Efficient temporal alignment of video sequences using unbiased bidirectional dynamic time warping

Cheng Lu and Mrinal Mandal
University of Alberta, Department of Electrical and Computer Engineering, Edmonton, Alberta, T6G 2V4, Canada

Abstract. *We propose an efficient technique for temporally aligning video sequences of similar activities. The proposed technique is able to synchronize view-variance videos from different scenes performing similar 3-D activities. Unlike existing techniques that just consider unidirectional alignment, the proposed technique considers symmetric temporal alignment and computes the optimal alignment by eliminating any view-based bias. The advantages of our technique are validated by experiments conducted on synthetic and real video data. The experimental results show that the proposed technique out-performs existing techniques in several test video sequences.*
© 2010 SPIE and IS&T. [DOI: 10.1117/1.3488415]

1 Introduction

Temporal alignment of video sequences is important in applications such as superresolution imaging,[1] robust multiview surveillance,[2] and mosaicking. In some applications, it is required to align video sequences from two similar scenes, where analogous motions have different trajectories through the video sequence. Figure 1 illustrates two similar motions occurring in related 3-D planar scenes with respect to time. Camera 1 views 3-D scene $X(X_1, Y_1, Z_1, t_1)$ in view 1(v_1) and acquires video $I_1(x_1, y_1, t_1)$. Camera 2 views another 3-D scene$X(X_2, Y_2, Z_2, t_1)$in view 2 (v_2) and acquires video $I_2(x_2, y_2, t_2)$. Note that the motions in these two scenes are similar but have dynamic time shift. The homography matrix **H** is typically used to represent the spatial relationship between these two views.

A typical schematic for temporal alignment is shown in Fig. 2. Note that for the sake of correlating two videos and representing the motions, features are extracted and tracked separately in each video. Robust view-invariance tracker methods are used to generate feature trajectories $\mathcal{F}_1(x_1, y_1, t_1)$ and $\mathcal{F}_2(x_2, y_2, t_2)$ from video I_1 and I_2, respectively.

Existing techniques vary on how to compute the temporal alignments. Giese and Poggio[3] computed the temporal alignment of activities of different people using dynamic time warping (DTW) between the feature trajectories, but limited their technique to a fixed viewpoint. Rao et al.[4] used a rank-constraint-based technique (RCB) in DTW to calculate the synchronization. Such techniques only consider unidirectional alignment,[3,4] i.e., they project the trajectory from one scene to the other, which designates one view as the reference for computing the temporal alignment. Such techniques introduce the bias toward the reference trajectory, i.e., due to the noise and imperfection of the obtained reference trajectory, such a technique will produce erroneous alignment. Therefore, for the sake of minimizing the bias, one should consider computing the alignment in a symmetric way. Singh et al.[5] formulated a symmetric transfer error (STE) as a functional of regularized temporal warp. The technique determines the time warp that has the smallest STE. It then chooses one of the symmetric warps as the final temporal alignment. The STE technique provides better results than unidirectional alignment schemes. The accuracy of the temporal alignment can be improved further, since the STE technique does not really eliminate the reference-view bias between two sequences.

In this work, we propose an unbiased bidirectional dynamic time warping (UBDTW) technique that can remove biasing and provide more accurate results.

2 Proposed Technique

The schematic of the proposed temporal alignment technique is shown in Fig. 3. The technique consists of three steps which are explained in the following sections.

2.1 *Bidirectional Projections*

Since feature trajectories represent the activities in the video sequences, we compute the projections of the feature trajectories $\mathcal{F}_1$ from scene 1 to 2 and $\mathcal{F}_2$ from scene 2 to 1 using Eq. (1) as follows:

$$\mathcal{F}_2^p(x_1', y_1', t_1) = H_{1\to 2} \cdot \mathcal{F}_1(x_1, y_1, t_1),$$

$$\mathcal{F}_1^p(x_2', y_2', t_2) = H_{2\to 1} \cdot \mathcal{F}_2(x_2, y_2, t_2), \quad (1)$$

where $H_{1\to 2}$ and $H_{2\to 1}$ are the homographies from scene 1 to 2 and scene 2 to 1, respectively. Homographies are independent of the scene structure and can be computed

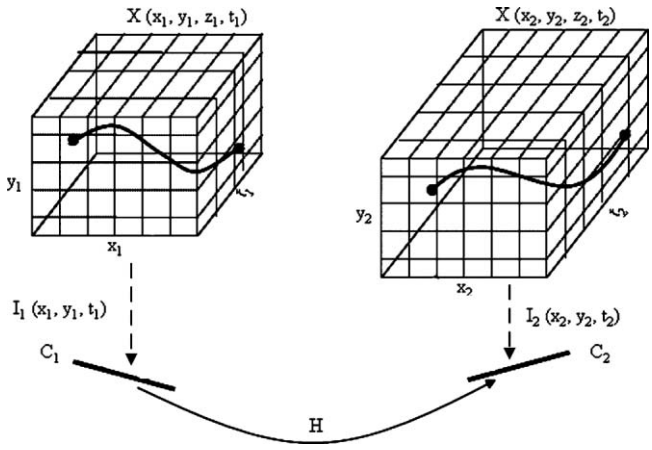

Fig. 1 Illustration of two distinct scenes acquired using two distinct cameras.

