Robust Features for Accurate Spatio-Temporal Registration of Video Copies

R. Roopalakshmi and G. Ram Mohana Reddy
Information Technology Department
National Institute of Technology Karnataka (NITK)
Surathkal, Mangalore, 575025, India
Email: {roopanagendran2002, profgrmreddy}@gmail.com

***Abstract*—Pirate copies of movies are proliferating on the Internet and causing huge piracy issues. Any anti-piracy strategy requires not only copy detection but also precise frame alignment of pirate video with master content, prior to the estimation of geometric distortions and capture location in a theater. Most studies in pirate video registration focus on the alignment of watermarked sequences, while few efforts are made to align non-watermarked videos using content-based features. In this paper, we propose a spatio-temporal scheme for aligning pirate and master contents using visual features, which consists of three stages: First, a video sequence is compactly represented using 1-D SURF (Speeded-Up Robust Features) signatures; Second, temporal frame alignments are computed using sliding window based dynamic programming method; Third, robust SURF descriptors are employed to generate spatial frame alignments. The results demonstrate the improved registration accuracy of the proposed method against various transformations.**

***Index Terms*—Video copy, temporal registration, geometric alignment, SURF, dynamic programming.**

I. Introduction

With the rapid growth of online streaming activities, illegal videos can be easily created and distributed, which cause a huge loss to the movie industry. CMPDA-2011 report (Canadian Motion Picture Distributors Association) alarms that, 133M pirated movies are watched in Canada in 2010 [1]. Thus rigorous forensic analysis frameworks and countermeasures are required for preventing illegal movie captures.

In case of camcorder capture in a theater, significant mismatches exist between the master and pirate video sequences. A master sequence corresponds to a database video, whereas a pirate sequence corresponds to a transformed video, derived from the master content by means of various transformations such as cropping, camcording and frame rate changes. The mismatches between two video sequences could be temporal, geometric or combination of both. Thus, for a number of applications such as estimating distortion model, identifying capture location and detecting embedded forensic watermarks, we need to first register the copied sequence with the master content [2]-[4].

Delannay et al. [5] proposed a registration technique based on keyframes to temporally align two video sequences. Cheng [6] presented a temporal registration algorithm for aligning two video sequences, which is severely affected by transformations such as noise addition. Cheng and Isnardi [7] developed a spatial, temporal and histogram based registration algorithm for detecting digital forensic watermarks.

Chupeau et al. [8] presented a temporal registration scheme to align two video sequences using color histograms. However, this method performs poor for region-based transformations. Baudry et al. [9] employed both the global and local fingerprints for registering video sequences, but this method gives poor results for low motion frames and complex transformations such as letter-box insertion and subtitles. In [10], the same authors focused on the temporal registration of two video sequences by utilizing wavelets-based fingerprints, which are computationally expensive.

Accurate spatio-temporal alignment of master and pirate contents is prerequisite, to carry out video forensic activities such as estimating distortion model and identifying camcorder capture location in a theater. However, existing registration schemes are concentrating on the alignment of watermarked documents, while very few attempts are made to align non-watermarked sequences using content-based features. Furthermore, CBCD algorithms proposed so far, may not address frame-to-frame alignment of two video sequences, once a copy is detected.

In this paper, we focus on the spatio-temporal alignment of master and pirate video sequences by exploiting robust visual signatures, in order to achieve accurate frame-to-frame mappings of both the video sequences. In this work, we utilize 1-D SURF signature derived from SURF interest points [11] for the temporal alignment task, which is compact compared to existing multi-dimensional SURF signatures [12], [13]. In addition, sliding window based dynamic programming technique is employed to reduce the fingerprint matching cost of the proposed framework.

