# SEMI-BLIND AUTOMATIC TEMPORAL REGISTRATION FOR VIDEO WATERMARKING SYSTEMS

*Bertrand Chupeau, Séverine Baudry, and Gwenaël Doërr*

Technicolor R&D France – Security & Content Protection Labs

## ABSTRACT

Despite the academic focus on blind detection, many forensic watermarking systems actually operate in a non-blind fashion to accommodate for real-life desynchronization attacks. In other words, the original content is needed to perform watermark detection e.g. to re-align the pirate sample with the original content prior to watermark decoding. A somehow intermediate strategy, that could be coined 'semi-blind', only requires access to a condensed representation of the original content i.e. content fingerprints. In this paper, we combined a variant of the dynamic time warping algorithm together with multi-dimensional temporal fingerprints to realign pirate video samples. Reported experimental results clearly demonstrate improved accuracy compared to state-of-the-art registration methods for a number of synthetic and real-life attack scenarios.

***Index Terms—*** Temporal registration, temporal fingerprints, dynamic time warping

## 1. INTRODUCTION

Synchronization between a watermark embedder and its associated detector is critical: if the two components are not aligned, the system will not see embedded watermarks even if they are present. As a result, it is common practice to realign the pirate sample with the original content prior to attempting recovering a watermark. Watermarking systems are classified as *non-blind*, *semi-blind* or *blind* depending on whether the detector requires access to, respectively, (i) the original content, (ii) a condensed representation of the original content e.g. some content fingerprints, or (iii) nothing. A watermark detector typically combines a watermark decoder and a resynchronization module and each one of those two components can be classified as non-/semi-/blind independently [1].

Forensic watermarking consists in embedding an imperceptible and robust signal in multimedia content in order to identify the recipient whom this piece of content has been delivered to. In this context, a real-life pirate video sample is likely to have been subject, either naturally or deliberately, to various photometric, spatial and temporal transforms [2]. We will solely focus on temporal distortions in this paper. A camcorder capture changes the frame rate, blends adjacent frames, and introduces interlacing artifacts. A screencast acquisition produces frame repeats and deletions to cope with a different, and possibly variable, frame rate. Video compression alone may randomly drop frames. Finally, a pirate may remove video segments here and there in an attempt to confuse the watermark detector. In view of this complexity, many watermarking system resort to non-blind or semi-blind resynchronization modules.

Section 2 briefly reviews several recent proposals to automatically register a video copy to the corresponding master and details the baseline Dynamic Time Warping (DTW) algorithm. In Section 3, we propose a number of modifications whose impact on registration accuracy is assessed in Section 4. Eventually, Section 5 summarizes the results that we obtained and provides an outlook for future work.

## 2. PRIOR ART

Early temporal realignment techniques relied on sparse matching of key-frames and local estimates of affine temporal transforms [3, 4]. However, with such an approach, the accuracy of the temporal mapping heavily depends on the density of extracted key frames. To mitigate this limitation, follow-up works introduced a dynamic programming framework which minimizes the Mean Square Error (MSE) between aligned frames while explicitly incorporating contextual constraints on feasible paths [5]. Subsequently, variants using content fingerprints instead of full frames were proposed in an attempt to minimize the amount of forensic metadata required to perform temporal realignment prior to watermark detection and to lower the sensitivity to spatial desynchronization [6, 7].

These realignment techniques typically use dynamic time warping (DTW), a generic algorithm to find the optimal mapping between a *reference* temporal sequence $\mathcal{R} = \{\mathbf{r}_i\}_{1 \leq i \leq n}$ and a *candidate* temporal sequence $\mathcal{C} = \{\mathbf{c}_j\}_{1 \leq j \leq m}$ which minimizes a global matching cost using dynamic programming [8]. To do so, a $n$-by-$m$ matrix is constructed recursively using the following Equation:

$$\gamma_{i,j} = \min\{\gamma_{i-1,j-1}, \gamma_{i,j-1}, \gamma_{i-1,j}\} + \mathrm{d}(\mathbf{r}_i, \mathbf{c}_j), \qquad (1)$$

where $\mathrm{d}(\mathbf{r}_i, \mathbf{c}_j)$ is some distance between the matched elements $(\mathbf{r}_i, \mathbf{c}_j)$. As a result, the matrix cell $\gamma_{i,j}$ corresponds to the accumulated cost of the warping path with minimal cost leading to this match. When the entire matrix has been filled, the optimal warping path can then be efficiently traced back, starting from $\min_i \gamma_{i,m}$.

