Joint Spatio-Temporal Alignment of Sequences

Ferran Diego, Joan Serrat, and Antonio M. López

***Abstract*—Video alignment is important in different areas of computer vision such as wide baseline matching, action recognition, change detection, video copy detection and frame dropping prevention. Current video alignment methods usually deal with a relatively simple case of fixed or rigidly attached cameras or simultaneous acquisition. Therefore, in this paper we propose a joint video alignment for bringing two video sequences into a spatio-temporal alignment. Specifically, the novelty of the paper is to formulate the video alignment to fold the spatial and temporal alignment into a single alignment framework. This simultaneously satisfies a frame-correspondence *and* frame-alignment similarity; exploiting the knowledge among neighbor frames by a standard pairwise Markov random field (MRF). This new formulation is able to handle the alignment of sequences recorded at different times by independent moving cameras that follows a similar trajectory, and also generalizes the particular cases that of fixed geometric transformation and/or linear temporal mapping. We conduct experiments on different scenarios such as sequences recorded simultaneously or by moving cameras to validate the robustness of the proposed approach. The proposed method provides the highest video alignment accuracy compared to the state-of-the-art methods on sequences recorded from vehicles driving along the same track at different times.**

***Index Terms*—Direct-based, feature-based, image registration, Markov random fields, synchronization, video alignment, video retrieval.**

I. Introduction

VIDEO alignment relates video sequences in temporal and spatial dimensions (Fig. 1). Hence, the video is decomposed in temporal and spatial alignment. The former, or synchronization, estimates a discrete temporal correspondence that associates the frames of a sequence to the frames of another maximizing a similar image content. The latter, usually called registration, takes a pair of corresponding frames and maps the image coordinates of one frame to another based on some similarity measure and a spatial deformation. Video alignment has received considerable attention for many years since the seminal paper by Stein [1] and plays an important role in applications as wide baseline matching [2], high dynamic range video and video mating [3], action recognition [4], change detection [5], detecting abandoned objects [6], and frame dropping prevention [7].

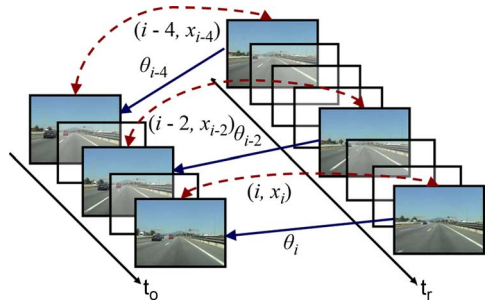


Fig. 1. Example of the most general video alignment problem: the alignment of video sequences recorded from independent moving cameras at different times. An estimation of a non-parametric temporal correspondence of frame pairs (i, x_i) (temporal alignment), where i and x_i are the i^{th} and x_i^{th} frame in the observed and reference sequence respectively; and one per frame unknown and non-fixed geometric transformation parameter denoted by $\boldsymbol{\theta}_i$ (spatial alignment) is required.

Most video alignment methods deal with sequences recorded simultaneously by cameras fixed or rigidly attached to each other that involve a fixed parametric model *and* a fixed geometric transformation across the sequences [2], [8]. The problem can be formulated as some minimization over a small number of parameters: time correspondence (e.g., offset and frame rate ratio) and geometric transformation. However, the video alignment problem becomes more challenging (and general) when this constraint for spatial and temporal mapping is broken because the sequences have been acquired at different times by independently moving cameras. This challenge has been tackled in the literature dividing the problem in two steps: a video synchronization and an image registration for each pair of corresponding frames. However, the errors on the first stage are propagated to the second without exploiting the complementarity between both steps.

Therefore, in this paper, we propose a joint spatio-temporal video alignment to handle the challenging problem of aligning sequences. Specifically, the video alignment is formulated as a unique inference problem on a huge number of parameters, a non-parametric temporal correspondence and non-fixed geometric transformation, instead of independently tackling either temporal or spatial alignment. This simultaneously satisfies a frame-correspondence, or synchronization, and a frame-alignment, or image registration, along the whole sequence. Hence, this joint similarity accurately discriminates among successive frames that share a huge similarity of content; thus reducing

Manuscript received June 23, 2011; revised June 13, 2012 and September 21, 2012; accepted November 08, 2012. Date of publication February 14, 2013; date of current version September 13, 2013. This work was supported by the Spanish Ministry of Education and Science under project TRA2011-29454-C03-01, and research program Consolider Ingenio 2010: MIPRCV (CSD200700018), the FPU MEC grant AP2007-01558 and Cellular Networks at the University of Heidelberg. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Christophe De Vleeschouwer.

F. Diego is with the Heidelberg Collaboratory for Image Processing (HCI), University of Heidelberg, Heidelberg, Germany (e-mail: ferran.diego@iwr.uni-heidelberg.de; fdiego@cvc.uab.es).

J. Serrat and A. M. López are with the Computer Vision Center and Computer Science Department, Edifici O, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Valls, Spain (e-mail: joans@cvc.uab.es; antonio@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2247390

spatio- and temporal-misalignments. This reduction is reinforced by exploiting the similarities between neighbor frames. The way we do it is by integrating the estimation of the spatio-temporal parameters into a standard pairwise Markov random filed (MRF); thus restricting the frame correspondence and spatial transformation according to the neighborhood. Finally, this new formulation is able to handle the alignment of sequences recorded at different times by independent moving cameras that follows a similar trajectory, and also generalizes the 'classic' cases that of fixed geometric transformation and/or linear temporal mapping.

This paper is organized as follows: before going more deeply into the details of our method, the past works are reviewed in Section II. The joint formulation of the temporal and spatial alignment is presented in Section III. The inference of the most likely non-parametric temporal correspondence and a non-fixed (and unknown) geometric transformation is described in Section IV. Experiments on different video sequence pairs are presented in Section V. In addition, the proposed approach is compared with regard to the most close related work [5], [9], [10]. Finally, the main conclusions are drawn in Section VI.

II. Related Works

Several video alignment methods have been proposed in the literature. They are distinguished based on the assumptions made by each method. Table I classifies the methods based on the assumptions of the temporal correspondence and the relation between cameras. The most studied case assumes that the sequences are recorded by static or rigidly fixed cameras. Hence, these methods [2], [7], [8], [11]–[18] rely on an unknown and fixed geometric transformation, i.e., homography or fundamental matrix, along the whole sequence. Furthermore, most of these works also handle video sequences recorded simultaneously involving that the temporal correspondence is a simple constant time offset $t_r = t_o + \beta$ [12], [14], [16] or a linear relationship $t_r = \alpha t_o + \beta$ [2], [8], [13], [17], [18] to account for different camera frame rates. For instance, Caspi and Irani [2] presented video alignment solutions for static or jointly moving cameras. They align feature trajectories like in [15], [16], [18] or register directly the whole intensity manifolds, in order to estimate jointly an homography and an affine temporal mapping. A few works [11], [15] aim to align video sequences captured at different times showing slightly different object motions, such as the same action performed by different people. The main challenge of these works is the non-parametric temporal correspondence due to the difference between motions. For instance, Rao *et al.* [15] used a rank constraint of corresponding points to measure the similarity between trajectories together with a dynamic time warping; thus allowing to find the non-linear time-warping function. Moreover, these works are also differentiated by the input data, which can be either more or less difficult to obtain. Feature-based methods require: tracking of one or more characteristic points along both whole sequences [11], [15], [16], [19], or detecting interest points in space or space-time [12], [13], [17], [18]. In contrast, the so-called direct methods are based solely on the image intensity [2], Fourier transform of image intensity [14] or dynamic texture [8].

