ADAPTIVE VIDEO FINGERPRINTS FOR ACCURATE TEMPORAL REGISTRATION

Séverine Baudry, Bertrand Chupeau, Frédéric Lefèbvre

Security Competence Center
Thomson R&D France
1 avenue de Belle-Fontaine, 35576 Cesson-Sévigné Cedex, France

ABSTRACT

Designing watermarks that can withstand geometric and temporal attacks is a notoriously difficult problem. The registration of a pirate content with the reference content, prior to the decoding of an embedded forensic mark, is therefore preferred. Semi-automatic registration systems, requiring the expert user to manually select and match control points through a suitable user interface, are not acceptable for content owners and do not usually provide very accurate results. A fully automated registration scheme is therefore chosen to synchronize pirate and master contents. By extracting robust and unique features, digital fingerprinting provides a relevant framework for automatic semi-blind content registration. In this paper we focus on temporal registration. In order to guarantee an accurate alignment while keeping the fingerprint size as small as possible, we propose a temporally adaptive approach. The registration scheme is designed around an adaptive motion description based on hierarchical encoding of the wavelet coefficients computed on the difference of successive frames.

Index Terms—Video fingerprint, temporal registration, watermarking, progressive encoding, SPIHT

1. INTRODUCTION

Throughout its life from capture to display to the final user, a video content may undergo a wide range of transformations, for instance compression (MPEG2, H264, divX …), geometric and temporal transforms (HD/SD conversion, edition …) and colorimetric alterations. Some applications may need to perform accurate spatial, temporal, or luminance registration on the distorted videos. As detailed in [1], one watermarking challenge (especially in robust blind schemes) is temporal and spatial synchronization. A second watermark template can be added to the main watermark as a helper to estimate the transformations [2], but when invisibility becomes a critical constraint (e.g. in digital cinema or postproduction mastering), this may be problematic. Moreover, when the content undergoes huge distortions (for instance after camcording), the reference watermark may not be robust enough to give an accurate estimation of the transformation, and the synchronization between the reference watermark and the candidate is lost. In the scope of semi-blind watermark detection where the title of the pirated content is known (for instance in forensics watermarking), it is possible to perform registration using the original content itself [3]. However, when one comes to large scale data, processing, storing and accessing the whole original content may no be convenient. Therefore, it is useful to base the registration process rather on a “condensate” of the original video, called visual hash or visual fingerprint. In the design of such a fingerprint, one wants to minimize the size of the fingerprint data to store, while keeping the robustness and accuracy of the registration as high as possible. In this paper, we address the problem of temporal registration. When the video has been heavily distorted, detecting the watermark may require a very accurate registration to minimize the amount of additional noise. However, precise frame alignment can be difficult on static scenes with low motion. In section 2 we describe and give the rationales of the proposed fingerprints. Section 3 describes the algorithm to accurately register videos using fingerprints. Section 4 gives experimental results.

2. FINGERPRINT DESIGN

2.1. Information content of the fingerprint

Fingerprinting can be seen as an extreme compression scheme, which keeps only the relevant information for registration. For temporal registration purposes, the relevant information is the difference between consecutive frames, noted $D_t(x, y) = I_{t+1}(x, y) - I_t(x, y)$ ($I_t$ being the $t^{th}$ image of the video). Our fingerprint will thus be a highly compressed version of the frame difference $D_t(x, y)$. If there is high motion, $D_t$ energy and entropy will be large. On the contrary, on static scenes, $D_t(x, y)$ will be almost null everywhere, except on small areas corresponding to moving parts. To enable frame-accurate registration, the fingerprint must describe finely $D_t(x, y)$ at low motion, whereas a very coarse description is sufficient at large motion. In other
words, we want to encode $D_t(x,y)$ at fixed (and very small) bit-rate, to keep only the features relevant for temporal registration.