## 3. CONTRIBUTIONS

### 3.1. Introducing Penalties

Equation (1) ignores the fact that horizontal and vertical transitions in the mapping path correspond to frame deletion and insertion respectively. In such situations, computing a distance between reference and candidate samples may have no physical meaning and it is preferable to introduce explicit penalties in the cost function, e.g.

$$\gamma_{i,j} = \min\{\gamma_{i-1,j-1} + \mathrm{d}(\mathbf{r}_i, \mathbf{c}_j), \gamma_{i,j-1} + \pi_{\mathrm{ins}}, \gamma_{i-1,j} + \pi_{\mathrm{del}}\}, \qquad (2)$$

where $\pi_{\mathrm{ins}}$ and $\pi_{\mathrm{del}}$ are the penalties assigned to horizontal and vertical transitions. This modification is expected to yield much more accurate mapping path when strong video editing (scene cut and insertion) has been applied to the reference video sequence.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

### 3.2. Penalties Iterative Adjustment

Assigning the right values to the penalties which we have just introduced is not straightforward. Intuitively, the penalty assigned to frame deletion/insertion should correspond to a distance that falls outside the probability distribution of 'normal' pairings. In order to get an estimation of this distribution, we first run the DTW algorithm with arbitrary penalty values, e.g. $\pi_{\text{ins}} = \pi_{\text{del}} = 1$. It is then possible to compute the mean and standard deviation $(\hat{\mu}, \hat{\sigma})$ of the distances $\text{d}(\mathbf{r}_i, \mathbf{c}_j)$ between matched elements along the diagonal portions of the warping path (i.e. excluding horizontal and vertical transitions). We then update the penalties for the next iteration of DTW as follows:

$$\pi_{\text{ins}} = \pi_{\text{del}} = \hat{\mu} + K.\hat{\sigma}, \quad K \geq 0, \tag{3}$$

and the process is iterated until the change ratio of the penalty values falls below a specified threshold (e.g. $10^{-3}$). In practice, we observed that convergence is obtained after five iterations in most cases. The parameter $K$ specifies how far the mapping path can deviate from a straight line and can thus be regarded as some kind of *rigidity* parameter.

### 3.3. Multidimensional Fingerprint Temporal Signatures

In the context of video temporal realignment, one could consider using the full frames as the input of the DTW algorithm. However, this strategy does not account for the fact that it might be difficult to get access to the reference video sequence to perform forensic watermark extraction. Additionally, it induces a significant computational burden. As a result, it is common practice to consider a condensed representation of the video signal and more particularly the temporal evolution of some still image fingerprint, aka. the fingerprint temporal signature [6, 7].

While a number of fingerprints have been proposed in the literature, we will focus in the remainder of this paper on the so-called RASH descriptor due to its natural robustness to photometric and geometric distortions [9]. This fingerprint essentially reduces to computing the variance of the luminance along radial strips passing through the image center, yielding a 180-D vector with a 1° angle discretization. Previous works on temporal realignment reduced the temporal dynamics of RASH to the norm of its gradient:

$$\text{S}(\mathcal{F}, n) = \sqrt{\sum_{\theta=1}^{180} \Big(\text{RASH}_\theta(\mathbf{F}_n) - \text{RASH}_\theta(\mathbf{F}_{n-1})\Big)^2}, \tag{4}$$

where $\mathcal{F} = \{\mathbf{F}_n\}$ is a sequence of video frames and $\text{RASH}_\theta(.)$ is the component of the RASH vector associated to the direction $\theta$.