TABLE I
The Reviewed Algorithms are Classified Based on the Relationship Among Cameras and the Assumed Temporal Correspondence

		Temporal correspondence	
		affine $t_r = \alpha t_o + \beta$	unknown
cameras	Static	[12], [14], [2], [16], [18] [17], [8], [13], [7]	[15], [11]
	Moving	[20], [19]	[3], [5], [9], [6], [10]

A few works [3], [5], [6], [9], [10], [19], [20] address the challenging case of independently moving cameras. Their major challenge is to estimate a non-fixed geometric transformation along the whole sequence. That is, the geometric transformation between corresponding frame pairs changes from frame to frame. Lei and Yang [20] and Tuytelaars and VanGool [19] simplified the problem assuming an affine temporal correspondence since the videos recorded the same physical event by mobile cameras, e.g., football match broadcasting or athletic events. In particular, Tuytelaars and VanGool [19] find a temporal offset that minimizes the distance between the backprojected lines that have been tracked successfully through both sequences. In contrast, Lei and Yang [20] evaluate a tri-ocular geometric constraint among point or lines manually selected for a large number of hypothesis of temporal offsets. Other works [3], [5], [6], [9], [10] deal with sequences recorded at different times that entails the challenging task of estimating a non-parametric temporal correspondence. Kong *et al.* [6] only consider GPS information for associating independently corresponding frame pairs that are registered using a modified RANSAC. Sand and Teller [3] and Liu *et al.* [9] estimated a temporal and spatial alignment frame-by-frame minimizing a robust image alignment for many possible pairs of frames or a dense scene alignment for few retrieved pairs of frames, respectively. One of our previous work [5] decomposes the video alignment to estimate globally a temporal correspondence fusing the information obtained from two different sensors, a camcorder and a GPS receiver, and estimate the geometric transformation for each pair of corresponding frames frame-by-frame. Finally, the other previous work [10] estimates the temporal correspondence as a maximization of some overall similarity measure between candidates of corresponding frames, with regard to the parameters of the geometric transform. This estimation is based on a divide-and-conquer algorithm in order to avoid an exhaustive search. However, the shortcoming of this approach is that an error in earlier stages of the algorithm is propagated through the next estimations of the temporal mapping. Therefore, instead of tackling independently the spatial and temporal alignment or each frame independently as mentioned in the previous works, the proposed algorithm formulates jointly the spatial and temporal alignment exploiting the knowledge among neighbor frames.

III. Joint Video Alignment

Suppose we are given two video sequences $\mathbf{S}^o$ and $\mathbf{S}^r$ designated as observed and reference sequence, respectively. The reference provides the spatial and temporal reference; whereas the observed is mapped to match it. The observed sequence is assumed to be entirely contained within the reference and hence all the observed frames has associated a corresponding frame

in the reference sequence. The video alignment consists of associating the most similar frame in the reference sequence to each frame in the observed sequence (temporal alignment) together with the estimation of an unknown geometric transformation that relates the image-coordinates of this frame pair (spatial alignment). The synchronization, or temporal alignment, is formulated as an estimation of a temporal correspondence $\mathbf{x}_{1:n_o} = [x_1 \dots x_{n_o}]$, being $x_i \in \{1, \dots, n_r\}$ the index of the corresponding frame in reference sequence to the i^{th} frame in the observed sequence, and n_r and n_o are the number of frames in the reference and the observed sequence, respectively (Fig. 1). The registration, or spatial alignment, estimates an unknown parameters $\boldsymbol{\theta}_i$ of the geometric transformation model which relates the image coordinate systems of such pair of corresponding frames (i, x_i). For instance, $\boldsymbol{\theta}_i$ would be the 8 parameters defining the homography between the i^{th} pair of corresponding frames. Therefore, the video alignment consists of estimating the following parameters $\boldsymbol{\Theta} = [\mathbf{x}_{1:n_o}; \boldsymbol{\theta}_{1:n_o}]$.

Specifically, our main scenario is focused on aligning sequences that are recorded at different times from independent moving cameras. The trajectories of the cameras must be similar but not necessarily coincident; hence allowing trajectory dissimilarities between the sequences. Therefore, the problem of video alignment must deal with the challenging task of estimating a non-parametric temporal correspondence *and* an unknown and geometric transformation, e.g., an homography or affine transformation. In this scenario, a pair of corresponding frames are ideally assumed to be recorded at the same position but with different camera pose; thus under the assumptions that the two cameras have the same intrinsic parameters, the principal point (the origin of the image coordinate system) is at the image center and the focal length for the x and y axis are equal, the image-coordinates of this pair is related by a conjugate rotation $H = KR(\boldsymbol{\theta}_i)K^{-1}$ [5], being $K = diag(f, f, 1)$ and f the focal length. The relative camera pose $R(\boldsymbol{\theta}_i)$ is parametrized by the Euler angles $\boldsymbol{\theta_i} = [\theta_{x,i}, \theta_{y,i}, \theta_{z,i}]^T$.

Given an observed sequence $\mathbf{S}^o$, the joint video alignment $\boldsymbol{\Theta}$ is posed as a maximum *a posteriori* Bayesian inference problem,

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta} \in \mathcal{L}} p(\boldsymbol{\Theta}|\mathbf{S}^o), \tag{1}$$

where $\boldsymbol{\Theta}^*$ is the most likely spatio-temporal mapping between the observed and the reference sequence, and $\mathcal{L}$ is the set of all temporal correspondence and geometric transformation parameters. The posterior probability density $p(\boldsymbol{\Theta}|\mathbf{S}^o)$ is properly decomposed assuming Gibbs distributions [21] because this distribution allows to describe the pdf as a sum parts of energy functions. This decomposition is written as follows:

$$\begin{aligned} p(\boldsymbol{\Theta}|\mathbf{S}^o) &\propto p(\mathbf{S}^o|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) && (2)\\ &\propto p(\mathbf{S}^o|\mathbf{x}_{1:n_o}, \boldsymbol{\theta}_{1:n_o})p(\mathbf{x}_{1:n_o})p(\boldsymbol{\theta}_{1:n_o}) && (3)\\ &\propto \frac{1}{Z(\boldsymbol{\Theta})} e^{-E(\mathbf{S}^o;\boldsymbol{\Theta}) - E(\mathbf{x}_{1:n_o}) - E(\boldsymbol{\theta}_{1:n_o})}, && (4) \end{aligned}$$

where $Z(\boldsymbol{\Theta})$ is the partition function, $E(\mathbf{S}^o; \boldsymbol{\Theta}) = -\log p(\mathbf{S}^o|\boldsymbol{\Theta})$ is the energy of joint video alignment between the observed and the reference sequence, and $E(\mathbf{x}_{1:n_o}) = -\log p(\mathbf{x}_{1:n_o})$ and $E(\boldsymbol{\theta}_{1:n_o}) = -\log p(\boldsymbol{\theta}_{1:n_o})$ are the energy of regularization terms in temporal and spatial dimensions, respectively. For simplicity, the temporal correspondence $x_{1:n_o}$ and the spatial correspondence $\boldsymbol{\theta}_{1:n_o}$ are assumed to be independent since the relative longitudinal motion between the cameras described by $x_{1:n_o}$ has not a direct correlation with relative dissimilarities on the camera trajectories described by $\boldsymbol{\theta}_{1:n_o}$. Therefore, this independence allows to include simple assumptions on the dynamics of the camera avoiding complex priors. The meaning of these likelihoods and these assumptions are defined below. To summarize, the joint video alignment $\boldsymbol{\Theta}$ is reformulated as follows:

$$\begin{aligned} \boldsymbol{\Theta}^* &= \arg\min_{\boldsymbol{\Theta} \in \mathcal{L}} -\log p(\boldsymbol{\Theta}|\mathbf{S}^o) && (5)\\ &= \arg\min_{\boldsymbol{\Theta} \in \mathcal{L}} E(\mathbf{S}^o; \boldsymbol{\Theta}) + E(\mathbf{x}_{1:n_o}) + E(\boldsymbol{\theta}_{1:n_o}). && (6) \end{aligned}$$