Paper 10040LR received Mar. 12, 2010; revised manuscript received Aug. 2, 2010; accepted for publication Aug. 13, 2010; published online Dec. 21, 2010

1017-9909/2010/19(4)/040501/3/$25.00 © 2010 SPIE and IS&T.

Downloaded From: http://electronicimaging.spiedigitallibrary.org/ on 11/01/2012 Terms of Use: http://spiedl.org/terms

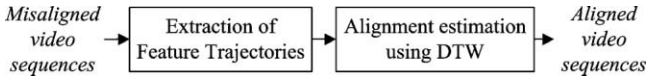

Fig. 2 A typical schematic of existing techniques.

from $X(x_1, y_1, z_1, t_1) \cap X(x_2, y_2, z_2, t_2)$ using the direct linear transform (DLT) algorithm.[6]

2.2 *Computation of Symmetric Warps*

Once we obtain two pairs of feature trajectories, $(\mathcal{F}_1, \mathcal{F}_2^p)$ and $(\mathcal{F}_1^p, \mathcal{F}_2)$, we compute the symmetric warps $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$ using regularized DTW. We construct the warp $\mathcal{W}$ as follows:

$$\mathcal{W} = w_1, w_2, \ldots, w_L \quad \max(\mathcal{L}_1, \mathcal{L}_2) \leq L < \mathcal{L}_1 + \mathcal{L}_2, \tag{2}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are the length of trajectories $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. The L'th element of the warp $\mathcal{W}$ is $w_L = (i, j)$, where i and j are the time indices of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. The optimal warp is the minimum distance warp, where the distance of a warp is defined as follows:

$$\text{dist}(\mathcal{W}) = \sum_{k=1}^{L} \text{dist}[\mathcal{F}(i_k), \mathcal{F}^p(j_k)], \tag{3}$$

where $\text{dist}[\mathcal{F}(i_k), \mathcal{F}^p(j_k)]$ is the distance between the two values of the given time indices (i, j) in the k'th element of the warp. We propose a regularized distance metric function as follows:

$$\text{dist}[\mathcal{F}(i), \mathcal{F}^p(j)] = ||\mathcal{F}(i) - \mathcal{F}^p(j)||^2 + w \text{ reg}, \tag{4}$$

$$\text{reg} = ||\partial\mathcal{F}(i) - \partial\mathcal{F}^p(j)||^2 + ||\partial^2\mathcal{F}(i) - \partial^2\mathcal{F}^p(j)||^2, \tag{5}$$

where $\partial\mathcal{F}$ and $\partial^2\mathcal{F}$ are the first and second derivatives of $\mathcal{F}$. The regularization term can be considered a smoothness penalty, where w is the weight (normally, $w = 25$).

To find the optimal warp, an accumulated distance matrix is created. The value of the element in the accumulated distance matrix is:

$$\mathcal{D}(i, j) = \text{dist}[\mathcal{F}(i), \mathcal{F}^p(j)] + \min(\phi), \tag{6}$$

$$\phi = [\mathcal{D}(i-1, j), \mathcal{D}(i-1, j-1), \mathcal{D}(i, j-1)]. \tag{7}$$

A greedy search technique is employed to find the optimal warp $\mathcal{W}$, such that $\text{dist}(\mathcal{W})$ is minimum. We can now obtain the symmetric warps $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$ for $(\mathcal{F}_1, \mathcal{F}_2^p)$ and $(\mathcal{F}_1^p, \mathcal{F}_2)$, respectively.

2.3 *Optimal Warp Calculation*

Note that we calculated symmetric warps $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$, and corresponding distance matrixes $\mathcal{D}_{1,2p}$ and $\mathcal{D}_{1p,2}$ in the last step. However, the warps still have bias ($\mathcal{W}_{1,2p}$ biased toward $\mathcal{F}_1$ while $\mathcal{W}_{1p,2}$ is biased toward $\mathcal{F}_2$). To minimize the effect of biasing on alignment, we first combine

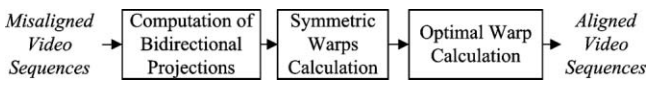

Fig. 3 The schematic of UBDTW.

