Video Alignment for Phylogenetic Analysis

Silvia Lameri, Paolo Bestagini, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy

***Abstract*—The possibility of studying multiple objects at once for forensic analysis has paved the way to the development of multimedia phylogeny algorithms. Concerning video phylogeny, a fundamental step at the base of many applications is multiple video alignment. This is, given a pool of near-duplicate video sequences partially overlapping in the temporal domain, find the relative time delay between all of them. As phylogeny methods typically takes into account huge quantities of data, the used alignment algorithms must be computationally efficient. In this paper, we propose a solution for multiple video alignment based on the minimisation of a least-square cost function. The proposed solution can be computed in closed form with reduced computational complexity. Moreover, we propose two possible solutions for refining the estimated alignment based on the removal of outlier measurements.**

I. Introduction

Thanks to the rapid diffusion of multimedia sharing platforms, the amount of video sequences distributed online is constantly increasing. Moreover, as editing software tools are at everyone's hand, video content modifications can be easily applied by everyone. This possibility has determined the spread of near-duplicate (ND) video objects, i.e., different edited versions of the same original content. In order to regulate this huge amount of information, the forensics community has developed a wide set of algorithms and tools for video analysis and authentication [1].

Many of these methods are based on the analysis of single video objects to study their past history. As an example, in [2], [3], [4], the authors study video coding history. Video tampering detection and localization are tackled in [5], [6], [7]. Moreover, the problem of video recapture is faced in [8], [9], [10] according to different hypothesis.

However, with the diffusion of near-duplicate contents, the forensics community has started developing methods that synergically exploit information coming from all of them to perform deeper analysis. For example, in [11], [12], the authors focus on reconstructing the video phylogeny tree. This is an acyclic directed graph representing the ancestral relationships between all possible video pairs in the analysed pool. This enables to detect the user that originally posted some illicit material, or to solve ownership issues. Additionally, in [13], a system for reconstructing the original sequence used to generate a set of ND videos was developed. This enables to reconstruct some video content no more available online in its totality, as well as to shed an interesting insight on the way content has been distributed and re-used.

A key step at the base of the aforementioned video phylogeny algorithms is the temporal alignment of ND videos. As a matter of fact, in [12], videos must be pair-wise aligned to be frame-wise compared. A wrong alignment is proved to bring to very inaccurate video phylogeny tree reconstructions. Moreover, in [13], the concept of video alignment is brought to a different level. Indeed, alignment is not required for video pairs only, but a global alignment consistent for all the videos in the analysis pool is requested.

In addition to video phylogeny, video temporal alignment has been deeply studied also for many other applications, such as video retrieval [14], event retrieval [15], gesture recognition [16], security [17] and so on. However, video alignment algorithms developed in these fields, as [18], [19], [20], are typically computationally expensive. Indeed, the alignment procedure often comes as part of the solution of a more complex problem (e.g., processing of videos from very different view-points, alignment of videos at different frame rate, video matching in huge catalogues, etc.). Therefore, even though these techniques are very accurate, they are not suitable for video phylogeny purposes.

In this paper, we propose an algorithm for global temporal alignment of a pool of near-duplicate videos, specifically tailored to video phylogeny problems. In particular, our proposed method first estimates the misalignment between each video pair in the NDs pool. Then, it exploits all pair-wise alignment information to estimate the global alignment consistent for all videos in closed-form. The ability of solving the problem in closed-form enables the algorithm to be computationally efficient, which is of paramount importance for phylogeny applications. Finally, two procedures for detecting possible outlier measurements in pair-wise alignment are proposed. These guarantee an accurate global video alignment estimation even when some video pairs are not correctly aligned.

An experimental campaign on 2100 video sequences and additional simulated data has been conducted to validate the developed technique. More specifically, we compared the performances of the proposed solution with other alignment algorithms whose computational complexity is comparable to ours.

The rest of the paper is structured as follows. Section II reports the formal definition of the problem. Section III describes each step of the proposed algorithm, from pair-wise alignment to outlier removal. Section IV shows the results achieved in our experimental campaign. Finally, Section V concludes the paper.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

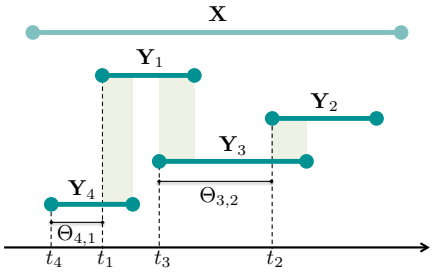

Fig. 1: Problem setup. A set of ND videos $\mathbf{Y}_i$ are generated from $\mathbf{X}$. As some of them overlap in time, temporal alignment is achievable even though $\mathbf{X}$ is not available.