A. Joint Video Alignment Energy

Joint video alignment energy $E(\mathbf{S}^o; \mathbf{x}_{1:n_o}, \boldsymbol{\theta}_{1:n_o})$ aims to evaluate the similarity between a pair of sequences given a spatio-temporal mapping, $\mathbf{x}_{1:n_o}$ and $\boldsymbol{\theta}_{1:n_o}$. For simplicity, $E(\mathbf{S}^o; \mathbf{x}_{1:n_o}, \boldsymbol{\theta}_{1:n_o})$ is evaluated independently given a pair of corresponding frames and a geometric transformation (Fig. 2) as follows:

$$E(\mathbf{S}^o; \mathbf{x}_{1:n_o}, \boldsymbol{\theta}_{1:n_o}) = \sum_{i=1}^{n_o} E(S_i^o; x_i, \boldsymbol{\theta}_i), \tag{7}$$

where $E(S_i^o; x_i, \boldsymbol{\theta}_i)$ measures similarity between a pair of frames (i, x_i) considering the geometric transformation $\boldsymbol{\theta}_i$. Due to the combination of discrete and continuous hidden variables, this likelihood is broken into two components as follows:

$$E(S_i^o; x_i, \boldsymbol{\theta}_i) = \varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i), \tag{8}$$

where $\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ denoted as temporal-correspondence measures the correspondence based on feature points between pair of frames (i, x_i) given the geometric transformation $\boldsymbol{\theta}_i$; whereas $\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i)$ denoted as spatial-correspondence or frame-alignment evaluates the spatial alignability according to $\boldsymbol{\theta}_i$ and image intensities given the most likely corresponding frames (i, x_i). Both terms will be described below. Hence, $E(S_i^o; x_i, \boldsymbol{\theta}_i)$ combines jointly temporal and spatial alignment together with the benefits of feature-based and direct-based methods. Therefore, the spatio-temporal alignment in a pair of sequences is defined to satisfy simultaneously the similarity of content (feature-based method) and the image registration (direct-based method) along the whole sequence. One important point to note here is that neither the temporal correspondence nor spatial-correspondence by themselves are good models for the joint spatio-temporal similarity $E(S_i^o; x_i, \boldsymbol{\theta}_i)$. Only their combination results in a good model, e.g., for video alignment.

1) Temporal-Correspondence: The aim of temporal-correspondence $\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ is to measure the similarity of content between the i^{th} frame in the observed sequence warped according to $\boldsymbol{\theta}_i$ and the x_i^{th} frame in the reference sequence.

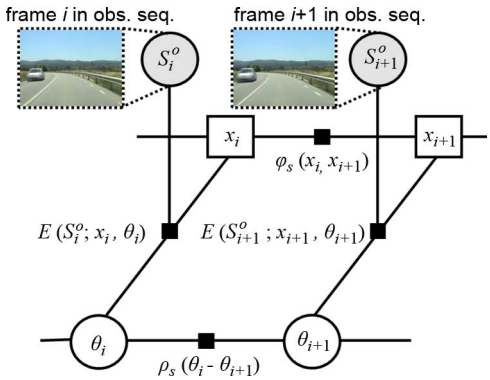


Fig. 2. Illustrative example of the joint video alignment formulation. $\varphi_S(x_i, x_{i+1})$ and $\rho_S(\boldsymbol{\theta}_{d,i+1} - \boldsymbol{\theta}_{d,i})$ allow us to restrict the smoothness of the parameters in temporal and spatial dimensions, respectively; whereas $E(S_i^o; x_i, \boldsymbol{\theta}_i)$ measures the similarity likelihood between two frames. Square nodes represent discrete variables and circular nodes continuous variables. Shaded nodes denote observed variables; non-shaded nodes are hidden variables.

This similarity is based on counting the number of similar feature descriptors between this pair. This is computed efficiently based on image retrieval techniques [22], [23] as follows: first, SIFT descriptors [24] are extracted for each frame in the reference sequence. Second, a dictionary of n visual words (VW) is learned from a reference sequence by running K-means on the extracted SIFT descriptors. The value of n is typically set to 5% of extracted descriptors. Hence, each image can be described simply as a set of visual words in $\{1, \ldots, n\}$. Finally, all frames of the reference sequence are set up as an inverted file. That is, the q^{th} row of the inverted file corresponds to the q^{th} visual word learned by K-means; then this row contains the indexes of frames, $l_q \in \{1, \ldots, n_r\}$, where the q^{th} visual word appeared in the frames of the reference sequence. Therefore, the correspondence-similarity between pairs of frames is defined as the inverse of the weighted sum of VW ocurred in both frames:

$$\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i) = \frac{1}{\sum_{q=1}^{n_f} \omega_{f_q} \mathbf{1}_{(l_{f_q})}(x_i)}, \tag{9}$$

where $S_i^o(u; \boldsymbol{\theta}_i)$ is i^{th} frame in the observed sequence warped according $\boldsymbol{\theta}_i$ as defined below, $u = (x, y)$ is the pixel coordinates of an image, n_f is the number of descriptors in $S_i^o(u; \boldsymbol{\theta}_i)$, f_q is the VW corresponding to the q^{th} SIFT descriptor, $\mathbf{1}_{(l_{f_q})}(x_i)$ is a indicator function that is equal to 1 if x_i is indexed in l_{f_q} and 0 otherwise, and ω_{f_q} is the inverse of the total number of frames indexed in l_{f_q} as *inverse document frequency* (idf) scores used in text retrieval [25] to rely mainly on the VWs that appear less times.

2) Spatial-Correspondence: The aim of spatial-correspondence $\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i)$ is to measure the similarity of spatially aligning the frame i in the observed sequence to the frame x_i in the reference sequence. In order to make it more robust to alignment outliers, the similarity is computed as the sum of intensity-differences evaluated by a robust function Ω_D as follows:

$$\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i) = \sum_u \Omega_D\left(S_{x_i}^r(u) - S_i^o(u; \boldsymbol{\theta}_i)\right), \tag{10}$$

where $S_i^o(u; \boldsymbol{\theta}_i)$ is equal to $S_i^o(u + \mathcal{U}(u; \boldsymbol{\theta}_i))$, $\mathcal{U}(u; \boldsymbol{\theta}_i) = \mathcal{X}(u) \cdot \boldsymbol{\theta}_i$ is the quadratic approximation of the geometric transformation between corresponding frames [26], [27], $\mathcal{X}(u)$ is a matrix which depends only on the pixel coordinates as shown in (13), $S_{x_i}^r(u)$ is the x_i^{th} frame in the reference sequence. Under the assumptions that these angles are small and the focal length is large enough, $S_i^o(u; \boldsymbol{\theta}_i)$ is approximated using a Taylor series expansion as follows:

$$S_i^o(u; \boldsymbol{\theta}_i) \approx S_i^o(u) + \nabla^T S_i^o \mathcal{X}(u) \boldsymbol{\theta}_i, \tag{11}$$

where $\nabla S_i^o = [\partial S_i^o / \partial x, \partial S_i^o / \partial y]^T$ is the gradient of the i^{th} frame in the observed sequence along x- and y-directions. Without loss of generality, the quadratic approximation of the conjugate rotation [5], [10], [26] is written as follows:

$$\mathcal{U}(u; \boldsymbol{\theta}_i) = \mathcal{X}(u) \boldsymbol{\theta}_i, \tag{12}$$

$$= \begin{bmatrix} -\frac{xy}{f} & f + \frac{x^2}{f} & -y \\ -f - \frac{y^2}{f} & \frac{xy}{f} & x \end{bmatrix} \begin{bmatrix} \theta_{x,i} \\ \theta_{y,i} \\ \theta_{z,i} \end{bmatrix}. \tag{13}$$

B. Temporal and Spatial Regularization Energy

Temporal and spatial regularization aim to introduce a prior knowledge of the sequence pair into the temporal and spatial alignment (Fig. 2). This knowledge restricts the estimation of the spatio-temporal parameters $\boldsymbol{\Theta}$. On one hand, temporal regularization $E(\mathbf{x}_{1:n_o})$ favors only temporal mappings that satisfy assumptions on the camera motion, e.g., cameras moves forward. For simplicity, this is evaluated considering only consecutive frame correspondences as follows:

$$E(\mathbf{x}_{1:n_o}) = \sum_{i=1}^{n_o - 1} \lambda_T \varphi_S(x_i, x_{i+1}), \tag{14}$$

where λ_T is a trade-of between the frame-correspondence and the influence of smoothing the temporal correspondence. Specifically, the vehicle movement and the content of the videos are assumed not to reverse its motion in none of the sequences though it allows each of them to stop at any time independently and, of course, move forward. Hence, temporal alignment must increase monotonically. This assumption is modeled by setting $\varphi_S(x_i, x_{i+1})$ as follows:

$$\varphi_S(x_i, x_{i+1}) = \begin{cases} 0 & \text{if } x_{i+1} \geq x_i \\ \beta & \text{otherwise,} \end{cases} \tag{15}$$

where β penalizes frame correspondences that breaks the monotonically increasing constraint.