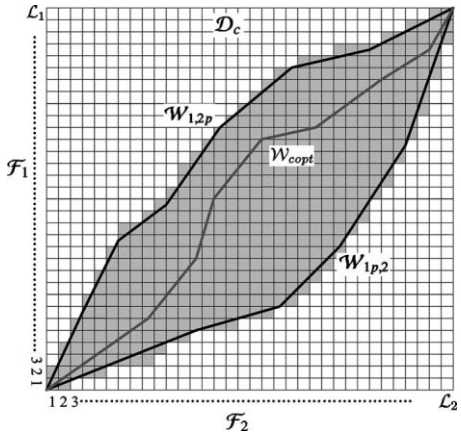

Fig. 4 An intuitive illustration of the UBDTW.

$\mathcal{D}_{1,2p}$ and $\mathcal{D}_{1p,2}$ to make a new distance matrix $\mathcal{D}_c$ as follows:

$$\mathcal{D}_c = \mathcal{D}_{1,2p} + \mathcal{D}_{1p,2}. \tag{8}$$

Once $\mathcal{D}_c$ is obtained, global constraint based on $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$ is added into this matrix. Denote W_c as the warps under the global constraint, which is restricted by the symmetric warps, as follows:

$$\min(\mathcal{W}_{1,2}, \mathcal{W}_{2,1}) \leq \mathcal{W}_c \leq \max(\mathcal{W}_{1,2}, \mathcal{W}_{2,1}). \tag{9}$$

Finally, warp $\mathcal{W}_c$, which satisfies the following equation, is chosen as the final warp.

$$\mathcal{W}_{\text{copt}} = \arg_{\mathcal{W}_c} \min[\text{dist}(\mathcal{W}_c)]. \tag{10}$$

Figure 4 shows an intuitive explanation for the proposed optimal warp calculation. The grid represents the distance matrix $\mathcal{D}_c$. The horizontal and vertical axes represent the index of trajectories $\mathcal{F}_2$ and $\mathcal{F}_1$, respectively. The two bold lines represent the symmetric warps $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$. The gray area represents the search area constrained by $\mathcal{W}_{1,2p}$ and $\mathcal{W}_{1p,2}$. The warp $\mathcal{W}_{\text{copt}}$ inside the global constraint is considered as the final unbiased warp.

3 Experiments and Comparative Analysis

We evaluated our technique using both synthetic and real videos and compared it with RCB[4] and STE techniques.[5]

3.1 *Synthetic Data Evaluation*

In the synthetic data evaluation, we generate planar trajectories 100 frames long using a pseudorandom number generator. These trajectories are then projected onto two image planes using user-defined camera projection matrices. A 60-frames-long time warp is then applied to a section of one of the trajectory projections. The temporal alignment techniques are then applied to the synthetic trajectories. The test was repeated on 100 different synthetic trajectories and

Downloaded From: http://electronicimaging.spiedigitallibrary.org/ on 11/01/2012 Terms of Use: http://spiedl.org/terms

Table 1 Performance improvement of the proposed technique over existing techniques

	Origin	RCB	STE	Proposed
DynDate	25.50	7.33(71.25%)	3.11(87.80%)	**1.98(92.24%)**
SynData with noise	25.50	17.18(32.63%)	3.45(86.47%)	**2.97(88.35%)**

100 similar trajectories with noise added. The added noise was a normally distributed random variate, with zero mean and variance $\sigma^2 = 0.1$. The mean absolute error between the warp obtained by different techniques and the ground truth is computed as the evaluation metric. The results are shown in Table 1. The percentage in the parentheses represents the improvement obtained by an alignment technique with respect to the original error. Figure 5 shows a synchronization result with a synthetic trajectory. The performance of the RCB, STE and the proposed techniques are compared. It is clear that the proposed technique outperformed the other techniques.

3.2 *Real Data Evaluation*

For the real video test, we use two videos (54 frames and 81 frames long, respectively) capturing the activity of lifting a coffee cup by different people. We tracked the coffee cup

