. Roopalakshmi, G. Ram Mohana Reddy, A novel approach to video copy detection using audio fingerprints and PCA, Elsevier Procedia Computer Science Journal, 2011. http://dx.doi.org/10.1016/j.procs.2011.07.021.

[27] Meinard Müller, Information Retrieval for Music and Motion, XVI, edition I, Springer Press, 2007.

[28] Pavel Senin, Dynamic Time Warping Algorithm Review, Information and Computer Science Department, University of Hawaii, 2008.

[29] TRECVID 2010 Guidelines [Online]. Available: 〈http://www.nlpir.nist.gov/projects/tv2010/tv2010.html〉.

[30] CC_WEB_VIDEO: Near-Duplicate Web Video Dataset. 〈http://vireo.cs.cityu.edu.hk/webvideo/〉.