On the other hand, spatial regularization $E(\boldsymbol{\theta}_{1:n_o})$ penalizes high variations among successive parameters of the geometric transformation $\boldsymbol{\theta}_{1:n_o}$. Specifically, this variations may be smooth along time since the relative pose between cameras changes slightly from frame to frame in most cases. This constraint is due to the smooth motion of the camera. Therefore, the spatial regularization term is evaluated as follows:

$$E(\boldsymbol{\theta}_{1:n_o}) = \sum_{i=1}^{n_o - 1} \sum_{d}^{N_d} \lambda_S \rho_S(\theta_{d,i+1} - \theta_{d,i}) \tag{16}$$

where λ_S controls the influence of the regularization term to restrict the smoothness of geometric transformation among successive frames, and N_d is the number of variables in the geometric transformation, e.g., 3 parameters in a conjugate rotation. Each geometric variable is evaluated by the robust function ρ_S, e.g., Charbonnier or L-2 penalty function. Notice that the temporal and spatial regularization terms are a discrete and continuous pairwise MRF, respectively.

IV. Optimization

The estimation of joint video alignment parameters $\boldsymbol{\Theta}$, a non-parametric temporal correspondence and non-fixed geometric transformations, consists of minimizing the following equation:

$$\begin{aligned}\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}\in\mathcal{L}} &\sum_{i=1}^{n_o-1} \lambda_T \varphi_S(x_i, x_{i+1}) \\ &+ \sum_{i=1}^{n_o} \varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)\rho_D(S_i^o, \boldsymbol{\theta_i}; x_i) \\ &+ \sum_{i=1}^{n_o-1} \sum_{d=\{x,y,z\}} \lambda_S \rho_S(\theta_{d,i+1} - \theta_{d,i}). \end{aligned} \quad (17)$$

The optimization of (17) is however not jointly convex in the pair $(\mathbf{x}_{1:n_o}, \boldsymbol{\theta}_{1:n_o})$, but convex with respect to each of the two variables ($\mathbf{x}_{1:n_o}$ and $\boldsymbol{\theta}_{1:n_o}$) when the other one is fixed. A natural approach to solving this problem is to alternate the minimization of $\mathbf{x}_{1:n_o}$ or $\boldsymbol{\theta}_{1:n_o}$ since (8) is broken into frame-correspondence and frame-alignment. That is, a temporal alignment estimates globally the most likely frame correspondence based on the reliability of the geometric transformation; whereas the geometric transformations are estimated based on the reliability of the temporal correspondence. These iterative minimization proceeds as follows:

1. For $\boldsymbol{\theta}_{1:n_o}$ being fixed, (17) is solved for $\mathbf{x}_{1:n_o}$:

$$\begin{aligned}\min_{\mathbf{x}_{1:n_o}} &\sum_{i=1}^{n_o-1} \lambda_T \varphi_S(x_i, x_{i+1}) \\ &+ \sum_{i=1}^{n_o} \rho_D(S_i^o, \boldsymbol{\theta}_i; x_i) \cdot \varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i) \end{aligned} \quad (18)$$

   where $\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i)$ is evaluated for all the reference frames $x_i \in \{1, \ldots, n_r\}$, keeping the given $\boldsymbol{\theta}_i$ fixed. The most likely frame correspondence $\mathbf{x}_{1:n_o}$ is inferred using the min-sum inference algorithm [28], [29] due to the efficient computation of $\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ for all x_i and $\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i)$. This inferred correspondence $\mathbf{x}_{1:n_o}$ is a good initialization to estimate $\boldsymbol{\theta}_{1:n_o}$ since the corresponding frames maximizes the similarity of content w.r.t. the frames in the observed sequence.

2. For $\mathbf{x}_{1:n_o}$ being fixed, (17) is solved for $\boldsymbol{\theta}_{1:n_o}$:

$$\begin{aligned}\min_{\boldsymbol{\theta}_{1:n_o}} h(\boldsymbol{\theta}_{1:n_o}) = \min_{\boldsymbol{\theta}_{1:n_o}} &\sum_{i=1}^{n_o-1} \sum_{d}^{N_d} \lambda_S \rho_S \left(\theta_{d,i+1} - \theta_{d,i}\right) \\ &+ \sum_{i=1}^{n_o} \sum_{x} \beta_i \cdot \Omega_D \left(S_{x_i}^r(x) - S_i^o(x; \boldsymbol{\theta_i})\right) \end{aligned} \quad (19)$$

   where $\beta_i = \varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ is the similarity between a pair of corresponding frames (i, x_i), keeping fixed, and $S_i^o(u; \boldsymbol{\theta}_i)$ is approximated as (11). Therefore, the parameters of the geometric transformation are estimated considering the frame-correspondence likelihood $\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ in order to compensate spatially low reliabilities of the temporal correspondence. The optimization of (19) may in fact be non-convex, and difficult to minimize depending on the choice of the robust penalty functions $\Omega_D(\cdot)$ and $\rho_S(\cdot)$. Instead of finding a global optimum in (19), a simple local optimization is performed by computing the gradient and setting to zero. Then, the partial derivatives with respect to $\boldsymbol{\theta}_i$ are:

$$\begin{aligned}\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \rightarrow \lambda_S &\left[\left(C_i^S - C_{i+1}^S\right)\boldsymbol{\theta}_i + C_{i+1}^S \boldsymbol{\theta}_{i+1} - C_{i-1}^S \boldsymbol{\theta}_{i-1}\right] \\ &+ C_i^D(\boldsymbol{\theta}_i) \cdot \boldsymbol{\theta}_i + b_i(\boldsymbol{\theta}_i) \end{aligned} \quad (20)$$

   with

$$C_i^D = \sum_x \beta_i \tilde{\Omega}_D(S_t) \tilde{\mathcal{U}}^T \tilde{\mathcal{U}}, \quad (21)$$

$$b_i = \sum_x \beta_i \tilde{\Omega}_D(S_t)(S_t) \tilde{\mathcal{U}}^T, \quad (22)$$

$$C_i^S = [\tilde{\rho}_S(\theta_{1,i} - \theta_{1,i-1}), \ldots, \tilde{\rho}_S(\theta_{N_d,i} - \theta_{N_d,i-1})]^T, \quad (23)$$

   where $\tilde{\mathcal{U}}$ is defined as $\nabla^T S_i^o \mathcal{X}(u)$ in the previous section, $S_t = S^r(u) - S_i^o(u; \boldsymbol{\theta}_i)$ denotes the pixel difference between a pair of corresponding frames, and $\tilde{\rho}_S(y)$ and $\tilde{\Omega}_D(y)$ is defined as $\rho_S'(y)/y$ and $\Omega_D'(y)/y$, respectively. Note that the partial derivative w.r.t. $\boldsymbol{\theta}_i$ is a function that depends on the current spatial variables $\boldsymbol{\theta}_i$ and the consecutive adjacent spatial variables $\boldsymbol{\theta}_{i-1}$ and $\boldsymbol{\theta}_{i+1}$. Hence, the estimation of $\boldsymbol{\theta}_i$ is not tackled independently. Finally, all partial derivatives are set to 0, and the terms are regrouped in matrix-vector form as follows:

$$[C^D(\tilde{\boldsymbol{\theta}}) + \lambda_S C^S(\tilde{\boldsymbol{\theta}})] \cdot \tilde{\boldsymbol{\theta}} = -b_i(\tilde{\boldsymbol{\theta}}) \quad (24)$$

   where $\tilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}_{1,1:n_o}, \ldots, \boldsymbol{\theta}_{N_d,1:n_o}]^T$ is the spatial parameters warped into a vector, and $C^D(\tilde{\boldsymbol{\theta}})$ and $C^S(\tilde{\boldsymbol{\theta}})$ are sparse matrices ordered properly that depend on $\tilde{\boldsymbol{\theta}}$, and $b_i(\tilde{\boldsymbol{\theta}})$ is a vector that also depends on $\tilde{\boldsymbol{\theta}}$. In order to estimate $\tilde{\boldsymbol{\theta}}$, (24) is made linear by keeping $C^D(\tilde{\boldsymbol{\theta}})$, $C^S(\tilde{\boldsymbol{\theta}})$ and $b_i(\tilde{\boldsymbol{\theta}})$ fixed, and then solve the resulting linear equation system using a standard technique [30]. After finding a new estimate, this procedure is iterated until a fixed point is reached.

