Efficient video sequences alignment using unbiased bidirectional dynamic time warping

Cheng Lu a, Meghna Singh a, Irene Cheng b, Anup Basu b, Mrinal Mandal a,⇑

a Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4
b Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2V4

ABSTRACT

In this paper, we propose an efficient technique to synchronize video sequences of events that are acquired via uncalibrated cameras at unknown and dynamically varying temporal offsets. Unlike other existing techniques that just take unidirectional alignment into consideration, the proposed technique considers symmetric alignments and compute the optimal alignment. We also establish sub-frame accuracy video alignment. The advantages of our approach are validated by tests conducted on several pairs of real and synthetic sequences. We present qualitative and quantitative comparisons with other existing techniques. A unique application of this work in generating high-resolution 4D MRI data from multiple low-resolution MRI scans is described.

1. Introduction

Alignment of video sequences plays a crucial role in applications such as super-resolution imaging [1], 3D visualization [2], robust multi-view surveillance [3] and mosaicing [4]. Most video alignment techniques deal with video sequences of the same scene and hence assume that the temporal offset between the video sequences does not change over time, i.e., a simple temporal translation is assumed. However, for applications such as video search, video comparison and enhanced video generation, we need to align video sequences from two different (but similar) scenes. In such scenarios, the temporal offset between the video sequences is dynamically changing, and cannot be estimated by a translational offset. Consider the problem of generating high resolution visualization of MRI (magnetic resonance imaging) data from multiple low resolution videos. Due to the tradeoff between acquisition speed and image resolution in MRI, we can only acquire a few frames of functional events such as swallowing (4–7 fps). We can, however, acquire multiple swallowing videos that differ from each other spatially and temporally. The objective then is to compute a robust spatial and temporal alignment between the MRI videos in order to fuse them together for better visualization. Since the video sequences are of repeated swallows, the temporal offset between the video sequences does not remain constant, and cannot be estimated by a translational offset. In fact, the temporal offset varies dynamically with time and depends on the stage of the swallowing process.

Several techniques have been proposed to synchronize the video sequences with the same or similar scenes. Giese and Poggio [5] proposed a dynamic time warping based technique to align motions performed by different people between the feature trajectories. Perperidis et al. [6] proposed to use the spline for accurate alignment for locally warp cardiac MRI sequences. Rao et al. [7] proposed to use rank constraint as the distance measure in a dynamic time warping technique to align multiple sequences (we refer this technique as RCB technique in this paper). Their technique is limited to integer frame alignment of video sequences. However, with low acquisition rates, as in our case, sub-frame alignment is crucial. Singh et al. [8] proposed a video alignment technique based on minimization of the symmetric transfer error (STE) between two commutative alignments. The STE technique provides a good performance with sub-frame accuracy. But the accuracy can be improved further since the technique does not completely eliminate the biasing between two sequences.

The main contribution of this work lies in formulating the alignment problem as unbiased bidirectional dynamic time warping (UBDTW), in which the optimal alignment is computed in the combined symmetric accumulated distance matrix. UBDTW eliminates any bias that is introduced by choosing one sequence as the
reference sequence, to which target sequences are warped. In addition, the proposed technique allows us to compute sub-frame accurate and view-invariant alignment of video sequences that have a dynamically varying temporal offset between them and also helps us to reduce the occurrence of singularities\(^1\) in the alignment.

The rest of this paper is organized as follows. Section 2 presents related work in video alignment, and justifies the necessity for this work. In Section 3, we propose the UBDTW technique. Experimental setup and present comparative results are shown in Section 4. Application of our work in 4D MRI visualization is presented in Section 5, which is followed by a summary and conclusions in Section 6.

2. Review of previous work

Past literature in video alignment and temporal registration can be broadly classified into two categories: video sequences of the same scene or video sequences of similar scenes; differing primarily on the assumptions made with respect to the temporal offset between sequences. It can be seen from Table 1, that our work addresses all three scenarios (view-invariance, dynamic time shifts and sub-frame accuracy), while previous works have only addressed a subset of these scenarios.