II. Problem Formulation

Let $\mathbf{X}$ be an original video sequence relative to a particular event. We define a near-duplicate (ND) video of $\mathbf{X}$ a time-clipped version of $\mathbf{X}$, to which content preserving transformations (e.g., cropping, compression, resize, brightness enhancement, logo addition and so on) have been applied. With reference to Figure 1, let $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ denote a set of K near-duplicate videos generated from $\mathbf{X}$. We can associate to each video $\mathbf{Y}_i$ a starting time t_i on a common timeline. The misalignment of video $\mathbf{Y}_j$ with respect to video $\mathbf{Y}_i$ is defined as

$$\Theta_{i,j} = t_j - t_i. \tag{1}$$

In this work we propose a method that aims to produce a global temporal alignment of a collection of near-duplicate videos $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$, starting from their analysis and without any prior knowledge on the original content $\mathbf{X}$. Aligning a set of sequences consists in finding the starting times $t_1, \ldots, t_K$, up to an additive constant. To this purpose, we assume that at least a minimum number of ND video pairs $(\mathbf{Y}_i, \mathbf{Y}_j)$ share a temporal overlap. For example in Figure 1 the sequence $\mathbf{Y}_1$ does not share any temporal overlap with $\mathbf{Y}_2$, but the sequence $\mathbf{Y}_3$ partially overlaps with both $\mathbf{Y}_1$ and $\mathbf{Y}_2$. It is therefore possible to align the three of them, even though not all of them have some frames in common. Anyway, if additional overlaps occur (e.g., between $\mathbf{Y}_1$ and $\mathbf{Y}_2$), redundant information is available. It is therefore possible to exploit it for even more robust global alignment estimation.

As a matter of fact, given a pool of overlapping ND sequences, it is possible to detect pairs of videos $(\mathbf{Y}_i, \mathbf{Y}_j)$ that share common frames (as in [13]) and estimate their pair-wise mutual delay Θ. Then, t values can be reconstructed by merging Θ estimates. The more the available Θ values, the better the t estimate. As an example, with reference to Figure 1, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are only linked through $\mathbf{Y}_3$. Therefore, if $\Theta_{3,2}$ is wrongly estimated, the global delay between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ cannot be correctly computed. Conversely, if an additional overlap between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ would be present, $\Theta_{1,2}$ could be computed. By exploiting both $\Theta_{3,2}$ and $\Theta_{1,2}$ values, it is possible to reduce the effect of the wrong $\Theta_{3,2}$ estimate in computing the global delay between $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

In the next section, we present all the steps of the proposed alignment algorithm that exploits data redundancy being computationally efficient, thus suitable for phylogeny applications.

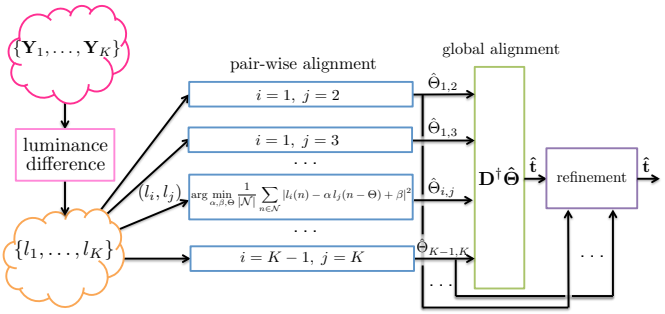

Fig. 2: Diagram of the proposed system.

III. Alignment Algorithm

To deal with the aforementioned problem of video temporal alignment we propose an algorithm divided in four basic steps as depicted in Figure 2: (i) each video is represented over time as a monodimensional signal; (ii) pair-wise alignments $\Theta_{i,j}$ are estimated by comparing video monodimensional representations; (iii) global alignment t_k of each sequence is estimated combining all the available pair-wise $\Theta_{i,j}$ alignments; (iv) a final refinement step is applied to remove the effect of possible outlier measurements $\Theta_{i,j}$. In the following, a detailed description of each step is given.