A. Implementation Details

The linearized brightness for S^o in (11) uses the first order Taylor approximation valid only on small displacements. Hence, the estimation of $\boldsymbol{\theta}$ involves a coarse-to-fine strategy. The image pyramid is constructed using a scale factor of 0.5 to speed-up the convergence. The number of pyramid levels is set to 5. Furthermore, the proposed algorithm needs a warping scheme at each pyramid level to solve (19) in order to deal with image non-linearities. The robust function Ω_D and ρ_S are the Charbonnier penalty function defined as $\sqrt{x^2 + \epsilon^2}$ and uses $\epsilon = 0.001$ for both terms. This robust function has been proved to be useful in optical flow. Moreover, the trade-off parameters

TABLE II
THE CLASSIFICATION OF THE EIGHT VIDEO SEQUENCE PAIRS USED TO VALIDATE THE PROPOSED METHOD AND THE NOMENCLATURE OF EACH GROUP IS INSIDE THE PARENTHESIS

		Temporal correspondence	
		affine $t_r = \alpha t_o + \beta$	unknown
cameras	Static	'water' in [2] (SS)	'jump' in [32] and 'dancing' in [15] (SD)
	Moving	'indoor–2' in [33] (MS)	'highway', 'campus', 'back–road' in [5] and 'indoor–1' in [3] (MD)

TABLE III
PERFORMANCE OF DIFFERENT VIDEO ALIGNMENT ALGORITHMS UNDER DIFFERENT SCENARIOS WITH DISSIMILAR IMAGE CONTENT

Scenarios	Algorithm	% of frames with		
		$err = 0$	$err < 1$	$err < 2$
All	Diego *et al.* [5]	58.12	66.79	70.79
	Serrat *et al.* [10]	73.2	81.86	83.04
	Liu *et al.* [9]	77.26	87.30	91.38
	Our Approach	**84.20**	**93.04**	**96.22**
Campus	Diego *et al.* [5]	78.08	90.77	94.23
	Serrat *et al.* [10]	76.47	**95.81**	**98.9**
	Liu *et al.* [9]	77.95	89.04	93.46
	Our Approach	**82.05**	91.03	94.87
Highway	Diego *et al.* [5]	49.55	54.92	59.53
	Serrat *et al.* [10]	66.15	70.69	72.92
	Liu *et al.* [9]	76.06	87.16	91.49
	Our Approach	**87.51**	**96.16**	**98.74**
Back–road	Diego *et al.* [5]	43.53	51.28	55.00
	Serrat *et al.* [10]	76.99	79.40	82.57
	Liu *et al.* [9]	77.77	85.05	88.97
	Our Approach	**83.11**	**92.02**	**95.04**

λ_S and λ_T control the influence of the regularization terms. Thus, λ_S used for aligning sequences recorded by independent moving cameras, which is set empirically to 500, is lower than λ_S used for aligning sequences recorded by static cameras, which is set to 2000, since a large λ_S involves approximately a fixed geometric transformation along the whole sequence. The final parameter β in (15) controls the influence of temporal smoothness on the optimization framework. Therefore, setting β to 0 implies that there is no temporal assumption on the optimization in (18), e.g., each pair of corresponding frames are estimated independently; whereas setting $\beta > 0$, in practice equal to 100, enforces that the temporal correspondence increase monotonically based on β. Finally, the optimization starts setting the $\boldsymbol{\theta}_{1:n_o}$ to zero and followed by computing the frame-correspondence using (9) and the spatial-correspondence using (10) for all the pairs of frames. Then, the optimization alternates between the estimation of the temporal mapping and the spatial correspondence until the parameters do not change significantly or reach a maximum number of iterations. In practice, the stopping criterion relies on the relative decrease (typically 10^{-3}) in the cost function in (16) and the maximum number of iterations equal to 10. The algorithm is currently implemented in Matlab code using VLFeat library [31] on an Intel(R) Core(TM)2 Duo CPU E6850 at 3 GHz with 2 Gb of RAM.

V. EXPERIMENTS

In this section, experiments to validate the proposed approach are conducted on different pairs of video sequences. First, the synchronization performance is evaluated on sequences recorded at different times by independent moving cameras following similar trajectories, and compared to the closest related works [5], [9], [10]. Second, the proposed method is applied to other synchronization cases that involve an unknown and fixed geometric transformation or a parametric temporal mapping to validate its robustness and applicability.

Algorithm 1 Optimization of Joint Video Alignment

1: **Input** : S^r (reference sequence), S^o (observed sequence), λ_S and λ_T (regularization parameters), β (temporal smoothness) and the geometric transformation to use (e.g., homography, affine, etc.).

2: **Initialization**: $\boldsymbol{\theta}_{1:n_o} \leftarrow 0$

3: **While** (not reaching stopping criteria) **do**

4: **Update** $\mathbf{x}_{1:n_o}$:

5: Compute $\rho_D(S_i^o, \boldsymbol{\theta}_i; x_i)$ using (10).

6: Compute $\varphi_D(S_i^o, x_i; \boldsymbol{\theta}_i)$ using (9).

7: Compute $\mathbf{x}_{1:n_o}$ using min-sum inference algorithm on (17)

8: **Update** $\boldsymbol{\theta}_{1:n_o}$:

9: **While** (not reaching a fixed point) **do**

10: Compute $\boldsymbol{\theta}_{1:n_o}$ using (23)

11: **end while**

12: **end while**

13: **Output**: $\mathbf{x}_{1:n_o}$ (temporal mapping between sequences), $\boldsymbol{\theta}_{1:n_o}$ (geometric transformation between corresponding frames)

A. Datasets

Experiments are conducted on eight video sequence pairs. Those pairs are divided in 4 groups based on the relation between cameras and the assumed temporal correspondence as shown in Table II. The first group denoted by MD consists of 4 video sequence pairs recorded at different times by independent moving cameras on different scenarios with a huge dissimilar image content. Three of these sequences ('highway', 'campus' and 'back-road') are recorded with a facing-camcorder attached at the windscreen at different times. The vehicle must keep on the same lane, that is, the lateral displacement of the camera is bounded to about ± 2.5 meters. Variations in the camera pose at the same place occur because of the different heading of the vehicle. They are modeled by a 3D rotation in which the most important factor is the yaw angle. Large variations in yaw angle occur in short intervals but the constant and maximum allowed

Fig. 3. Sampled frames from the fusion of three pairs of aligned videos ('campus', 'highway' and 'backroad'). The greenish and purple regions of unnatural color appear because of either slight misalignment of objects common to both corresponding frames or the different scene content (vehicles). The three complete original video sequence pairs and their fusion can be viewed at *www.cvc.uab.es/~fdiego/Joint/*.

relative pose between cameras is 3° since a vehicle will cross a road lane in 50 meters (3 sec at 50 kph) if a dissimilar trajectory of 3°, which means a 90% of frame overlapping, is fixed. In contrast, the other ('indoor–1') available in [3] is recorded by hand-held cameras in a indoor scenario. The first three pairs constitute a challenging task for video alignment because frequent and large velocity differences exist between reference and observed sequences at the same location, e.g., differences of