2.1. Video alignment of same scene

In synchronizing videos of the same scene, the temporal offset is considered to be an affine transform \([1]\), such as \(t' = s \cdot t + \Delta t\), where \(s\) is the ratio of frame rates and \(\Delta t\) is a fixed translational offset. Dai et al. [9] use 3D phase correlation between video sequences, whereas Tuytelaars and VanGool [10] compute alignment by checking the rigidity of a set of five (or more) points. Tresadern and Reid [11] also follow a similar approach of computing a rank-constraint based rigidity measure between four non-rigidly moving feature points. Caspi and Irani [1] recover the spatial and temporal relation between two sequences by minimizing the SSD error over extracted trajectories that are visible in both sequences. Carceroni et al. [12] extend [1] to align sequences based on scene points that need to be visible only in two consecutive frames. Singh et al. [2] also build on [1] to develop parameterized event models of discrete trajectories and align the event models based on an SSD measure for sub-frame alignment. However, they do not address the problem of view-invariance and dynamic temporal offsets.

2.2. Video alignment of different scenes

When aligning video sequences of different scenes, albeit sequences correlated via motion, one has to factor in the dynamic temporal scale of activities in the video sequences. Giese and Poggio [5] approach alignment of activities of different people by computing a dynamic time warp between the feature trajectories. They did not address the problem when the activity sequences are from varying viewpoints and their approach is a one-to-one frame correspondence. Perperidis et al. [6] attempt to locally warp cardiac MRI sequences, by extending Caspi’s work [1] to incorporate spline based local alignment. Though their approach does lead to good alignment of time-varying sequences, it has two main drawbacks: (1) the computation space for spline based registration is quite large and the authors need to compute points of inflexion in the cardiac volume change; and (2) the alignment is still a one-to-one frame correspondence and not sub-frame accurate. The RCB technique proposed by Rao et al. [7] will not work for planar motion activity since the fundamental matrix is singular in this case. Singh et al. [8] proposed an STE-based video alignment technique. In a two videos framework, the STE is calculated in two steps. In the first step, one video is considered as the reference, the second video is aligned using a DTW technique, and a warping function is obtained. In the second step, the second video is considered as the reference, the first video is aligned using the DTW technique, and a warping function is obtained. The STE is calculated based on the difference between the two warping functions. It is expected that two videos temporally aligned well would have small STE. The STE technique provides better performance over RCB techniques. However, the STE technique selects one of the two symmetric warps as the final alignment which may not be accurate. In addition, due to its iterative STE calculation, the computational complexity of this technique is high compared to existing techniques.

3. Proposed technique

Suppose that two cameras C1 and C2 view two independent scenes of similar activities, as shown in Fig. 1. Camera 1 views 3D scene \(X(x_1, y_1, z_1, t_1)\) in View 1 and acquires video \(I_1(x_1, y_1, t_1)\). Camera 2 views another 3D scene \(X(x_2, y_2, z_2, t_2)\) in View 2 and acquires video \(I_2(x_2, y_2, t_2)\). Note that the motions

![Fig. 1. Illustration of two different scenes acquired using two distinct cameras. The projections of scenes onto the reciprocal cameras are also shown.](image-url)
in these two scenes are similar but have dynamic time shift. In this paper we assume that the scene is planar and the homography matrix \( H \) is used to represent the spatial relationship between these two views. These assumptions are not limiting since the proposed technique can be adapted to account for epipolar geometry for non-planar scenes and robust correspondence techniques can be used for wide-baseline cameras [16]. Additionally, we assume that a single feature trajectory of interest is available to us\(^2\) such that the beginning and end points of the activity are marked in the trajectory, similar to the assumption made in [7,8].

The schematic of the proposed alignment technique is shown in Fig. 2. The technique consists of five modules which are explained in the following sections.

3.1. Feature trajectories extraction

Note that for the sake of correlating two video sequences and representing the motions between the video sequences, features are extracted and tracked from two video sequences. A robust view-invariance tracker is used to generate feature trajectory \( F_1(x_1, y_1, t_1) \) and \( F_2(x_2, y_2, t_2) \) from video \( I_1(x_1, y_1, t_1) \) and \( I_2(x_2, y_2, t_2) \), respectively. The feature trajectories are illustrated in the Fig. 1. On their own these feature trajectories are discrete representations of the event in the scene. However, to align video sequences to sub-frame accuracy, we need to interpolate between the discrete representations. Most techniques used linear interpolation between the discrete points; however, we chose the technique proposed by Singh et al.[2] to generate continuous models from discrete points.

3.2. Event models calculation

We derive event models from the discrete features as proposed by Singh et al. [2]. An event model is a global representation of the activity based on all the extracted feature points and can be tailored to address varying sub-frame alignment needs. Using event models instead of discrete feature points allows the proposed technique to adapt to increasing complexity of event dynamics, which is demonstrated in our test cases.