A. Monodimensional descriptors

In order to compare pairs of sequences within a pool of K near-duplicate videos $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ in a computationally efficient way, we resort to a monodimensional description of a video over time, as suggested in [21]. More specifically, given an arbitrary video sequence $\mathbf{Y}_i$, we compute the difference between the average luminance of adjacent frames as

$$l_i(n) = \text{avgluma}(\mathbf{Y}_i(n)) - \text{avgluma}(\mathbf{Y}_i(n-1)), \tag{2}$$

where $\text{avgluma}(\cdot)$ extracts the average of the luminance component of a frame, and $\mathbf{Y}_i(n)$ is the n-th frame of sequence $\mathbf{Y}_i$.

B. Pair-wise alignment

Given two partially overlapped near-duplicate videos $\mathbf{Y}_i$ and $\mathbf{Y}_j$, we seek the temporal shift $\Theta_{i,j}$ between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ by comparing their monodimensional descriptors l_i and l_j, built as in (2). In particular we find the pair-wise temporal alignment between $\mathbf{Y}_i$ and $\mathbf{Y}_j$ by minimising a cost function. More formally, we compute

$$\hat{\Theta}_{i,j} = \arg\min_{\alpha,\beta,\Theta} \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} |l_i(n) - \alpha \cdot l_j(n - \Theta) + \beta|^2, \tag{3}$$

where $\mathcal{N}$ is the set of time lags that ensure overlap between l_i and l_j delayed by Θ, $|\mathcal{N}|$ denotes its cardinality, α is a scaling factor and β accounts for possible luminance shifts. The minimisation is carried out on the tuple (α, β, Θ). It is worth noticing that Θ can only assume a finite set of integer values. Indeed, time-shifts are measured in frames (integer values) and the two shifted sequences must overlap.

The proposed solution can be interpreted as an enhanced version of the one presented in [13]. As a matter of fact, in [13] the temporal alignment is estimated by looking at the position of the highest peak of the cross-correlation between l_i and l_j, which does not take into account neither α, not β.

C. Global alignment

At this point in the algorithm, given the set $\{\hat{\Theta}_{i,j}\}_{i,j=1,\dots,K,\ i\neq j}$, of pair-wise alignment measurements, we estimate the starting times $t_1,\dots,t_K$ of $\mathbf{Y}_1,\dots,\mathbf{Y}_K$ on a common timeline. To this purpose, we exploit the relationship between Θ and t values expressed by (1), which can be expressed in matricial form as

$$\begin{bmatrix} \Theta_{1,2} \\ \Theta_{2,3} \\ \vdots \\ \Theta_{K-1,K} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_{K-1} \\ t_K \end{bmatrix}, \quad (4)$$

$$\text{or} \qquad \boldsymbol{\Theta} = \mathbf{D}\mathbf{t}.$$

More specifically, as we do not know the true Θ values, but only their estimates $\hat{\Theta}$, our model becomes

$$\hat{\boldsymbol{\Theta}} = \mathbf{D}\mathbf{t} + \mathbf{e}, \quad (5)$$

where $\mathbf{e}$ denotes a noise term accounting for Θ measurement errors. Depending on hypothesis on the error $\mathbf{e}$, there are several ways to estimate $\mathbf{t}$ from the knowledge of $\hat{\boldsymbol{\Theta}}$ and $\mathbf{D}$. As we seek for a fast solution, we formulate our estimation problem in terms of least squares (LS). This means that the estimated $\mathbf{t}$ is obtained solving

$$\hat{\mathbf{t}} = \arg\min_{\mathbf{t}\in\mathbb{R}^K} \|\mathbf{D}\mathbf{t} - \hat{\boldsymbol{\Theta}}\|_2, \quad (6)$$

whose solution is known to be in closed-form.

As the matrix $\mathbf{D}$ is rank deficient with $rank(\mathbf{D}) = K-1$ [22], the LS problem is solved exploiting singular value decomposition (SVD). To this purpose, let $\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the SVD of $\mathbf{D}$. The LS solution of (6) is

$$\hat{\mathbf{t}} = \mathbf{D}^\dagger\hat{\boldsymbol{\Theta}} = \mathbf{V}_1\boldsymbol{\Sigma}_1^{-1}\mathbf{U}_1^T\hat{\boldsymbol{\Theta}}, \quad (7)$$

where $\mathbf{D}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{D}$, $\boldsymbol{\Sigma}_1$ is the square submatrix of $\boldsymbol{\Sigma}$ containing the $K-1$ positive singular values, and $\mathbf{U}_1$ and $\mathbf{V}_1$ are the first $K-1$ columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.