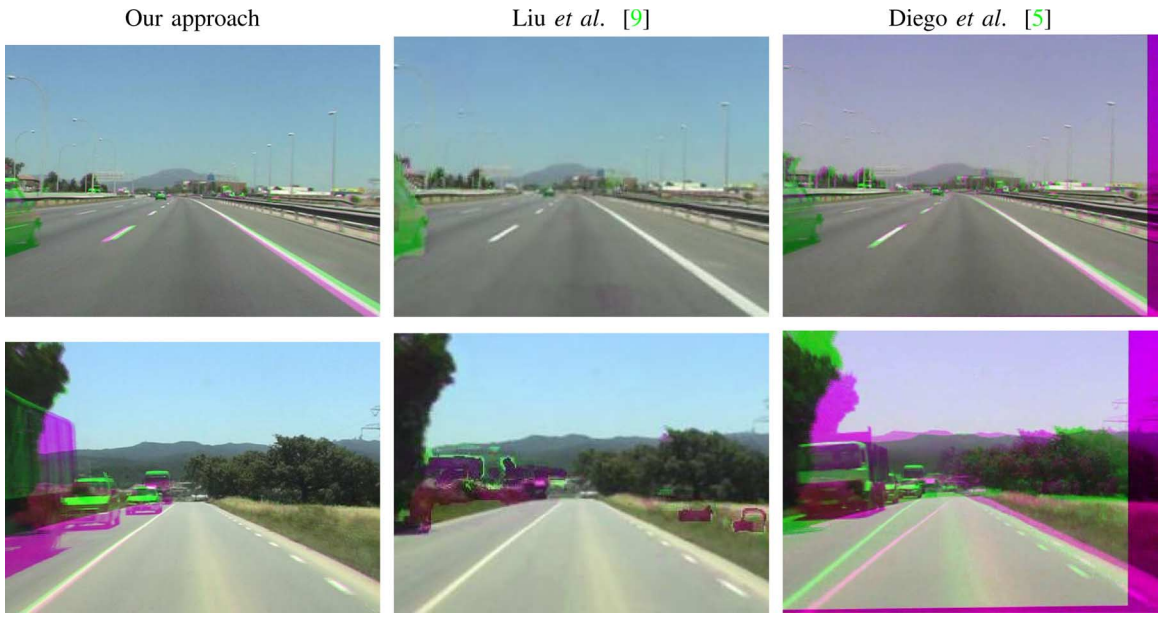


Fig. 4. Example results of the compared video alignment algorithms. More results can be viewed at http://www.cvc.uab.es/~fdiego/Joint/.

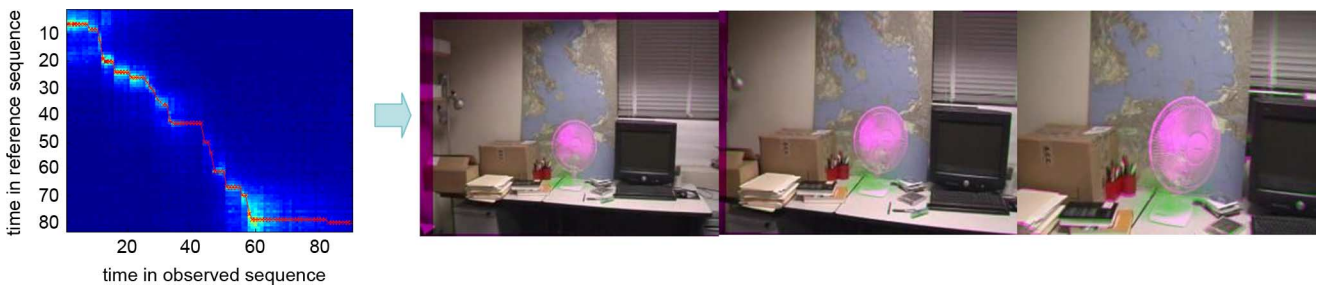


Fig. 5. From left to right: the joint video alignment energy computed by (8) inside the range defined by time (frames) x_i in reference sequence/time (frame) i in observed sequence. The red line is the most likely frame correspondence $\mathbf{x}_{1:n_o} = [x_1 \dots x_{n_o}]$ found using Algorithm 1. Example results of our approach on 'indoor-1' pair provided by Sand and Teller [3] in order to illustrate some representative results.

± 30 kph occur for relatively long intervals. The second group denoted by SD consists of 2 video sequence pairs ('jump' and 'dancing' available in [32] and [15], respectively) recorded at different times by static cameras. These sequences show slightly different human action done by different people and hence the temporal correspondence follows a non-parametric form. Finally, the last two groups denoted by SS ('water' available in [2]) and MS ('indoor—2' available in [33]) consists of one video sequence pair recorded simultaneously by static and moving cameras, respectively. 'Indoor–2' is not strictly recorded by moving cameras, but it consists of a stereo video sequence pair whose observed sequence is warped according to an affine transformation that changes smoothly along time.

We annotate manually a ground-truth (GT) for 'highway', 'campus' and 'back-road' pair to quantitatively assess the synchronization performance on the most general alignment problem. This assessment is also compared with the closest related work [5], [9]. The rest of sequences available has not any ground-truth associated since the previous works are only evaluated qualitatively. The annotated GT consists of a narrow interval $[l_t, u_t]$ in a reference sequence for one out every 10 frames in the observed sequence. This interval always contains the true corresponding frame because of the difficult decision of determining a single corresponding frame. The frames without an available ground-truth are interpolated linearly from time $[l_t, u_t]$ to $[l_{t+10}, u_{t+10}]$. The average length of that interval is 3 frames long. Hence, the synchronization error at time t, given the corresponding frame x_t, is defined as the distance of x_t to the closest ground-truth interval boundary as follows:

$$\mathrm{err}(t, x_t) = \begin{cases} 0 & \text{if } l_t \leq x_t \leq u_t \\ \min\left(|l_t - x_t|, |u_t - x_t|\right) & \text{otherwise.} \end{cases} \tag{25}$$

B. Performance Evaluation

In this section, the proposed algorithm is validated by evaluating the synchronization performance whether regularization terms (second and third term in (4)) are included or not. Furthermore, it is compared with the three closest related works [5], [9], [10]. Our previous work [5] estimates a complete temporal correspondence maximizing a image- and location-similarity based on image appearance descriptors and GPS data, respectively; whereas the geometric transformation is estimated frame-by-frame using Lukas-Kanade algorithm [34]. In contrast, the other previous work [10] estimates the temporal correspondence as a maximization of some overall similarity measure between candidates of corresponding frames, with regard to the parameters of the geometric transform. This estimation is based on

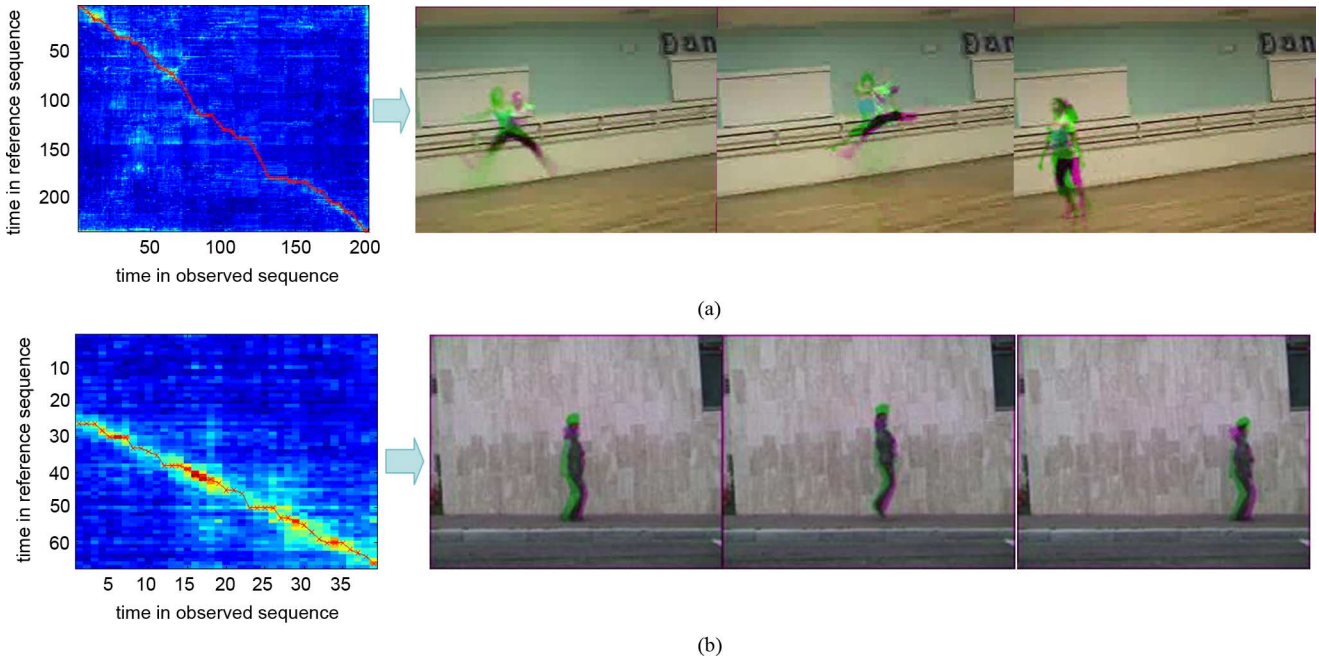


Fig. 6. From left to right: the joint video alignment energy computed by (8) inside the range defined by time (frames) x_i in reference sequence/time (frame) i in observed sequence. The red line is the most likely frame correspondence $\mathbf{x}_{1:n_o} = [x_1 \dots x_{n_o}]$ found using Algorithm 1. Example results of our approach on a) 'dance' and b) 'jump' pair provided by Rao *et al.* [15] and Gorelick *et al.* [32], respectively.