II. Proposed Framework

We propose a spatio-temporal framework shown in Fig. 1 to align master and pirate video contents, which consists of two phases namely, temporal and geometric frame alignments. In the first phase, when a copy clip is given, we segment the master sequence using a sliding window of length equal to the copy clip. After this, the similarity between the copy clip and each windowed segment is computed based upon their 1-D SURF signatures. Then, the windowed segment with

978-1-4673-2014-6/12/$31.00 ©2012 IEEE

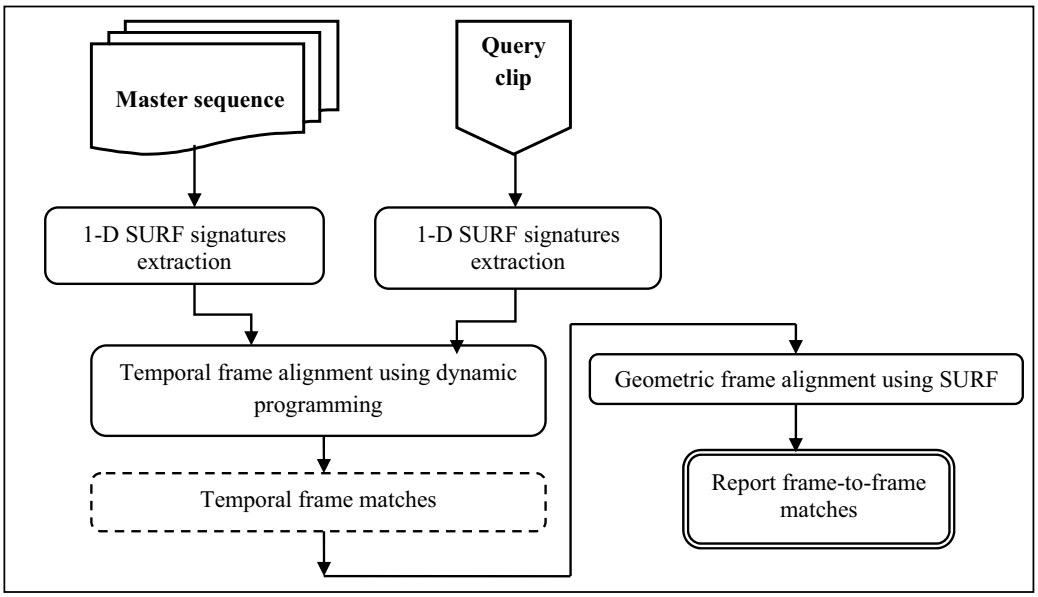

Fig. 1. Overview of the proposed registration framework

minimum distance score is denoted as a candidate segment, and optimal frame-to-frame alignments are computed using dynamic programming, which is detailed in Section III.

In the second phase, SURF descriptors of temporally aligned frames are mapped using enough control points in order to achieve robust geometrical alignments, which is explained in Section IV.

III. Temporal Alignment of Frames

A. Fingerprints Extraction

In the proposed framework, we utilize 1-D SURF signatures for the temporal registration task. SURF is an interest point based feature [11], which is popularly used in CBCD literature to detect illegal video clips [12], [13].

SURF descriptor associates each interest point with a high dimensional feature vector, which is typically 64 integers per interest point. Since each frame contains multiple interest points, there would be too much of information to index and search. In addition, direct comparison of SURF descriptors across all frames would be computationally expensive. On the other hand, robust visual signature describing both spatial and temporal information is required to achieve accurate frame-to-frame alignments.

In order to solve these issues, we compute a 1-D SURF signature by combining spatial and temporal information. More precisely, a video frame is segmented into $k \times k$ regions and the 1-D SURF signature is computed as the mean value of region-wise count of SURF interest points of a frame. However, the segmentation of a frame into $k \times k$ regions, plays a significant role in determining registration accuracy. In this study, the value of k is set as 3, after implementing experiments for different k values ranging from 2-5.

B. Introduction to Dynamic Programming

Dynamic programming is an efficient recursive technique, which is popularly used in sequence-to-sequence alignment and comparison methods [14]. Two feature sequences can be optimally aligned using dynamic programming as follows:

a) Computing minimum score matrix: In order to specify the optimal alignment between two sequences, first 2-D score matrix computation is needed. An element $SM(i,j)$ of score matrix SM gives minimum matching cost to match the subsequences [0,1,.., i] with [0,1,.., j], which is recursively computed as,

$$SM(i,j) = Min \begin{cases} SM(i-1,j-1) \\ SM(i,j-1) + W_h \\ SM(i-1,j) + W_v \end{cases} + D(i,j) \quad (1)$$

where W_h, W_v are penalties associated with horizontal and vertical directions and $D(i,j)$ is the difference between two feature sequences associated with the elements i and j.

b) Determining the optimal alignment path: optimal frame-to-frame matches are computed, by performing the trace-back step starting from diagonal element to the top.