In practice, this crude reduction of the temporal dynamics to a scalar value fails to provide accurate registration for some videos with slow motion. In order to increase the granularity of the motion information captured by the temporal signature, we computed the norm of the gradient on disjoint portions of the RASH vector:

$$\text{S}(\mathcal{F}, n, k) = \sqrt{\sum_{\theta=\theta_k+1}^{\theta_{k+1}} \Big(\text{RASH}_\theta(\mathbf{F}_n) - \text{RASH}_\theta(\mathbf{F}_{n-1})\Big)^2}, \tag{5}$$

where $\theta_k = k.\lfloor\frac{180}{D}\rfloor$, $0 \leq k < D$. Geometrically, each component of this multidimensional signatures capture the motion activity in a given angular sector, whose aperture is specified by the parameter $D$. The larger $D$ is, the narrower is the associated angular sector.

| | Exact matches (%) | | | Mean deviation (frame) | | |
|---|---|---|---|---|---|---|
| | $D$=1 | $D$=2 | $D$=3 | $D$=1 | $D$=2 | $D$=3 |
| *Welcome to the Roses* | 98.3 | 99.8 | 99.9 | 4.04 | 0.00 | 0.00 |
| *The Pink Panther* | 98.1 | 99.1 | 99.5 | 9.00 | 0.04 | 0.02 |
| *Beauty and the Beast* | 98.5 | 98.7 | 99.0 | 0.10 | 0.04 | 0.03 |
| *Aladdin* | 99.3 | 99.6 | 99.4 | 2.95 | 0.25 | 0.04 |
| *The Valet* | 97.8 | 99.7 | 99.9 | 1.97 | 0.60 | 0.58 |
| *Kill Bill* | 97.4 | 97.5 | 97.6 | 10.24 | 10.23 | 9.66 |
| **Average** | **98.2** | **99.1** | **99.2** | **4.72** | **1.86** | **1.72** |

**Table 1**. Impact of the dimensionality of the fingerprint temporal signature on registration accuracy when segments of videos are deleted/inserted.

## 4. EXPERIMENTAL RESULTS

In order to assess the impact of the changes which we suggested, a number of experiments have been conducted for various use cases. The accuracy of the temporal registration process is measured using two metrics: (i) the percentage of frames that are correctly realigned according to some available ground truth, and (ii) the average frame deviation between the correct frame mapping and the ground truth.

### 4.1. Video Editing

Pirates routinely edit the content of video sequences in an attempt to evade watermarking and fingerprinting techniques. To simulate such process, we used excerpts taken from 6 feature movies in DVD format (720 × 576 @ 25 fps) and, for each excerpts, we derived two alternate versions by randomly deleting segments of variable lengths in a manner similar to [10]. In the absence of geometric or photometric distortions between the reference and candidate sequences, no noise is expected when matching corresponding elements. As a result, in this experiment, we set the rigidity parameter $K$ to 0, thereby allowing the mapping path to significantly deviate from the straight line. Eventually, we run the realignment algorithm described in Section 3 with temporal signatures of various dimensions $D$.

The introduction of penalties in the DTW framework has a drastic impact on registration performances. Even when using scalar fingerprint temporal signatures, the percentage of frames that are perfectly re-aligned rockets up to 98.2% compared to 73.5% when using the baseline DTW algorithm. At the same time, the average frame deviation drops from 70.54 frames down to 4.72 frames. On top of that, Table 1 clearly illustrates that increasing the dimensionality of the fingerprint temporal signature enables to further improve registration accuracy. In most cases, quasi-perfect realignment can be achieved with a 3-dimensional signature. The only notable exception is the sequence *Kill Bill*. Figure 1 illustrates the issue: the sequences share a small segment of 100 frames squeezed in-between two large scene insertion and deletion. As a result, DTW fails to realign the small segment and thus biases the whole statistic. This being said, it is possible to attract the estimated mapping path toward the ground truth by increasing the dimensionality of the fingerprint temporal signature.