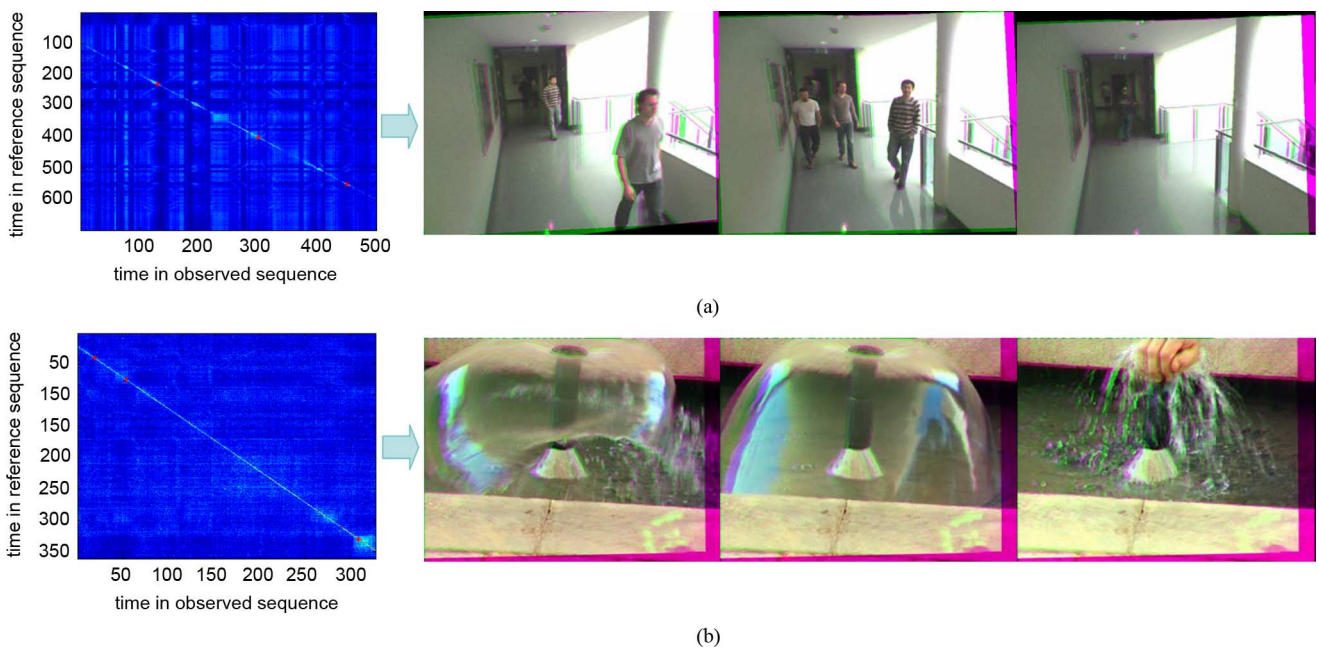


Fig. 7. From left to right: the joint video alignment energy computed by (8) inside the range defined by time (frames) x_i in reference sequence/ time (frame) i in observed sequence. The red line is the most likely frame correspondence $\mathbf{x}_{1:n_o} = [x_1 \dots x_{n_o}]$ found using Algorithm 1. Example results of our approach on 'indoor-2' and 'water' pairs provided by Kelly *et al.* [33] and Caspi and Irani *et al.* [2], respectively.

a divide-and-conquer algorithm in order to avoid an exhaustive search. The third algorithm [9] is adapted to video alignment. This consists of estimating a temporal correspondence and the spatial alignment frame-by-frame as follows. First, for each frame in the observed frame, the top-20 frames are retrieved matching a spatial histogram of quantized SIFT. Then, these 20 frames are re-ranked based on the flow energy using a robust dense scene alignment; thus the frame with the lowest energy is considered as the corresponding frame.

A few sample frames of the proposed algorithm on 'highway', 'campus' and 'back-road' pairs are shown in Fig. 3. The performance of the algorithms is outlined in Table III. From the results we derive that the proposed algorithm is robust under different scenarios. When the regularization terms are removed, the average performance of perfect alignment is 78.87% of observed frames. That result shows that (9) accurately discriminates corresponding frames, and improves the best performance of Liu's method 77.26% and our pre-

vious works 58.12% and 73.2%. The best Liu's performance is obtained by retrieving the top-20 frames and choosing the frame with the best alignment. However, when it retrieves just one frame as we do, the average performance of Liu's method decreases from 77.26% to 67.88% frame; thus reinforcing the frame discrimination of the proposed algorithm. In addition, the average performance of the proposed method increases from 78.87% to 84.20% when the regularizations are included. More than 91% of observed frames of all video sequences pairs have a synchronization error up to 1-frame. Hence, the performance w.r.t. the state-of-the-art algorithms improves approximately more than 6%. Moreover, the computational time is the main drawback of Liu's method because only the robust dense scene alignment at half-resolution takes more than 30 sec per image pair that does not consider the computational time for retrieving the 20-most likely frames. For instance, the alignment between two video sequences with 1,000 frames using Liu's method takes approximately 7 days because Liu's method needs to estimate 20 times a robust dense scene alignment for each frame in the observed sequence; thus it requires to estimate 20,000 scene alignments that needs an average of 30 sec per alignment (approx. 600,000 sec). In contrast, the frame correspondence in (18) takes 5 sec at most for each observed frame; whereas the frame-alignment implemented in MATLAB requires 15 sec for each pair of corresponding frames. That is, the total computational time required is equal to 20 sec per frame. Although the computational time is similar in both methods, the proposed algorithm requires 3 iterations at most to converge; whereas Liu's method needs 20 iterations to just retrieve a single frame.

To assess the quality of the spatial alignment, a simple image fusion that assigns the observed frame to the red and blue channels and the registered corresponding frame to the green channel is performed. Fig. 4 shows a qualitative comparison using some aligned frames for each video alignment algorithm on 'highway', 'campus' and 'back-road'. Fig. 5 shows a few alignment samples of the proposed algorithm on 'indoor-1' pair. Using the proposed method, Figs. 3 and 4 shows the estimation of the geometric transformation with high sub-pixel accuracy dealing with high dissimilar image content since this estimation considers the adjacent geometric transformations and the reliability of the temporal correspondence estimation. The work proposed by Liu *et al.* [9] improves qualitatively the pixel-wise correspondence at the expense of a slight loss of accuracy in matching dissimilar image contents. Note that the goal on [9] is a pixel-wise dense alignment instead of estimating a geometric transformation.