Given a discrete trajectory \( F(x, y, t_0) \), we compute linear model parameters \( \beta \) such that a continuous estimate \( \hat{F}(x, y, t) \) can be derived in the least square sense as shown in Eqs. (1),(2), where the symbolizes estimation and \( e \) represents the error term.

\[
F = F\beta + e \tag{1}
\]

\[
\hat{F} = F\hat{\beta} \tag{2}
\]

In a weighted least square sense, with a weight matrix \( W \), the best linear estimate of \( \beta \) is given by:

\[
\hat{\beta} = (F^TWF)^{-1}F^TWF \tag{3}
\]

The weights are computed iteratively and the effect of outliers on the event model estimation is minimized, for details see [2]. Using event models has two additional advantages. Firstly, trajectory information for non-integer frames are now no longer based on linearly interpolated values from adjacent frames, and hence videos with low acquisition rates can also be accurately synchronized to non-integer values, as shown in [2]. Secondly, computation of the event models inherently smoothes the trajectory to remove outliers due to noise or poor tracking. The continuous feature trajectories are represented as \( F_1 \) and \( F_2 \) and the resolutions are higher than the discrete trajectories \( F_1 \) and \( F_2 \), respectively.

\(^2\) In typical videos, multiple object trajectories will be generated and an additional task of the synchronization technique will be used to find corresponding feature trajectories in the multiple video sequences. This is an open problem in vision research and one that we will not address in this paper.
where \( w \) is the weight (normally, \( w = 50 \)), \( \partial F \) and \( \partial^2 F \) are the first and second derivatives of \( F \), and are computed numerically as follows.

\[
\partial F = \frac{F(i) - F(i-1) + |F(i+1) - F(i-1)|}{2}
\]

\[
\partial^2 F = \frac{\partial F(i) - \partial F(i-1) + |\partial F(i+1) - \partial F(i-1)|}{2}
\]

The purpose of the regularization term is twofold: (i) it allows us to factor in a smoothness constraint on the warping and (ii) it also reduces the occurrence of singularities in the alignment.

In order to find the optimal warp, a \( L \times L \) accumulated distance matrix is created. The value of the element in the accumulated distance matrix is:

\[
D(i,j) = \text{Dist}[F(i), F^p(j)] + \min(\phi)
\]

where

\[
\phi = [D(i - 1, j), D(i - 1, j - 1), D(i, j - 1)]
\]

A greedy search technique is then employed in the accumulated distance matrix to find the optimal warp \( W \) such that the \( \text{Dist}(W) \) is minimum.

Using RDTW described above, we are able to obtain the symmetric warps (alignments) \( W_{1,2p} \) and \( W_{1p,2} \) for \( (F_1, F^p_2) \) and \( (F^p_1, F_2) \), respectively. However, these two warps do have bias: \( W_{1,2p} \) is biased toward \( F_1 \), whereas \( W_{1p,2} \) is biased toward \( F_2 \).

### 3.5. Computation of optimal alignment

Note that we calculated symmetric warps \( W_{1,2p}, W_{1p,2} \), and corresponding distance matrices \( D_{1,2p} \) and \( D_{1p,2} \) in the last step. However, the warps do have bias. In order to eliminate the bias, we first
combine distance matrices $D_{1p2}$ and $D_{1p2}$ to make a new distance matrix, named combined symmetric accumulated distance matrix, $D_c$ as follows:

$$D_c(i,j) = D_{1p2}(i,j) + D_{1p2}(i,j)$$

Once $D_c$ is obtained, global constraint based on $W_{1p2}$ and $W_{1p2}$ are added onto this matrix. Denote $W_c$ as the warps under the global constraint and the constraint are formulated as follows:

$$\min(W_{1,2}, W_{2,1}) \leq W_c \leq \max(W_{1,2}, W_{2,1})$$

$$0 < L_c < \min(L_{1,2}, L_{2,1})$$

where $L_c, L_{1,2},$ and $L_{2,1}$ are the length of the warp $W_c$, $W_{1p2}$, and $W_{1p2}$, respectively. Finally, the warp $W_c$ which satisfies the following equation is chosen as the final optimal warp (alignment) for video sequences $I_1$ and $I_2$:

$$W_{copt} = \arg \min_{W} \text{Dist}(W_c)$$

**Table 3**

<table>
<thead>
<tr>
<th>Noise</th>
<th>RCB</th>
<th>STE</th>
<th>Proposed technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.14</td>
<td>3.12</td>
<td>1.97</td>
</tr>
<tr>
<td>0.0001</td>
<td>7.53</td>
<td>3.17</td>
<td>2.08</td>
</tr>
<tr>
<td>0.001</td>
<td>8.02</td>
<td>3.29</td>
<td>2.36</td>
</tr>
<tr>
<td>0.01</td>
<td>9.32</td>
<td>3.37</td>
<td>2.74</td>
</tr>
<tr>
<td>0.1</td>
<td>12.45</td>
<td>3.45</td>
<td>2.91</td>
</tr>
</tbody>
</table>

The gray area represents the search area constrained by $W_{1p2}$ and $W_{1p2}$. The warp $W_{copt}$ which is inside the global constraint is considered as the final unbiased warp.