### 4.2. Temporal Jitter and Lossy Compression

When manipulating videos, pirates naturally introduce temporal jitter and lossy compression artifacts. To simulate such distortion, we introduced a controlled 5% temporal jitter to the trailer of *Casino Royale* (8,900 frames, 1920×1080, 23.976 fps) priori to re-encoding

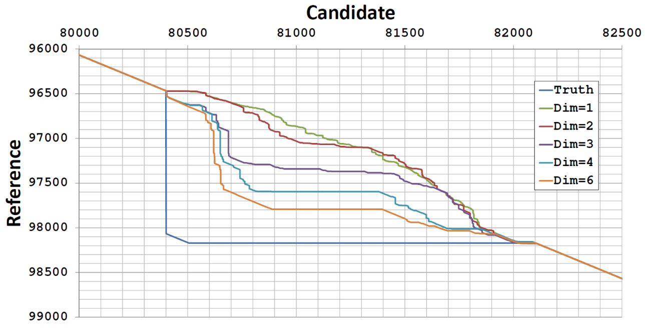

**Fig. 1**. Zoom on a critical section of the estimated registration path for the sequence *Kill Bill* where DTW fails to lock on the ground truth.

| | Exact matches (%) | | Mean frame error | |
|---|---|---|---|---|
| | Eq. (1) | Eq. (2) | Eq. (1) | Eq. (2) |
| 300 kbps @ 512×288 | 75.7 | 81.2 | 1.14 | 0.40 |
| 500 kbps @ 512×288 | 77.2 | 82.6 | 1.53 | 0.30 |
| 300 kbps @ 854×480 | 76.0 | 82.4 | 0.83 | 0.31 |
| 500 kbps @ 854×480 | 80.4 | 83.6 | 0.70 | 0.30 |
| 1000 kbps @ 854×480 | 85.3 | 85.2 | 0.50 | 0.25 |
| 2000 kbps @ 854×480 | 86.8 | 85.7 | 0.41 | 0.23 |

**Table 2**. Accuracy of the realignment process with a scalar fingerprint temporal signature in case of temporal jitter and lossy compression.

the video file using Mencoder with the lavc codec and mpeg4 filter at several bit rates and spatial resolutions in a manner similar to [11]. In practice, it implies that one frame out of 20 is either deleted or duplicated in average. We then run both the baseline and modified DTW algorithms using various dimensions $D$ for the fingerprint temporal signature. Empirical observations showed that a rigidity parameter $K$ set to 1 yielded the best registration results in this setup.

The accuracy of the realignment process using scalar fingerprint temporal signatures ($D$=1) is reported in Table 2. The general trend is that the quality of the registration improves when the strength of the attack decreases (higher bit rate and/or lighter resizing). Moreover, introducing penalties in the DTW process consistently enhances performances, if only with respect to the average frame deviation error. Figure 2 then illustrates the impact of increasing the dimensionality of the fingerprint temporal signature. Regardless of the attack strength, the percentage of perfectly realigned frames rapidly improves before flattening. For instance, a 30-D signature yields on average 6.5%additional correct frame mapping compared to a scalar signature.

### 4.3. Camcorder capture

Another popular piracy technique consists in camcording the display device. To simulate this attack, we camcorded the trailer of *Casino Royale* that was being projected onto a screen, with an acquisition geometry which introduces noticeable keystone distortion. We then cropped the recorded sequence to exclude irrelevant portions of the video, e.g. the surroundings of the projection screen. A human operator then visually inspected the sequence frame by frame to manually derive the ground truth by comparing the reference and candidate videos displayed side-by-side. Even if the 'true' mapping

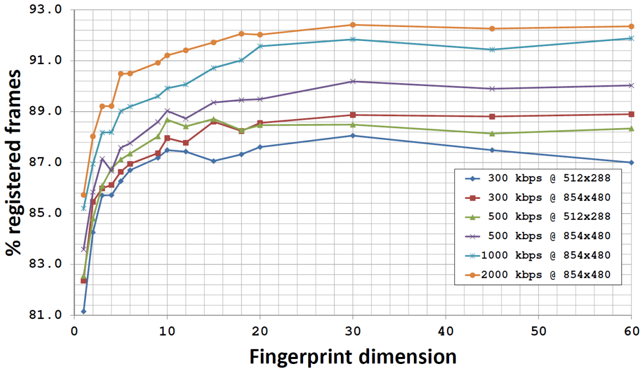

**Fig. 2**. Accuracy of the modified DTW process ($K$=1) with fingerprint temporal signatures of various dimensions in case of temporal jitter and lossy compression.