C. Sliding Window Based Dynamic Programming

The computational complexity of dynamic programming 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 discrepancy, we perform frame matching between the copy clip and a candidate segment instead of the entire master sequence. Fig. 2 explains the steps used to select the candidate segment of master sequence.

1: Segment the master sequence into overlapping blocks of length equal to the query clip.
2: Extract 1-D SURF signatures for each segment using the procedure explained in Section III. A.
3: Let a master sequence

$$MS = \{s_1, s_2, s_3, .., s_m\}, \quad (2)$$

where s_i is the i^{th} segment and m is the total segments of MS. Here, the 1-D SURF signatures of s_i are denoted as,

$$s_i \in \{SF_i^k\}_{k=1}^n, \quad (3)$$

where SF_i^k is k^{th} visual signature of segment s_i.
4: Let QS be a query clip and the 1-D SURF signatures of QS are denoted as,

$$QS \in \{QSF^k \mid k = 1, 2, ..., n\} \quad (4)$$

where QSF^k is k^{th} visual signature and n is the total SURF signatures of QS.
5: The similarity sim_{seg} between QS and the segment s_i is computed using Manhattan distance as follows,

$$sim_{seg}(s_i, QS) = \sum_{k=1}^{n} |SF_i^k - QSF^k| \quad (5)$$

6: A master segment with minimum sim_{seg} value (i.e. $sim_{seg} \leq$ threshold) is selected as the candidate segment of MS. In this work, the threshold is set as 0.48, after implementing experiments for different values ranging from 0.30-0.60.

Fig. 2. Algorithm: Candidate segment selection

D. Frame Alignment using Dynamic Programming

Consider CS be a candidate segment of the master sequence and QS be a query clip. Let SF_{cs}^k and QSF^k are the 1-D SURF signatures of segments CS and QS denoted by,

$$CS \in \{SF_{cs}^k\}_{k=1}^n \quad , \; QS \in \{QSF^k\}_{k=1}^n \quad (6)$$

The distance between the visual signatures of CS and QS is computed using comparative Manhattan distance as follows,

$$dist_{surf}(CS(j), QS(j)) = \frac{|SF_{cs}^j - QSF^j|}{|SF_{cs}^j| + |QSF^j|} \quad (7)$$

where $j = [1 : n]$ and n is the total SURF signatures of two video segments. Then score matrix SM is computed using Equations (1) and (7). After this step, the optimal alignment path is determined and temporal frame alignments (TFA) based on SURF signatures is computed as,

$$TFA \in \{cv_x, qv_y\}, \quad 1 \leq x \leq n, \quad 1 \leq y \leq n \quad (8)$$

Here, cv_x and qv_y indicate the frame matches of candidate and query video segments respectively.

IV. Geometric Alignment of Frames

Performing geometric alignment across all temporally aligned frames is not feasible due to computational load. Furthermore, all video frames may not provide necessary interest points to enable accurate geometric registration.

In order to handle these issues, we exploit a small set of representative frames for the geometric registration framework. The SURF descriptors and the score matrices computed for the temporal alignment provide significant guidelines to select the representative frames.

More precisely, frame pairs with lower distance score (i.e. $dist_{surf}$ value of Equation (7)) are considered and mapped in terms of their descriptors, in order to provide accurate pixel correspondences of frames. Two control points are matched, only if the squared Euclidean distance between their feature vectors is minimum. Fig. 3 shows the sample candidate and query segment frames, which are geometrically mapped in terms of their interest point pairs. Here, copy video is created by applying random noise transformation.