Because of the rank-deficiency of $\mathbf{D}$, the found solution $\hat{\mathbf{t}}$ is non-unique. It is indeed quite evident that a constant can be added to each of the elements of the vector $\hat{\mathbf{t}}$ of starting times without changing the vector of pair-wise overlaps $\hat{\boldsymbol{\Theta}}$. As a matter of fact, all possible solutions up to an additive constant are equivalent for our problem, as stated in Section II. As a convention, we decided to set the first value of $\hat{\mathbf{t}}$ to zero, by subtracting $\hat{t}_1$ to all the elements of $\hat{\mathbf{t}}$. Also notice that, as values of $\hat{\mathbf{t}}$ must be integer numbers (delay is measured in frames), we apply a rounding operation to $\hat{\mathbf{t}}$.

As a final remark, also notice that not all the sequences in the near-duplicates set may overlap (as it was for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ in Figure 1). This means that only a subset of all the possible $\hat{\Theta}_{i,j}$ may be available. Nonetheless, the $\hat{\mathbf{t}}$ estimation procedure remains exactly the same, just removing from $\mathbf{D}$ all the rows corresponding to missing $\hat{\Theta}_{i,j}$ values.

D. Outlier removal

The LS procedure described above allows to estimate $\hat{\mathbf{t}}$ from $\hat{\boldsymbol{\Theta}}$. However, $\hat{\boldsymbol{\Theta}}$ may contain some outlier measurements $\hat{\Theta}_{i,j}$ due to incorrect pair-wise alignment estimations. In this case, it would be preferable to remove outliers before $\hat{\mathbf{t}}$ estimation.

To this purpose, we propose two possible outlier removal procedures. They both start with the computation of the LS residual of each measurement as

$$\mathbf{r} = \hat{\boldsymbol{\Theta}} - \mathbf{D}\hat{\mathbf{t}}, \quad (8)$$

where each element $r_{i,j}$ of $\mathbf{r}$ basically measures how well each $\hat{\Theta}_{i,j}$ fits the found LS solution. As a matter of fact, in case of noiseless measurements, $\mathbf{r} = 0$. On the contrary, outlier $\hat{\Theta}_{i,j}$ exhibit high $r_{i,j}$ values.

LS-driven minimum-spanning tree. The first outlier removal procedure consists in making use of $r_{i,j}$ as confidence values associated to each $\hat{\Theta}_{i,j}$. Formally, we build a graph where each node represents a video sequence $\mathbf{Y}_i$, and each edge from $\mathbf{Y}_i$ to $\mathbf{Y}_j$ has weight $r_{i,j}$. If two sequences do not overlap, no direct edge links them. We then run minimum-spanning tree algorithm on this graph to find the path at minimum cost that links each $\mathbf{Y}_i$ to each $\mathbf{Y}_j$. The $\hat{\mathbf{t}}$ values can than be computed summing $\hat{\Theta}_{i,j}$ values on the path linking $\mathbf{Y}_i$ to $\mathbf{Y}_j$.

Robust LS. Another possible procedure to detect outliers consists in comparing the standard deviation σ_r of $\mathbf{r}$ elements with a threshold Γ_r. Formally, we compute

$$\sigma_r = \sqrt{\frac{1}{|\mathbf{r}|}\sum_{r_{i,j}\in\mathbf{r}}(r_{i,j}-\mu_r)^2}, \quad (9)$$

where μ_r is the average value of elements in $\mathbf{r}$ and $|\mathbf{r}|$ is the number of elements in $\mathbf{r}$. If $\sigma_r \geq \Gamma_r$, we apply the refinement procedure. We select the highest $r_{i,j}$ element, and remove the associated $\hat{\Theta}_{i,j}$ from $\hat{\boldsymbol{\Theta}}$. We then estimate $\hat{\mathbf{t}}$ again using the refined measurements, and iterate the overall refinement procedure until $\sigma_r < \Gamma_r$, or the minimum number of $\hat{\Theta}_{i,j}$ to solve the LS problem is reached. The final estimation of $\hat{\mathbf{t}}$ is kept as solution.