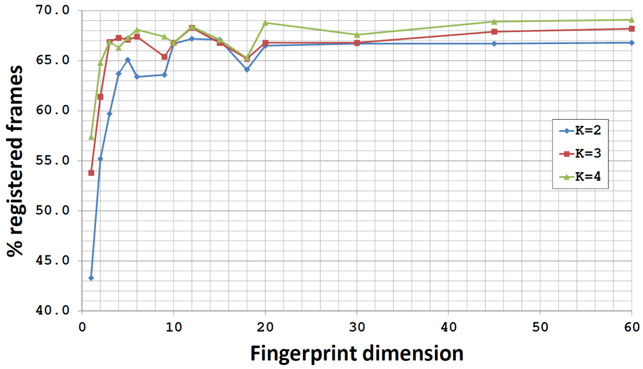

**Fig. 3**. Accuracy of the modified DTW process for different values of $K$ and $D$ in case of camcording.

path globally follows a linear trend due to the change of frame rate between the display and recoding devices (23.976 fps to 25 fps), close inspection reveals irregularities due to temporal integration of successive frames by the camcorder. An underlying assumption of DTW is that the two signals are sampled at the same frequency. If this not the case, a simple solution, which we adopted, is to upsample the sequence with the lowest frequency to match the other one before starting the DTW procedure.

We did not observe consistent results with the baseline DTW algorithm; we will therefore focus hereafter on the modified DTW framework that incorporates insertion/deletion penalties. Due to the larger amount of noise due to photometric and geometric distortions, we investigated several values for the rigidity parameter $K$ and various dimensions $D$ for the fingerprint temporal signature. Registration accuracy results are reported in Figure 3. The signature dimension appears to play a bigger role in performances than the rigidity parameter. The general trend remains the same as before but the percentage of perfectly realigned frames hes dropped below 70%. Still, most pairings are very close to the ground truth. For instance, for $K$=3 and $D$=12, 'only' 68% of the frames are perfectly registered, but 30% stand ±1 frame apart, 1% stand ±2 frames apart. The average registration error is in this case 0.3 frame.

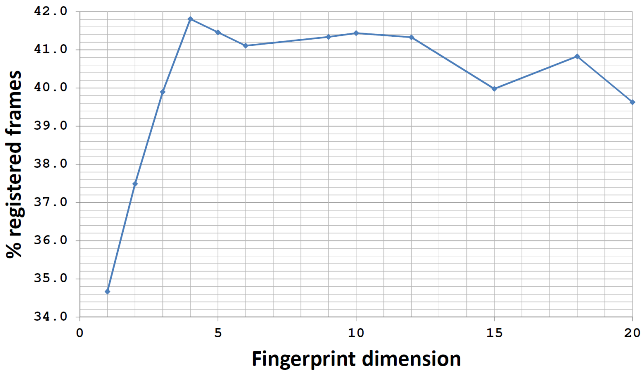

**Fig. 4**. Accuracy of the modified DTW process ($K$=0) in case of screencast capture.

### 4.4. Screencast Capture

A raising piracy threat is the use of screencasting software that records whatever is displayed on a PC screen. To analyze the effects of such software, we screencasted the trailer of *Cars 2* (3,000 frames, 642 × 274, 25 fps) and manually recovered the ground truth as previously. The analysis of the ground truth revealed that the temporal distortion model is highly complex and unpredictable. The frame rate varies a lot (piece-wise constant between 37 and 42 fps) due to CPU saturation and a number of frames are randomly repeated. In an attempt to cope with this, we upsampled the reference fingerprint temporal signature from 25 to 36 fps. On the other hand, geometric and photometric distortions are minimal which allows us to relax the rigidity parameter $K$ to 0 and thereby provide maximal elasticity to the mapping path. Empirical validation confirmed that this setting yields the best possible registration results in this case.