2.2. Hierarchical encoding of the fingerprint

To select a suitable compression scheme, we must take into account the constraints on the fingerprint. The compression scheme should preserve location information as well as amplitude information. It should also allow fine tuning of the level of information encoded. Scalable embedded coding schemes, like EZW (Embedded Zero-tree Wavelet) or SPIHT (Set Partitioning in Hierarchical Trees), meet these requirements.

EZW [4] and SPIHT [5] are state of the art compression methods based on wavelet decomposition. The main idea is to represent information in a hierarchical way, such that information is encoded in decreasing order of importance. A coarse approximation (first significance level) of low frequency coefficients will come first, and then refinement information will follow. Runs with no information (i.e. sets of null coefficients) are compactly encoded. Thus, as soon as the first bits of an EZW or SPIHT encoded stream are received, decoding can start to produce a very coarse approximation of the original image. Receiving further bits enables more and more refinement of the decoded image: this is a common property of every embedded (or progressive) coding scheme. Figure 1 shows examples of SPIHT encoding of the picture “Lena”, with increasing number of significance levels (bit-planes): the higher the bit-rate, the better the resolution and luminance precision.

For the fingerprint, we will select an energy threshold $E_T$. The difference image $D_t(x,y)$ will be wavelet transformed to yield the coefficients $W(x,y,l,o)$, with $l$ being the resolution level and $o$ the orientation (LL, LH, HL, HH) of the coefficient. These coefficients are then progressively encoded with the EZW or SPIHT algorithm. The encoding stops as soon as the energy of coefficients encoded so far exceeds $E_T$. Note that we use a fixed energy compression scheme rather than a fixed bit-rate scheme; however, bit-rate will be more or less constant due to the efficiency of the compression algorithm.

Figure 2 summarizes on an example the fingerprint computation. First, the difference between successive frames is computed (middle) and wavelet transformed. The wavelet coefficients are then progressively encoded (below). The approximation obtained by keeping only the coarser significance levels is displayed on left, while the approximation obtained by keeping more significance levels is shown on right.

3. TEMPORAL REGISTRATION USING FINGERPRINTS

3.1. Registration by dynamic programming

Dynamic programming is a state of the art algorithm which can be used to register any two sequences of data elements. It combines a priori and a posteriori information. A posteriori information is a similarity measure between pairs of data elements. A priori information introduces some “smoothness” constraint between registered samples. Temporal registration is performed by applying a dynamic programming algorithm between the original and the pirated fingerprint, in a similar way than [3]. SPIHT data corresponding to each image of both videos are decompressed, to yield approximations at significance level $n$ of the wavelet coefficients noted $F_t^n(x,y,l,o)$ for the original video and $G_t^n(x,y,l,o)$ for the pirated video. A
posteriori information is here given by the Euclidean distance:

\[
d(F^n_t, G^n_t) = \sqrt{\sum_{x,y,l,o} (F^n_t(x,y,l,o) - G^n_t(x,y,l,o))^2}
\]

Other similarity measures may be used, for instance a weighted Euclidean distance, to take into account the different sensitivity to noise of wavelet coefficients, depending on their level and orientation.

3.2. Confidence score and registration error

In some applications it is very important to give a confidence score along with the estimated registration. For instance, if registration is used to retrieve a watermark, the confidence score can be used to weight the confidence of the decoded watermark. Watermark detection rate may also be increased by trying to decode the watermark on adjacent frames if the registration confidence score is low on a given frame.

The registration confidence depends on a posteriori information (how close to the original one the registered pirate image is) as well as a priori information (how close from one another consecutive frames in the original video are): on a scene with low motion the registration algorithm is more prone to errors. A posteriori information can be approximated by the distance between original an pirated fingerprint \(d(F^n_t, G^n_t)\); a priori information by the distance between consecutive fingerprints of the original video, around the estimated index: \(d(F^n_{\hat{t}(t)}, F^n_{\hat{t}(t) + \tau})\).