IV. Results

In order to validate the proposed alignment algorithm, we built a dataset of 2100 near-duplicate videos. These have been generated starting from 7 well-known original video sequences, namely: *city*, *crew*, *foreman*, *mobile*, *mother*, *paris*, and *sign Irene* . These videos at CIF resolution (i.e., 352×288 pixels) range from 300 to 540 frames each. From each original sequence we created 30 different realisations of misaligned near-duplicates. Each realisation is composed by 10 misaligned ND videos generated by randomly applying transformations listed in Table I to the original sequence. ND videos

TABLE I: Parameters of transformations.

Transformation	Range
Gaussian blurring	std. dev. = [2,8]
Logo addition	area = [3%, 7%]
Global scaling	[90%,110%]
Cropping	[0%,5%]
Contrast adjustment	[-10%,10%]
Brightness adjustment	[-10%,10%]

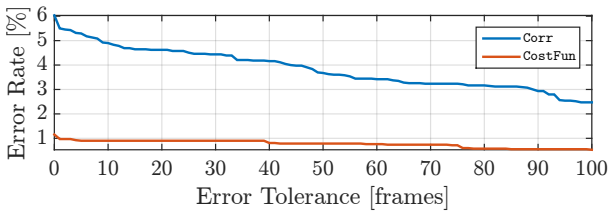

Fig. 3: Alignment error on video pairs using our method (`CostFun`) and the baseline (`Corr`).

in each realisation have been temporally trimmed in order to obtain a random misalignment among them. Algorithms have been evaluated both pair-wise and globally aligning the 10 near-duplicate videos in each of the $30 \times 7 = 210$ realisations.

It is worth noting that temporal trimming has been applied paying attention to the number of sequences sharing overlapped frames. As a matter of fact, in the easiest scenario, all the 10 videos in a realisation share some frames, thus providing 45 measurements $\hat{\Theta}_{i,j}$ (i.e., one for each video pair). However, we considered the more difficult case in which only a limited set of sequences share some frames. In particular, we considered realisations in which up to 32 video pairs (over 45) do not share any frame. Temporally overlapping sequences share at minimum 50 frames (i.e., less then 2 seconds).

A. Pair-wise alignment evaluation

In order to evaluate the accuracy of the proposed method in estimating the pair-wise misalignment $\hat{\Theta}_{i,j}$, we computed the monodimensional descriptor of each video as in (2) and applied (3) to all the overlapping ND videos. For rapidly minimising (3), we normalized each $l(n)$ to have zero mean and unitary standard deviation. This allowed us to constrain $\alpha \in [0.8, 2.5]$ and $\beta \in [-1, 1]$, thus shrinking the search space. For comparison, we also estimated $\hat{\Theta}_{i,j}$ using as baseline the cross-correlation-based method reported in [13].

Figure 3 shows the percentage of pair-wise alignment errors using the proposed method based on cost-function minimisation (`CostFun`), and using the baseline method (`Corr`). More specifically, we considered as errors all the measured $\hat{\Theta}_{i,j}$ that differ from the correct value $\Theta_{i,j}$ for a fixed number of tolerance frames. It is possible to see that, if the tolerance is set to 0 (i.e., we only consider as correct estimations $\hat{\Theta}_{i,j} = \Theta_{i,j}$), our method wrongly estimates the misalignment on approximately 1% of the sequences, whereas the baseline wrongly estimates the misalignment in the 6% of the cases. Even if we consider a tolerance of 40 frames (i.e., we consider as correct estimations all the $\hat{\Theta}_{i,j}$ that verify $|\hat{\Theta}_{i,j} - \Theta_{i,j}| < 40$), the baseline method has an error percentage higher than 4%, whereas our method wrongly estimates the misalignment in only approximately the 0.5% of the cases.