As for the camcorder capture, we did not obtain consistent results with the baseline DTW algorithm. The registration performances for this content using the modified DTW algorithm are depicted in Figure 4. We still observe the same profile: the accuracy of the realignment process increases sharply when increasing $D$ at the beginning, then flattens before slowly decreasing. This slow decreasing trend at the end illustrate a well-known trade-off in fingerprinting between robustness and diversity. The larger $D$ is, the more there is information in the temporal signature, the more this information is unstable. In this configuration, at best ($D$=4) 42% of the frames are perfectly realigned, 27% stand ±1 frame apart, 10% ±2 frames apart, and the average registration error is 2.0 frames. While such performances are noticeably poorer than in previous experiments, they are still better than what the one-dimensional fingerprint temporal signature achieves (35% perfect match, 2.7 frame average error).

## 5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a number of modifications to improve the registration performances of the baseline DTW algorithm. More specifically, we suggested to introduce insertion/deletion penalties in the DTW framework and to increase the dimension of the fingerprint temporal signature fed to the DTW algorithm. These modifications proved to noticeably increase the accuracy of the realignment process for a wide range of attacks, from simple synthetic editing to real-life re-capture.

The DTW algorithm remains however somewhat hampered by the change of frame rate. The fingerprint interpolation technique that we used is a simplistic, yet unsatisfactory, bypass and alternate strategies are likely to yield improved registration results. The optimal $(K, D)$ values seem to be heavily dependent on the attack scenario. Further research is needed to find a means to automatically adjust these parameters without having to test all possibilities.

Registration is still a key issue in watermarking today, in particular if the detector requires frame-accurate alignment [12, 13]. A shortcoming of DTW realignment is that the registration process is performed independently of the watermarking signal. In other words, all frames are realigned in the same fashion regardless of the amount of watermark information carried by each frames. A recent proposal investigated a mechanism to bind the two in an attempt to guarantee higher registration accuracy for frames carrying more watermark signal [11].

## 6. REFERENCES

[1] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, Morgan Kaufmann Publishers, 2nd edition, 2008.

[2] P. Schaber, S. Kopf, W. Effelsberg, and N. J. Thorwirth, "Semi-automatic registration of videos for improved watermark detection," in *Proc. ACM SIGMM Multimedia Systems Conf.*, Feb. 2010, pp. 23–34.

[3] D. Delannay, C. de Roover, and B. Macq, "Temporal alignment of video sequences for watermarking systems," in *Security and Watermarking of Multimedia Contents V*, Jan. 2003, vol. 5020 of *Proc. of SPIE*, pp. 481–492.

[4] O. Harmanci, M. Kucukgoz, and M. K. Mihcak, "Temporal synchronization of watermarked video using image hashing," in *Security, Steganography, and Watermarking of Multimedia Contents VII*, Jan. 2005, vol. 5681 of *Proc. of SPIE*.

[5] H. Cheng, "Temporal registration of video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, Apr. 2003, vol. 3, pp. 489–492.

[6] B. Chupeau, L. Oisel, and P. Jouet, "Temporal video registration for watermark detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, May 2006, vol. 2, pp. 157–160.

[7] S. Baudry, B. Chupeau, and F. Lefèbvre, "A framework for video forensics based on local and temporal fingerprints," in *Proc. IEEE Int. Conf. Image Proc.*, Nov. 2009, pp. 2889–2892.

[8] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2001.

[9] C. De Roover, C. De Vleeschouwer, F. Lefèbvre, and B. Macq, "Robust video hashing based on radial projections of key krames," *IEEE Trans. on Sig. Proc.*, vol. 53, no. 10, pp. 4020–4037, Oct. 2005.

[10] F. Thudor, I. Autier, B. Chupeau, F. Lefèbvre, and L. Oisel, "Automatic chaptering of VoD content based on DVD content," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Jun. 2012.

[11] S. Baudry, "Frame-accurate temporal registration for non-blind watermarking," in *Proc. ACM Multimedia and Security Workshop*, Sept. 2012, pp. 19–26.

[12] D. Zou and J. A. Bloom, "H.264/AVC substitution watermarking: a CAVLC example," in *Media Forensics and Security XI*, Jan. 2009, vol. 7254 of *Proc. of SPIE*.

[13] D. Zou and J. A. Bloom, "H.264 stream replacement watermarking with CABAC encoding," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Jul. 2010, pp. 117–121.