Spatio-Temporal Video Copy Registration Using Multimodal Features

[1]Sathish Kumar.K, [2]Ganapathy.V

[1]M.Tech Scholar, [2]Professor
[1,2]Information Technology, SRM University
[1,2]Kancheepuram, Tamil Nadu, India
[1]sathishk_25@yahoo.com

***Abstract*—The exponential growth of multimedia technologies and media streaming activities have increased video publishing and sharing activities tremendously. Because of this massive media consumption, huge number of pirated copies of videos are proliferating on the Internet causing huge piracy issues. Fighting video piracy requires copy detection followed by the accurate frame alignments of master and pirated videos, in order to estimate distortion model and capture location in the theatre. Existing research on pirate video registration utilizes only visual features for aligning pirate and master videos, while only less 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. To solve these issues, a novel and robust registration scheme using multimodal features is proposed, which is also suitable for content based methods such as Content-Based video Copy Detection (CBCD).**

***Index Terms*— Temporal registration, SURF, Spectral Centroid, Dynamic Time Warping.**

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.

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). CBCD techniques utilize content-based features of the media to detect illegal videos; hence, they are widely popular compared to digital watermarking.

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.

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 Motivation and Contribution:

Accurate temporal alignment of pirate and master sequence is a prerequisite step to carry out forensic analysis of video contents. However, existing temporal registration approaches are designed specifically for detecting forensic watermarks using visual features. The temporal registration using only the visual content of video files may not be sufficient to provide reasonable registration accuracy. In addition to visual features, audio content is an indispensable and essential information source of a video sequence. Also, from the movie piracy perspective, the audio content of a illegal video is less affected compared to the visual content.

Hence a novel temporal registration approach making use of multimodal features is required, which can be used even in the absence of forensic watermarks.

We propose a novel temporal registration framework that utilizes both the visual and audio features for matching copied and original video sequences. We used visual fingerprints extracted from SURF interest points [2] and audio signatures based on spectral centroid features [3] for the proposed frame matching task. Dynamic Time Warping (DTW) method is utilized to align the feature sequences of both the master and copied video contents. Sliding window scheme is implemented to reduce the fingerprint matching cost of proposed registration framework.

2. Proposed Framework

We propose a novel and robust framework for temporally registering the duplicate and master video contents, which is shown in Figure 1. To make this registration efficient, we used compact spatio-temporal signatures derived from SURF interest points and spectral centroid features. To make this registration cost effective, we used dynamic time warping method for providing accurate frame-to-frame matches and applied sliding window method for reducing the size of signatures.

When a query clip is given, we divide the master sequence into non overlapping segments of size equal to number of query frames. Then we perform segment-wise scanning of the master sequence using a sliding window of length equal to query clip. The similarity between the query clip and the windowed segment is computed based upon their 1-D signatures derived from SURF descriptors and spectral centroid features. The windowed segment with minimum dissimilarity score (score below a predefined threshold) is denoted as a most similar segment, and it is further analyzed using dynamic time warping method to determine frame-to-frame alignments.

3. Temporal Frame Alignment

3.1 Problem Formulation

Let $M = \{x_i \mid i = 1,2, ..., m\}$ be a master video sequence, where x_i is the i-th frame of master sequence. Let $Q = \{y_j \mid j=1,2, ..., n\}$ be a query clip, where y_j is the j-th frame of query clip and $m >> n$. Here Q is derived by applying different types of video transformations (blurring, rotation, scaling, mp3 compression etc.,) to one or more subsequences
of M. Our goal is to determine the exact location of the subsequence $R = \{r_k \mid k=1,2, ..., i+n-1\}$ in M , such that Q matches M and as a result frame-to-frame alignments of Q and R can be obtained. This temporal registration process consists of two steps. First, temporal signatures are derived from visual and acoustic features of two video files. Second, the