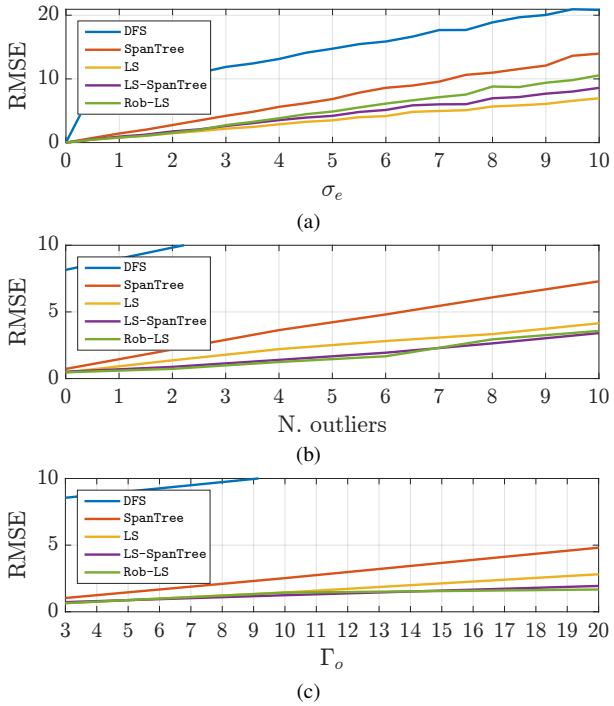

Fig. 4: RMSE in presence of additive noise and outliers on $\hat{\Theta}$: (a) no outliers; (b) $\sigma_e = 0.5$ and $\Gamma_o = 20$; (c) $\sigma_e = 0.5$ and 6 outliers.

B. Global alignment evaluation

In order to evaluate the global alignment methodology, we compared our methods with other possible solutions. To this purpose we selected the baseline depth-first search (`DFS`) used in [13]. Moreover, we considered a baseline minimum-spanning tree solution (`Span`). This is motivated by the fact that in [19] the authors make use of minimum-spanning tree algorithms to estimate t values from $\hat{\Theta}$ ones for event retrieval. However, as the complete algorithm would be too computationally expensive, we decided to simplify it. More specifically, we opt for a faster solution whose weights are based on the number of overlapping frames between sequence pairs. The rationale is that, the more the overlapping frames, the more robust the estimate of $\hat{\Theta}_{i,j}$.

Regarding our proposed methods, we denoted as: (i) `LS` the least-squares inversion without outlier-removal refinement; (ii) `LS-Span` the refinement procedure based on minimum-spanning tree driven by LS residuals; (iii) `Rob-LS` the refinement procedure that discards measurements according to the LS residuals standard deviation.

First, we conducted a simulative campaign to assess the behaviour of the algorithms under different perturbation conditions. To this purpose, we corrupted the noiseless ground-truth values $\Theta_{i,j}$ from our video dataset with additive i.i.d zero-mean Gaussian noise with standard deviation σ_e and the presence of outliers. Outliers were modelled as noise samples belonging to a Guassian distribution with standard deviation

TABLE II: Global alignment error on ND video sequences. Best results for each evaluation metric are reported in bold.

	Baseline		Proposed		
	DFS	SpanTree	LS	LS-Span	Rob-LS
RMSE	8.42	4.27	3.84	2.98	**2.79**
μ_{err}	6.04	2.58	2.75	2.02	**1.98**
σ_{err}	5.89	3.38	2.55	2.06	**1.85**

$\Gamma_o \cdot \sigma_e$ (i.e., Γ_o times greater than the additive noise one). As evaluation metrics we used the error bias, error standard deviation and the root-mean-square error (RMSE) defined as

$$\begin{aligned} \mu_{\text{err}} &= \mathrm{E}[\hat{t} - t], \\ \sigma_{\text{err}} &= \mathrm{E}[(\hat{t} - t - \mu_{\text{err}})^2], \\ \text{RMSE} &= \sqrt{\mathrm{E}[(\hat{t} - t)^2]}, \end{aligned} \tag{10}$$

where $\mathrm{E}[\cdot]$ expresses the average over all the estimations.

Figure 4(a) shows the RMSE for different additive noise standard deviations σ_e when no outliers are present. From these results it is evident that the two baseline procedures perform worse than the proposed ones. Moreover, as no outliers have been added, the best solution is to use LS which exploits all the measurements at best.

Figure 4(b) shows the RMSE for different number of outliers, when the noise standard deviation has been fixed to $\sigma_e = 0.5$ and $\Gamma_o = 20$. In this situation, it is clear that LS starts suffering from the presence of the outliers. On the other hand, the proposed refinement strategies LS-Span and Rob-LS outperforms all the other methods.

Finally, Figure 4(c) shows the RMSE when changing the "strength" of the outliers Γ_o, while keeping fixed $\sigma_e = 0.5$ and the number of outliers to 6. Also these results confirm that the proposed refinement procedures outperforms the other strategies, guaranteeing an RMSE less than 3 frames.