V. Experimental Setup and Results

A. Master Database and Query Dataset Construction

The proposed method is evaluated on 100h of TRECVID 2009 Sound & Vision data [15], plus another 30h of real data consisting of camcorded copies of master video sequences. All the video clips are converted into uniform format: 352×288 pixels and 15 frames/sec using resampling technique.

Table I lists the video transformations used in the proposed framework, which cover most of the manipulations specified in TRECVID-2009 copy detection task. From the master database, 45 video clips of duration 20-45s are randomly selected and Table I transformations are applied to generate the query dataset. In addition, 48 camcorded copies of duration 35-115s are generated from 25 master sequences. Thus, the resulting 633 ((45 ×13) + 48) video sequences are used as query clips for the proposed temporal registration task. Geometric registration is performed on a set of 32 representative frames selected from the temporally aligned query and candidate segments.

TABLE I
List of transformations used in the proposed framework

#	Category	Description
T1	Rotation	Rotating by 15°-20°
T2	Random noise	Add 20% gaussian noise
T3	Blurring	Blur by 21%
T4	Brightness change	Increase brightness by 15%
T5	Cropping	Crop top & bottom regions by 25% each
T6	Picture-in-picture	Insert smaller resolution picture
T7	Zoom in	Zoom in to the frame by 18%
T8	Slow motion	Halve the video speed
T9	Fast forward	Double the video speed
T10	Pattern insertion	Insert text pattern into selected frames
T11	Moving caption	Insert moving titles into entire video
T12	3 combined	20% cropping, 15% noise & moving caption
T13	5 combined	17% noise, 20% blurring, 17% brightness, cropping & pattern insertion

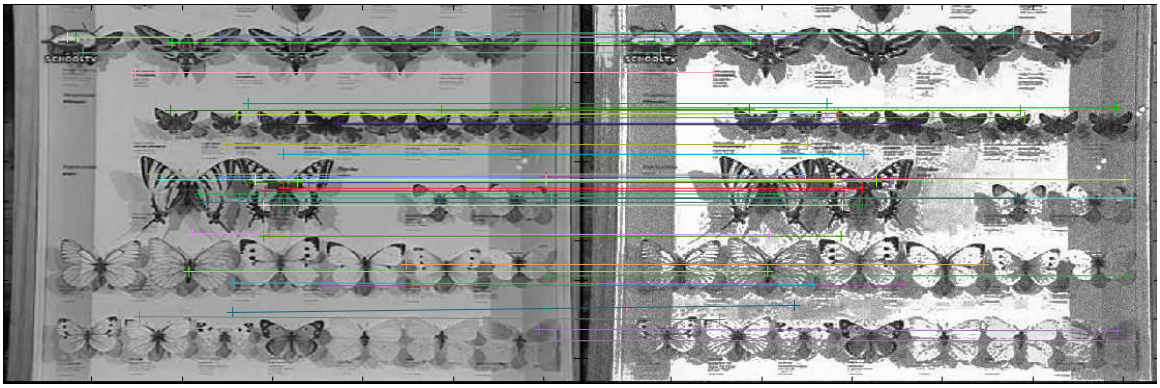

Fig. 3. Pairs of matched interest points of candidate (left) and query (right) frames. Random noise transformation is applied.

B. Overview of Evaluated Methods

We compare the accuracy of following three methods:

(1) 1-D SURF signatures (abbreviated as 'SURF');
(2) Chupeau et al.'s method [8] ('CHE');
(3) 1-D SURF signatures + sliding window ('ALL');

Our methods [methods (1) & (3)] evaluated different combinations of the proposed techniques. In method (1), 1-D SURF signatures of the query clip are matched with corresponding features of the entire master sequence (i.e. query clip is matched with all segments of the master sequence).

Chupeau et al.'s method [8] utilizes color histograms for calculating frame-to-frame correspondences between the query and master video sequences. 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 video files.

In method (3), we use sliding window mechanism to align SURF signatures of the query segment with that of the candidate segment instead of the entire master sequence. The candidate segment of the master sequence is selected using the procedure explained in Section III. C.