A pseudocode of the proposed technique is given below:

**Pseudocode of the proposed technique**

1. Extract feature trajectories $F_1$ and $F_2$
2. Compute event models $F_1$ and $F_2$
3. Compute spatial relationship, homography $H$ is used here
4. Project event models to derive projections $F_1^p$ and $F_2^p$
5. Set $w = 50$
6. Compute regularized warp $W_{1p2}$ from $(F_1, F_2^p)$
7. Compute regularized warp $W_{1p2}$ from $(F_1^p, F_2)$
8. Compute symmetric accumulated distance matrix $D_c$
9. Compute optimal warp $W_{copt}$ under global constraint
10. Report $W_{copt}$ as the final alignment

**4. Experiment and comparative analysis**

We tested our method on both synthetic and real image sequences. We also implemented the RCB and STE technique that deals with aligning videos of similar events. While the RCB technique cannot compute sequence alignment to sub-frame accuracy, we still compare our method with it for integer alignment. Our test cases and results are presented next. All experiments were run on a 2.4 GHz Intel Core II Duo CPU with 3GB RAM using MATLAB 7.04. Excluding the preprocessing time, the execution time of the RCB, STE and the proposed technique are summarized in Table 2. The proposed technique took 0.22 s to compute the optimal alignment.

![Performance evaluation using noisy trajectory](image_url)

**Fig. 6.** Performance evaluation using noisy trajectory. (a) The noisy trajectory (in right) is generated by adding noise with $\sigma^2 = 0.1$ to the clean trajectory (in left). (b) The alignment performance obtained using the RCB, STE, and the proposed technique.
between two synthetic sequences of lengths 100 and 160, whereas RCB technique took 0.20 s and the STE technique took 1.16 s. The processing time for real sequences of length 54 and 81 was 0.21 s, 0.19, and 1.03 s for the proposed, RCB, and STE technique, respectively. In other words, the complexity of the proposed technique is at a similar level to that of the RCB technique, and is much smaller compared to the STE technique.

4.1. Synthetic data evaluation

In synthetic data evaluation, we generate planar trajectories, 100 frames long, using a pseudo-random number generator. These trajectories are then projected onto two image planes using user defined camera projection matrices. The camera matrices are designed so that the acquisition emulates a homography, and are used only for generation purpose and not thereafter. A 60 frames long time warp is then applied to a section of one of the trajectory projections. The RCB, STE and the proposed techniques are then applied to the synthetic trajectories to compute the alignment between them. This process is repeated on 50 different synthetic trajectories. Figs. 4 and 5 show alignment results with simple and complex synthetic trajectories. The performance of the proposed technique and RCB technique are comparable in the case of simple trajectories. However, as the complexity of the trajectory increased, the proposed technique outperformed the RCB technique.

We also evaluated the effect of noisy trajectories on the RCB, STE, and the proposed technique. Normally distributed and zero mean noise with various values of variance ($\sigma^2$) was added to the synthetic feature trajectories. The results of alignment of smooth and noisy trajectories with the RCB, STE, and the proposed techniques are shown in Table 3, where the mean absolute differences between the actual and computed frame correspondence is reported as the alignment error. It should be noted that the STE and the proposed technique are marginally affected by the addition of noise. However, the performance of the RCB method degrades significantly. An example of noisy trajectory is shown in Fig. 6.