In order to validate the whole system, we evaluated the global alignment strategies using $\hat{\Theta}$ values extracted using (3) on the built near-duplicate video dataset. Results are shown in Table II. It is evident that baseline solutions have worse performances. Among the proposed ones, as expected, the refinement methods outperforms LS. More specifically, Rob-LS shows better performances than all the other solutions.

V. Conclusions

In this paper we faced the problem of multiple near-duplicate video alignment. To this purpose, we developed an algorithm that considers a pool of near-duplicate videos, estimates their pair-wise temporal misalignment, and finally globally aligns all of them. The proposed procedure also embeds the possibility of detecting outliers and discard them to reliably estimate the global alignment in two different ways. Reduced computational complexity enables this algorithm to be used for video phylogeny analysis.

The conducted experimental campaign proved the validity of the proposed methodology in different conditions, considering both simulated data and real video sequences. Future works will be devoted to the development of statistically motivated methods for outlier detection, and to testing in real-world scenarios.

References

[1] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e2, 2012.

[2] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Multiple compression detection for video sequences," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2012.

[3] D. Vazquez-Padin, M. Fontani, T. Bianchi, P. Comesana, A. Piva, and M. Barni, "Detection of video double encoding with GOP size estimation," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012.

[4] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Codec and gop identification in double compressed videos," *IEEE Transactions on Image Processing (TIP)*, vol. 25, pp. 2298–2310, 2016.

[5] D. Labartino, T. Bianchi, A. De Rosa, M. Fontani, D. Vazquez-Padin, A. Piva, and M. Barni, "Localization of forgeries in MPEG-2 video through GOP size and DQ analysis," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, s 2013.

[6] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Local tampering detection in video sequences," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[7] L. D'Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, "Video forgery detection and localization based on 3D patchmatch," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2015.

[8] W. Wang and H. Farid, "Detecting re-projected video," in *Information Hiding*. Springer Berlin Heidelberg, 2008.

[9] M. Visentini-Scarzanella and P. L. Dragotti, "Video jitter analysis for automatic bootleg detection," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2012.

[10] P. Bestagini, M. Visentini-Scarzanella, M. Tagliasacchi, P. Dragotti, and S. Tubaro, "Video recapture detection based on ghosting artifact analysis," in *2013 IEEE International Conference on Image Processing (ICIP)*, 2013.

[11] Z. Dias, A. Rocha, and S. Goldenstein, "Video phylogeny: Recovering near-duplicate video relationships," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2011.

[12] F. Costa, S. Lameri, P. Bestagini, Z. Dias, A. Rocha, M. Tagliasacchi, and S. Tubaro, "Phylogeny reconstruction for misaligned and compressed video sequences," in *IEEE International Conference on Image Processing (ICIP)*, 2015.

[13] S. Lameri, P. Bestagini, A. Melloni, S. Milani, A. Rocha, M. Tagliasacchi, and S. Tubaro, "Who is my parent? Reconstructing video sequences from partially matching shots," in *IEEE International Conference on Image Processing (ICIP)*, 2014.

[14] M. Esmaeili, M. Fatourechi, and R. Ward, "A robust and fast video copy detection system using content-based fingerprinting," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 6, pp. 213–226, 2011.

[15] J. Revaud, M. Douze, C. Schmid, and H. Jegou, "Event retrieval in large video collections with circulant temporal encoding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[17] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: establishing a common coordinate frame," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, pp. 758–767, 2000.

[18] A. Melloni, S. Lameri, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Near-duplicate detection and alignment for multi-view videos," in *IEEE International Conference on Image Processing (ICIP)*, 2015.

[19] M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid, "Circulant temporal encoding for video retrieval and temporal alignment," *International Journal of Computer Vision (IJCV)*, pp. 1–16, 2015.

[20] F. Padua, R. Carceroni, G. Santos, and K. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, pp. 304–320, 2010.

[21] Q. Xie, Z. Huang, H. T. Shen, X. Zhou, and C. Pang, "Efficient and continuous near-duplicate video detection," in *International Asia-Pacific Web Conference (APWEB)*, 2010.

[22] R. Schmidt, "Least squares range difference location," *IEEE Transactions on Aerospace and Electronic Systems (TAES)*, vol. 32, pp. 234–242, 1996.