C. Temporal Registration Results

Table II shows the temporal registration results of three compared methods for T1-T7 types. The results are indicated in terms of percentage of perfectly Matched Frames (denoted as 'MF') and Average Distance between true and estimated frame indexes ('AD').

Method (3) scores better for all seven transformations and improves registration accuracy up to 31.5% compared to the reference method. The integrated utilization of robust SURF signatures and the sliding window scheme is the exact reason for the improved performance of method (3). In addition, method (3) yields more accurate results compared to method (2), as the AD values are always less than two.

On the other hand, Chupeau et al.'s method [8] yields poor results for cropping and picture-in-picture types in terms of low MF and high AD rates. This is because, cropping introduces black borders on top and bottom regions; hence, very different video signatures are generated for master and query segments.

TABLE II
TEMPORAL REGISTRATION RESULTS FOR T1-T7 TYPES. MF: % OF PERFECTLY MATCHED FRAMES; AD: AVERAGE DISTANCE BETWEEN FRAME INDEXES

Attacks	SURF (1)		CHE (2)		ALL (3)	
	MF	AD	MF	AD	MF	AD
Rotation	75.18	2.0	56.69	3.1	81.18	1.1
Random noise	86.31	1.9	51.46	1.9	88.64	1.2
Blurring	76.76	2.3	59.34	2.5	80.15	1.2
Brightness	83.18	2.8	58.61	2.8	85.98	1.7
Cropping	79.54	3.4	34.28	3.6	82.59	1.2
Picture-in-picture	78.26	2.1	36.14	2.9	82.64	1.5
Zoom in	74.42	2.2	61.37	2.5	77.25	1.1

Table III lists the temporal registration results of three compared methods for T8-T13 types. Method (3) generally performs well for all eight transformations and improves the registration accuracy (by 34.53%) compared to the reference method.

TABLE III
TEMPORAL REGISTRATION RESULTS FOR T8-T13 TYPES. MF:% OF PERFECTLY MATCHED FRAMES; AD:AVERAGE DISTANCE BETWEEN FRAME INDEXES

Attacks	SURF (1)		CHE (2)		ALL (3)	
	MF	AD	MF	AD	MF	AD
Slow motion	87.76	5.1	50.28	4.2	88.11	3.1
Fast forward	85.45	4.8	54.67	5.4	87.34	4.7
Pattern insertion	78.63	1.4	45.19	2.6	80.10	1.2
Moving caption	66.36	1.5	48.31	2.1	69.86	1.3
3 combined	85.49	5.2	50.28	5.2	88.42	3.9
5 combined	79.61	4.6	40.59	5.1	82.68	2.9

For pattern insertion and 5 combined types, the MF rates of method (2) decline sharply. This is because, inserting patterns/captions substantially changes histogram bin values. In case of combined types, histogram signatures of query and candidate segments are severely affected by the addition of

TABLE IV
GEOMETRIC REGISTRATION RESULTS

Attacks	Mean dist	Max dist
Rotation	0.912	1.592
Random noise	0.713	1.373
Blurring	0.777	1.346
Brightness change	0.810	1.368
Cropping	0.735	1.345
Picture-in-picture	0.628	1.349
Zoom in	0.665	1.160
Pattern insertion	0.681	1.283
Moving caption	0.625	1.346
3 combined	0.854	1.459
5 combined	0.881	1.496

gaussian noise and insertion of text patterns. Yet, our methods (1) and (3) using SURF signatures are less affected by this category.

D. Geometric Registration Results

Table IV shows the geometric registration scores of the proposed method for different transformations such as rotation, cropping and combined types. The registration results are indicated in terms of mean and maximum pixel distances between geometrically mapped query and candidate segment frames.

The 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 accurate geometric alignments.

E. Computation Cost Comparison

The proposed method is evaluated in MATLAB using a PC with 2.8GHz and 3GB RAM. Table V lists the total computational cost of all three methods including fingerprint extraction and frame matching costs. They are measured by implementing the frame alignment of a 315s query clip with the 2493s master sequence.