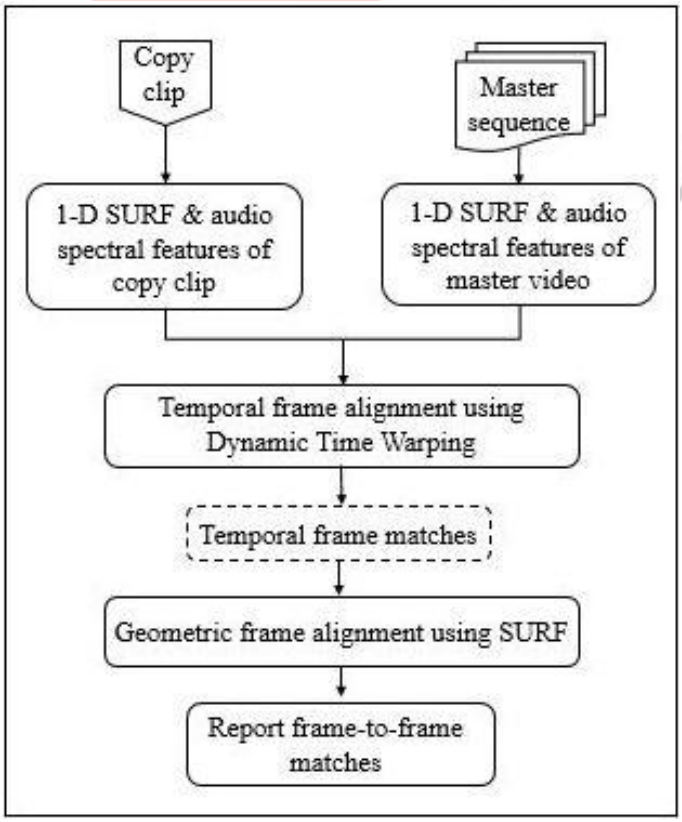

Fig. 1. Overview of proposed framework

temporal signatures of master and query segments
are matched using dynamic time warping, in order to obtain a list of frame-to-frame matches.

3.2 Video Signatures Generation

In the proposed framework, we utilize 1-D SURF signatures for the temporal registration task. SURF is an interest point based feature [2], which is popularly used in CBCD literature to detect illegal video clips [4], [5**].**

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×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.

3.3 Acoustic Signatures Generation

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 [6, 7]. Specifically, centroid is a highly robust spectral feature that describes brightness of a sound signal [8]; hence, it is popularly used in speech recognition applications [9]. 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.

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% [10]. From the power spectrum of the audio signal, the Spectral Centroid descriptor SC is computed using frequency distribution values as follows:

$$SC = \sum_{k=1}^{N} k \times X^d[k] \,/\, \sum_{k=1}^{N} X^d[k]$$

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 [10], 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.

3.4 Introduction to Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. It aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match between the two sequences is found.

For instance, similarities in walking patterns could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data which can be turned into a linear sequence can be analyzed with DTW.

3.5 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 most similar segments instead of the entire master sequence. Algorithm 1explains the steps used to select a most similar segments of the master sequence.

Algorithm 1. Selection of a most similar segments

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.

3: Let a master sequence MS be $MS \in \{S_i | 1 \le i \le m\}$

where S_i the i-th segment and m is total segments of MS. Here, each segment S_i of MS can be represented as

$$S_i \in \left\{\left(V_i^k \cup A_i^r\right) \middle| 1 \le k \le n,\ 1 \le r \le p\right\}$$

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

$$PS \in \left\{\left(QV^k \cup QA^r\right) \middle| 1 \le k \le n_q, 1 \le r \le p_q\right\}$$

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:

$$Seg_{sim}(S_k, PS) = PC_{dtw}(V_k, QV) + PC_{dtw}(A_k, QA)$$

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 most similar 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