More qualitative results for SD, MS and SS groups are shown in Figs. 6, 7(a) and 7(b), respectively. In particular, Figs. 7(b) and 7(a) show that the proposed algorithm handles the estimation of an affine temporal mapping on sequences recorded by static or moving cameras, respectively. The fixed geometric transformation shown in Fig. 7(b) is estimated by increasing substantially λ_S w.r.t. the value used for aligning sequences recorded by moving cameras shown in Fig. 7(a). Increasing the value allows the estimation of a fixed geometric transformation. Finally, Fig. 6 shows that the proposed method handles sequences of different human actions acquired at different times assuming there is enough motion in the field-of-view. Furthermore, the conjugate rotation has been only applied on 'highway', 'campus' and 'back-road' video sequence pairs since they satisfies the assumptions of this geometric transformation; whereas an homography has been applied as the unknown geometric transformation in the other sequences. A figure containing just a few frames of the aligned videos would be a poor reflection of the results. The original and fully aligned video sequences can be viewed at the web page http://www.cvc.uab.es/~fdiego/Joint/.

VI. Conclusions and Future Work

In this paper, joint video alignment is proposed to deal with the challenging problem of aligning video sequences; specifically these sequences are recorded by independently moving cameras at different times following approximately similar but no coincident trajectories without changing the direction. The temporal and spatial alignment are formulated jointly as a minimization over a huge number of parameters, e.g., non-parametric temporal correspondence and geometric transformation. The algorithm combines image-retrieval and image-registration techniques for computing frame-correspondence and frame-alignment, respectively. The best spatio-temporal alignment is solved iteratively alternating a min-sum algorithm and a gradient descent algorithm using the best result found in the previous result. The new video alignment formulation also generalizes other relative simple cases like the alignment of sequences recorded simultaneously or at different times by static cameras. Experiments conducted on different video sequence pairs show that the proposed video alignment deals with the presence of moving objects, different image content, variations on camera velocity and different scenarios, indoor and outdoor. The proposed method outperforms the synchronization accuracy w.r.t. the state-of-the-art on sequences recorded from vehicles driving along the same track at different times. As further work, a high-order prior can be introduced to model properly the dynamics of the camera motion.

References

[1] G. P. Stein, "Tracking from multiple view points: Self-calibration of space and time," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition*, 1999, pp. 1521–1521.

[2] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1409–1424, 2002.

[3] P. Sand and S. Teller, "Video matching," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 22, no. 3, pp. 592–599, 2004.

[4] Y. Ukrainitz and M. Irani, "Aligning sequences and actions by maximizing space-time correlations," in *Proc. ECCV*, 2006, vol. 3953, pp. 538–550.

[5] F. Diego, D. Ponsa, J. Serrat, and A. M. Lopez, "Video alignment for change detection," *IEEE Trans. Image Process.*, 2010.

[6] H. Kong, J.-Y. Audibert, and J. Ponce, "Detecting abandoned objects with a moving camera," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2201–2210, 2010.

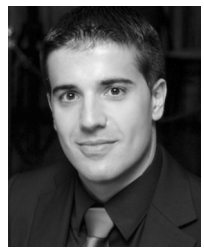
[7] D. Pundik and Y. Moses, "Video synchronization using temporal signals from epipolar lines," in *Proc. Comput. Vision ECCV 2010*, 2010, vol. 6313, pp. 15–28.

[8] A. Ravichandran and R. Vidal, "Video registration using dynamic textures," in *Proc. ECCV*, 2008, 2008, pp. 514–526.

[9] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, 2011.

[10] J. Serrat, F. Diego, F. Lumbreras, and J. M. Álvarez, "Synchronization of video sequences from free-moving cameras," in *Proc. 3rd Iberian Conf. Pattern Recognition and Image Anal. (IbPRIA 2007)*, 2007.

[11] M. Singh, I. Cheng, M. Mandal, and A. Basu, "Optimization of symmetric transfer error for sub-frame video synchronization," in *Proc. ECCV*.
[12] J. Yan, J. Yan, and M. Pollefeys, "Video synchronization via space-time interest point distribution," in *Proc. Advanced Concepts for Intell. Vision Syst.*, 2004.
[13] D. Wedge, D. Huynh, and P. Kovesi, "Using space-time interest points for video synchronization," in *Proc. IAPR Conf. Mach. Vision Applicat.*, 2007.
[14] C. Dai, Y. Zheng, and X. Li, "Accurate video alignment using phase correlation," *IEEE Signal Process. Lett.*, vol. 13, no. 12, pp. 737–740, 2006.
[15] C. Rao *et al.*, "View-invariant alignment and matching of video sequences," in *Proc. IEEE Int. Conf. Comput. Vision*, 2003, pp. 939–945. [Online]. Available: http://server.cs.ucf.edu/vision/projects/ViewInvariance/ViewInvariance.html.
[16] L. Wolf and A. Zomet, "Wide baseline matching between unsynchronized video sequences," *Int. J. Comput. Vision*, vol. 68, no. 1, pp. 43–52, 2006.
[17] P. A. Tresadern and I. D. Reid, "Video synchronization from human motion using rank constraints," *Comput. Vision Image Understand.*, vol. 113, no. 8, pp. 891–906, 2009.
[18] F. L. Padua, R. L. Carceroni, G. A. Santos, and K. N. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 304–320, 2010.
[19] T. Tuytelaars and L. VanGool, "Synchronizing video sequences," in *Proc. CVPR*, 2004, vol. 1, pp. 762–768.
[20] C. Lei and Y. Yang, "Trifocal tensor-based multiple video synchronization with subframe optimization," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2473–2480, 2006.
[21] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. New York, NY, USA: Springer, 2009.
[22] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*, J. Ponce, M. Hebert, C. Schmid, and A Zisserman, Eds. New York, NY, USA: Springer, 2006, vol. 4170, pp. 127–144.
[23] D. Nistr and H. Stewnius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognition*, 2006, pp. 2161–2168.
[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[25] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[26] M. Irani, "Multi-frame correspondence estimation using subspace constraints," *Int. J. Comput. Vision*, vol. 48, no. 3, pp. 173–194, 2002.
[27] R. Szeliski, Image alignment and stitching:A tutorial, Technical Report MSR-TR-2004-92, Microsoft Research, Tech. Rep. 4.
[28] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Pearson Education, 2003.
[29] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[30] T. A. Davis, "A column pre-ordering strategy for the unsymmetric-pattern multifrontal method," *ACM Trans. Math. Softw.*, vol. 30, no. 2, pp. 165–195, 2004.
[31] A. Vedaldi and B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms, 2008. [Online]. Available: http://www.vlfeat.org/.
[32] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[33] P. Kelly, N. O'Connor, and A. Smeaton, "A framework for evaluating stereo-based pedestrian detection techniques," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, pp. 1163–1167, Aug. 2008.
[34] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: a unifying framework," *Int. J. Comput. Vision*, vol. 56, no. 3, pp. 221–255, 2004.



Ferran Diego is a Postdoctoral researcher of Heidelberg Collaboratory for Image Processing at the University of Heidelberg in Heidelberg, Germany. He received the Higher Engineering in Telecommunications in 2005 from Politechnical University of Catalonia (UPC), the M.Sc. degrees in Language and Speech from Politechnical University of Catalonia and University of Edinburgh in 2005, the M.Sc. degrees in Computer Vision and Artificial Intelligence in 2007, and obtained his Ph.D. degree at Computer Vision Center and Universitat Autonoma de Barcelona in 2011. His main research interests include video alignment, low-level image processing, sensor fusion, machine learning, image retrieval, sparse and redundant representation modeling, structured sparsity and group and probabilistic graphical models for image analysis.

Joan Serrat received the Ph.D. degree in computer science from the Universitat Autonoma de Barcelona in 1990. He is associate professor at the computer science department of this university and also member of the Computer Vision Center. His current research interest is the application of probabilistic graphical models to computer vision problems like feature matching, tracking and video alignment. He has also been head of several machine vision projects for local industries and member of the IAPR Spanish chapter board.

Antonio M. López received the B.Sc. degree in computer science from the Universitat Politècnica de Catalunya in 1992, the M.Sc. degree in image processing and artificial intelligence from the Universitat Autònoma de Barcelona (UAB) in 1994, and the Ph.D. degree in 2000. Since 1992, he has been giving lectures in the Computer Science Department of the UAB, where he currently is an associate professor. In 1996, he participated in the foundation of the Computer Vision Center at the UAB, where he has held different institutional responsibilities, presently being responsible for the research group on advanced driver assistance systems by computer vision. He has been responsible for public and private projects, and is a coauthor of more than 50 papers, all in the field of computer vision.