The confidence level \(C(t)\) for pirate frame index \(t\) can then be given by an estimation of the SNR:

\[
C(t) = \frac{\min_{r \in \tau} d(F^n_{\hat{t}(t)}, F^n_{\hat{t}(r) + \tau})}{d(F^n_{\hat{t}(t)}, G^n_t)}
\]

with \(\hat{t}(t)\) being the (estimated) frame index in the original video corresponding to frame \(t\) of pirate video, and \(\tau\) defining a neighborhood around the estimated index.

3.3. Iterative registration

One disadvantage of the above registration is its cost in terms of CPU and memory. The SPIHT fingerprint must be decoded prior to computing the a posteriori distances. Moreover, the computation of distances may be costly at fine significance levels, because the number of coefficients taken into account can be large. To lower the CPU burden, we propose to perform the dynamic programming algorithm in two or more steps. The first step will perform a very rough estimation of registration, which will be refined in the later steps.

\[\text{In the first step, we decode only the coarser information of the SPIHT fingerprints, i.e. the most significant bits of the lowest resolution coefficients. Fingerprint decoding and distance computation are therefore here very fast. Dynamic programming is applied to yield a rough registration estimate.}\]

\[\text{On this first registration estimation, we compute the confidence score } C(t) \text{ for each frame, using roughly decoded fingerprints. We will then refine the registration only on images for which } C(t) \text{ is lower than a predefined threshold } T \text{ (see Figure 3). On these images, we further decode the SPIHT fingerprints to a finer level (by decoding more coefficients and with more significance bits). The dynamic programming algorithm is then applied to these finer decoded fingerprints, leading to more accurate results. The process can be iterated until the confidence is high enough on each frame.}\]

\[\text{Figure 3: Estimated matching after first registration step. Dotted boxes show the indexes with low confidence values, where registration will be refined during the 2nd step}\]

4. RESULTS

To assess the performances of our algorithm, we worked on the sequence “Talk Show” (see example images on Figure 2), which, since it contains mainly static plans with few motion, is quite challenging for temporal registration. This sequence is 500 images long. We compressed the sequence with divX at 400kbps, and then applied a temporal jitter: frames are randomly deleted or duplicated, at rate \(p = 0.05\).

Figure 4 shows the percentage of incorrectly registered frames after applying dynamic programming on fingerprints of various significance levels. Note that the accuracy increases with the number of significance levels. Even with very few significance levels the algorithm improves the registration accuracy (with no registration the percentage of misregistered frames is 95%).
An implementation of this concept using a SPIHT scalable embedded coding of inter-frame difference images showed a dramatic increase in the temporal registration accuracy, compared to fixed-size fingerprints. Moreover, a confidence measure is provided together with a given alignment, which allows designing an iterative registration scheme, refining the temporal analysis on low-confidence parts and thus concentrating the computational load where needed only. Such a semi-blind video temporal registration scheme can be seen as the pre-processing module of a more global forensic analysis framework [6]. Accurate alignment with master time-line is indeed a pre-requisite for a number of forensic application, not restricted to forensic watermark decoding. Future work could consist in applying the same dynamic description scheme to the audio signal and combining with video in a multimodal registration approach.

5. CONCLUSION

A novel scheme for accurate temporal registration of video copies with master content was presented. This semi-blind process is based on temporally adaptive fingerprints, finely describing low motion segments while coarsely describing high motion parts. An implementation of this concept using a SPIHT scalable embedded coding of inter-frame difference images showed a dramatic increase in the temporal registration accuracy, compared to fixed-size fingerprints. Moreover, a confidence measure is provided together with a given alignment, which allows designing an iterative registration scheme, refining the temporal analysis on low-confidence parts and thus concentrating the computational load where needed only. Such a semi-blind video temporal registration scheme can be seen as the pre-processing module of a more global forensic analysis framework [6]. Accurate alignment with master time-line is indeed a pre-requisite for a number of forensic application, not restricted to forensic watermark decoding. Future work could consist in applying the same dynamic description scheme to the audio signal and combining with video in a multimodal registration approach.

6. REFERENCES