4.2. Real data evaluation

For tests on real video sequences, we used two video sequences recording the activity of lifting a coffee cup by different people.
from different views (we refer these two video as LCC video). The first video is 54 frames long and the second video is 81 frames long captured from another viewpoint. We tracked the coffee cup which can represent the activity in the video sequences to generate feature trajectories. The feature trajectory of video LCC1 and LCC2 are shown in Figs. 7(a) and (b), respectively. The RCB, STE, and the proposed techniques were then applied to the real video data. Note that the ground truth information is not available. However, we can use visual judgment to check if the alignment is correct. Fig. 7(c) shows the results computed using the RCB, STE, and the proposed technique. The rectangle in Fig. 7(c) indicates the frames range for the frames as shown in Fig. 8(a)–(c). Fig. 8(a)–(c) shows some representative alignment results computed using the RCB, STE, and the proposed technique, respectively. We choose the 3rd, 6th, 9th, 13th, 16th frames from video LCC2 and their corresponding frames in video LCC1. Note that if the interested feature, i.e. the coffee cup, is at the same position in two frames, we marked it as matched, otherwise, mismatched. It is observed that there are two pairs of mismatched frames for the RCB technique (see Fig. 8(a)) whereas there is only one pair of mismatched frames for the STE technique (see Fig. 8(b)). This indicates that the existing techniques introduce erroneous alignments, whereas the alignment computed using the proposed technique (Fig. 8(c)) provides a more accurate temporal matching.
4.3. Unidirectional vs. bidirectional alignment

In the previous section we compared the proposed unbiased bidirectional technique against the unidirectional alignment technique proposed in [7]. The proposed technique and the RCB technique differ not only in the sub-frame alignment capability and reduction in singularities, but also in the capability of eliminating the bias. In order to highlight the advantage of the capability of eliminating the bias, we validate it against unidirectional alignment. The warps computed using both the bidirectional and unidirectional technique for two trajectories are shown in Fig. 9. Note that the two unidirectional warps are shown as black dot-dash line; whereas the blue dash line is the bidirectional warp computed using the proposed technique.

Note that the STE technique and the proposed technique both compute the bidirectional alignments. The STE technique computes the symmetric error between two unidirectional warps by turning the regularization factor iteratively until it reaches the minimum and then selects one of the warps as the final warp.

In the example shown in Fig. 9, it is clear that these two unidirectional warps \( W_{1,2p} \) and \( W_{1p,2} \) are close to the ground truth, but not accurate enough due to the bias towards the reference sequence. In the STE technique, either \( W_{1,2p} \) or \( W_{1p,2} \) will be selected as the final output. The proposed technique computes the warp using Eq. (12), and provides more accurate results.

5. Application: 4D MRI registration

One of our motivations behind this work is to build a 4D (volume + time) representation of functional events in the body using 2D planar acquisitions, specifically swallowing disorders. In our experiments, a subject lies prone inside an MRI scanner and is fed small measured quantities of water (bolus) via a system of tubing (water is displayed as white in the MRI images because of the high Hydrogen content in it). For each swallow, a time series of 2D images are acquired on a fixed plane. The acquisition plane is then changed and another series of 2D images are acquired. We acquire three such video sequences corresponding to left, right and center MRI slice planes, as illustrated in Fig. 10(a). The MRI video sequences are subjected to dynamic temporal offsets in the motion of the bolus. Also, since the acquisitions are at very low frame rates, limited by the technology to 4–7 fps, it is crucial to align the sequences to sub-frame accuracy. The proposed technique shows promising results in aligning the MRI sequences to generate a 4D representation. Implementation details of this application are discussed next.

The trailing and leading edges of the bolus are extracted from the MRI sequences using standard background separation techniques. The center of the trailing bolus is extracted using horizontal and vertical profiles, and is used to generate feature trajectories in the three sequences. After suitable event models have been computed for the trajectories, the proposed technique is applied to the video sequences to compute the alignment. The result of alignments with the proposed technique is shown in Fig. 11. Fig. 11 shows a few frames from the alignment computed between the center and left MRI slices that demonstrate sub-frame alignment. Frame 5 of the left MRI sequence is mapped to frame 5.5 of the center MRI sequence. Once the alignment has been computed, 4D visualization of the MRI data is carried out with a 4D model being shown in Fig. 10(b).

6. Conclusions

We proposed and successfully evaluated an efficient technique to align video sequences that are related by varying temporal offsets. Our formulation of alignment as the unbiased bidirectional dynamic time warping resulted in alignment that was not biased.
by the choice of the reference sequence. The regularized dynamic time warping successfully reduced the occurrence of false singularities and resulted in sub-frame accuracy synchronization. Comparative analysis with a rank-constraint based technique and symmetric transfer error based technique demonstrated a significant improvement in video alignment with our proposed technique. The proposed technique has a lower or similar computational complexity compared to existing techniques. An application of the proposed method in 4D MRI visualization was also presented. Although the proposed technique has been presented for a two-camera set-up, and the performance has been evaluated based on two-videos, the proposed technique can easily be extended to align more than two videos.

References