TABLE V
COMPARISON OF COMPUTATIONAL COST (IN SECONDS)

Computational Cost	SURF (1)	CHE (2)	ALL (3)
Fingerprint extraction	176.95	166.41	176.39
Frame matching	47.68	45.68	1.47
Total cost	224.63	212.09	177.86

The frame matching cost of method (3) is drastically reduced compared to other two methods. This is because, in method (3) query segment features are matched only with the respective candidate signatures instead of the entire master sequence; hence false positives are removed effectively and as a result frame matching cost is significantly reduced.

VI. CONCLUSION

In this article, we present an accurate spatio-temporal framework for aligning video contents by utilizing robust SURF features. The results prove that the proposed method significantly improves registration accuracy and widens the coverage to more number of transformations at the cost of a slight increase in fingerprint extraction cost. The proposed framework can be utilized for video forensic activities such as estimation of camcorder capture location in a theater.

Our future work will be focused on how to enhance the robustness of proposed scheme against attacks such as compression, strong encoding and gamma correction. For compression attacks, if the global or acoustic features are combined with the existing framework, then accuracy can be substantially improved.

ACKNOWLEDGMENT

The authors would like to thank reviewers for their valuable comments and suggestions, that improved the quality of this article. This research work is supported by Department of Science & Technology of Government of India under research grant no. SR/WOS-A/ET-48/2010.

REFERENCES

[1] "Economic consequences of movie piracy", CMPDA- Feb 2011 report. http://www.mpa-canada.org/press/IPSOS-OXFORD-ECONOMICS-Report_February-17-2011.pdf

[2] B. Chupeau, A. Massoudi and F. Lefèbvre, "Automatic Estimation and Compensation of Geometric Distortions in Video Copies", in proc. of SPIE, Visual Communication and Image Processing, vol. 6508, USA, 2007.

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

[4] D. Delannay, F. Delaigle, H. Demarty and M. Barlaud, "Compensation of Geometrical deformations for Watermark Extraction in Digital Cinema Applications", in proc. of SPIE Electronic Imaging 2001, Security and Watermarking of Multimedia Content III, vol.4314, 149-157, 2001.

[5] D. Delannay, C. de Roover and B. Macq, "Temporal alignment of video sequences for watermarking", in proc. of IS&T/SPIE's 15^{th} Annual Symposium on Electronic Imaging, Santa Clara, California, USA, Proc. Vol. 5020, pp. 481-492, 2003.

[6] H. Cheng, "Temporal Registration of Video Sequences", in proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, China, pp. 489-492, 2003.

[7] H. Cheng and M.A. Isnardi, "Spatial, temporal and histogram video registration for digital watermark detection", in proc. of Int. Conf. on Image Processing (ICIP 2003), Barcelona, Spain, pp. 735-738, 2003.

[8] B. Chupeau, L. Oisel, and P. Jouet, "Temporal video registration for watermark detection", in proc. of IEEE Int. conf. ICASSP 2006, vol. 2, pp. 157-160, France, 2006.

[9] S. Baudry, B. Chupeau and F. Lefèbvre, "A framework for video forensics based on local and temporal fingerprints", in proc. of IEEE Int. Conf. ICIP 2009, pp. 2889-2892, 2009.

[10] S. Baudry, B. Chupeau and F. Lefèbvre, "Adaptive Video Fingerprints for Accurate Temporal Registration", in proc. of IEEE Int. Conf. ICASSP 2010, pp. 1786-1789, 2010.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (surf)," Computer Vision and Image Understanding, 110(3), 346-359, 2008.

[12] G. Roth, R. Laganière, P. Lambert, I. Lakhmiri, and T. Janati, "A Simple but Effective Approach to Video Copy Detection", in proc. of Canadian Conf. Computer and Robot Vision, 2010.

[13] Z. Zhang, C. Cao, R. Zhang and J. Zou, "Video Copy Detection Based on Speeded Up Robust Features and Locality Sensitive Hashing", in proc. of IEEE Int. Conf. on Automation and Logistics, 2010.

[14] D. Sankoff, "The early introduction of dynamic programming into computational biology", Bioinformatics, 16, 1, 41-47, 2000.

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