The visual signatures of the most similar segment MS of the master sequence found using the algorithm in section 3.5 and the pirate sequence PS with n_s signatures are compared to find the dissimilarity score between the frames of two video sequences. The cost measure C_{vis} denoting the dissimilarity between two visual signatures is computed using comparative Manhattan distance metric as follows:

$$C_{vis}(MS, PS) = \frac{|(MS_s - PS_s)|}{|(MS_s)| + |(PS_s)|}, 1 \leq s \leq n_s$$

where MS_s, PS_s be the visual signatures of the most similar segment of master sequence and the pirate sequence respectively. The resultant frame matches FM_{vis} based on visual signatures is formulated as

$$FM_{vis} = \{(cv_i, pv_i) | 1 \leq i \leq n_s\}$$

where cv and pv indicate the matching frames of most similar segment and pirate video sequence, respectively.

4.2 Frame matching using acoustic signatures

The acoustic signatures of two sequences with n_a are compared to find dissimilarity score. The cost measure C_{aud} denoting the difference between two audio signatures is computed using squared Euclidean distance as follows:

$$C_{aud}(MS, PS) = |(MS_a - PS_a)^2|, 1 \leq a \leq n_a$$

where MS_a, PS_a be the acoustic signatures of the most similar segment of master sequence and the pirate sequence respectively. The resultant frame matches FM_{aud} based on acoustic signatures is formulated as

$$FM_{aud} = \{(ca_i, pa_i) | 1 \leq i \leq n_a\}$$

where ca and pa indicate the matching frames of most similar segment and pirate video sequences, respectively.

4.3 Decision fusion

Frames mapped by both the visual and audio signatures are considered as final frame matches of two video contents, which is given by

$$FM_{final} = \{\{FM_{vis}\} \cap \{FM_{aud}\}\}$$

Where FM_{final} provides frame-to-frame alignments of most similar segment of master sequence and

Fig. 2. Pairs of matched interest points of candidate (left) and query (right) frames. Random noise transformation is applied.

pirate 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.

5. 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 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. 2 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. Fig. 3 shows that only relevant content are matched between frames based upon the threshold value set.

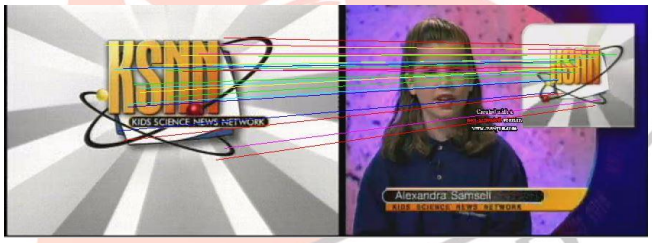

Fig. 3. Matching only the relevant content based on the threshold value.

6. 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 theatre. 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.

REFERENCES

[1] R. Roopalakshmi, G. Ram Mohana Reddy, Robust Features for Accurate Spatio-Temporal Registration of Video Copies, IEEE, 2012.

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

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

[4] G. Roth, R. Lagani` ere, 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.

[5] 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.

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

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

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

[9] 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.

[10] 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.

[11] 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.
[12] G. Yang, N. Chen, Q. Jiang, A robust hashing algorithm based on SURF for video copy detection, Elsevier Computers & Security 31 (2012) 33–39.
[13] Al-Naymat, G., Chawla, S., & Taheri, J. (2012). SparseDTW: A Novel Approach to Speed up Dynamic Time Warping.
[14] S.Baudry, Bertrand Chupeau and .Lef´bvreabc, "Adaptive Video Fingerprints for Accurate Temporal Registration", in proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010), pp. 1786–1789,2010.
[15] D. Delannay, C. de Roover and B. Macq, "Temporal alignment of video sequences for watermarking", IS&T/SPIE's 15 th Annual Symposium on Electronic Imaging, Santa Clara, California, USA, Proc. Vol. 5020, pp. 481-492, January 2003.
[16] S.Baudry, B.Chupeau and F. Lef´bvreabc, "A framework for video forensics based on local and temporal fingerprints", in proc. of IEEE International Conference on Image Processing (ICIP 2009), pp. 2889–2892, 2009.