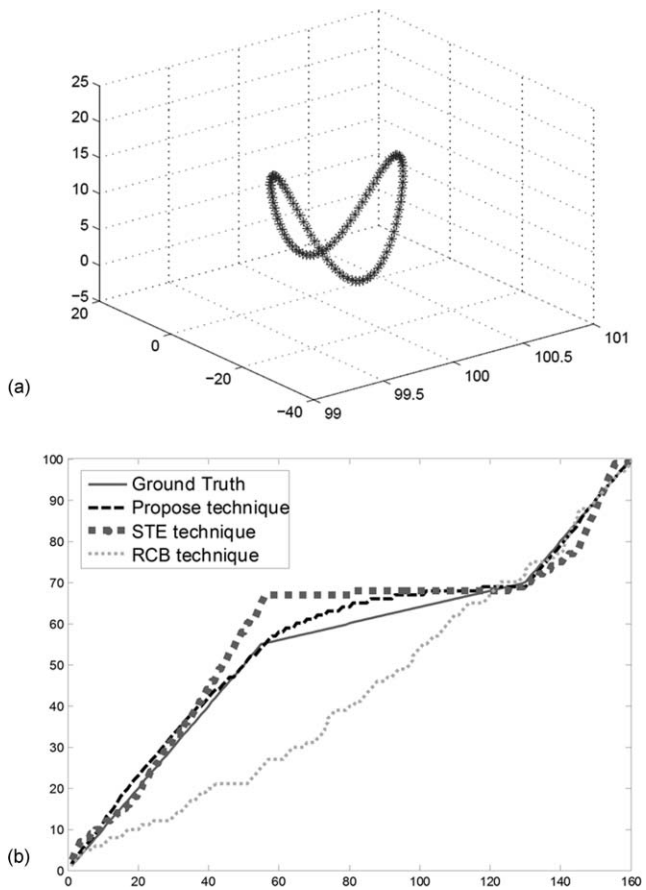

Fig. 5 (a) Synthetic trajectory; (b) result of temporal alignment for synthetic trajectories using RCB, STE and the proposed technique.

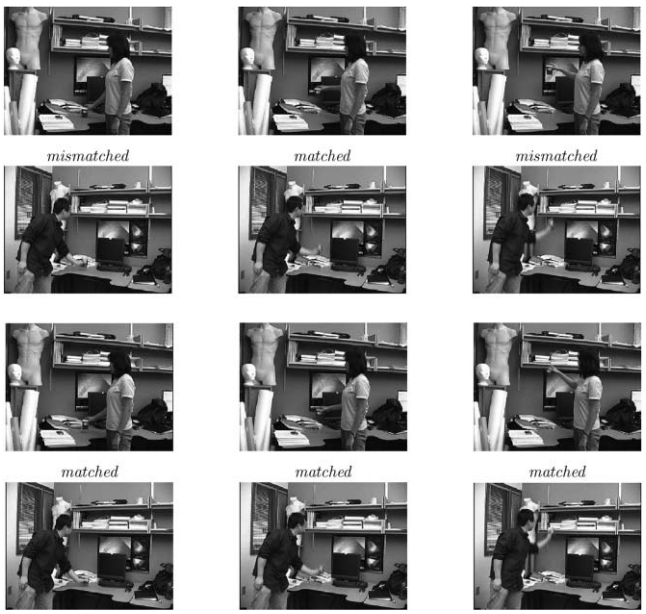

Fig. 6 Comparisons of temporal alignment results on real data using STE technique (shown in the first and second rows) and the proposed technique (shown in the third and fourth rows).

that can represent the activity in a video to generate feature trajectories. Since ground-truth information is not available, we used visual judgement to assess whether the alignment was correct or not.

Figure 6 shows some representative aligned frames in the 4th, 8th, and 12th elements of the alignment warp computed using the STE and the proposed technique. Note that if the coffee cup is at the same position in two frames, we marked it as "matched," otherwise, "mismatched." In the results obtained using the STE technique, only one pair of frames is matched, indicating that such technique can often result in erroneous alignments. The performance of the proposed technique is shown in the last two rows. It is observed that all the alignments are correct.

4 Conclusions

An efficient technique is proposed to synchronize video sequences captured from planar scenes and related by varying temporal offsets. The proposed UBDTW technique is able to remove the biasing and lead to accurate temporal alignment. In the future, we would like to extend this work to more general scenes.

References

1. Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(11), 1409–1424 (2002).
2. L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 758–767 (2000).
3. M. A. Giese and T. Poggio, "Morphable models for the analysis and synthesis of complex motion patterns," *Int. J. Comput. Vis.* **38**(1), 59–73 (2000).
4. C. Rao, A. Gritai, M. Shah, and T. F. S. Mahmood, "View-invariant alignment and matching of video sequences," *Proc. ICCV03*, pp. 939–945 (2003).
5. M. Singh, I. Cheng, M. Mandal, and A. Basu, "Optimization of symmetric transfer error for sub-frame video synchronization," *Proc. ECCV03*, 554–567 (2008).
6. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press, Cambridge, UK (2004).

Downloaded From: http://electronicimaging.spiedigitallibrary.org/ on 11/01/2012 Terms of Use: http://spiedl.org/terms