Multi-view Synchronization of Human Actions and Dynamic Scenes

Emilie Dexter[1]

Patrick Pérez

Ivan Laptev

INRIA, Vista Group
Centre Rennes-Bretagne Atlantique
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
http://www.irisa.fr/vista/

Abstract

This paper deals with the temporal synchronization of image sequences. Two instances of this problem are considered: (a) synchronization of human actions and (b) synchronization of dynamic scenes with view changes. To address both tasks and to reliably handle large view variations, we use self-similarity matrices which remain stable across views. We propose time-adaptive descriptors that capture the structure of these matrices while being invariant to the impact of time warps between views. Synchronizing two sequences is then performed by aligning their temporal descriptors using the Dynamic Time Warping algorithm. We present quantitative comparison results between time-fixed and time-adaptive descriptors for image sequences with different frame rates. We also illustrate the performance of the approach on several challenging videos with large view variations, drastic independent camera motions and within-class variability of human actions.

1 Introduction

When temporal alignment is unknown, synchronizing image sequences is a necessary and critical task for applications such as novel view synthesis, 3*D* reconstruction or analysis of dynamic scenes. It can be also useful for comparing sequences with dynamic contents that are similar up to speed variations. This is of interest for generic action analysis/recognition, or for more specific action analysis tasks like the comparison of different athletes' techniques in sport videos. The major difficulty lies in the drastic differences of visual appearance that the two sequences can exhibit. These changes can be due to different viewpoints, camera motions or even varying appearances of the moving objects. In addition to the inter-view variability, inter-scene variability is another source of concern when trying to synchronize sequences from two similar though not identical dynamic scenes.

In this work, we address two instances of this challenging synchronization problem in a common view-independent framework. The first one is **Action Synchronization**, i.e. synchronize sequences of different performances of an action under view changes. The second one is **Video Synchronization**, i.e. synchronize sequences of a same dynamic event seen from different views. In both cases, temporal alignment or synchronization consists in matching frames of the first sequence with frames in the second one.

BMVC 2009 doi:10.5244/C.23.122

1.1 Previous works

Video synchronization has mostly been addressed under assumptions of stationary cameras and linear time transformation. Some works jointly estimate the space-time transformations between two image sequences [4, 12, 16] while others focus only on temporal alignment [3, 11, 17, 18]. In the majority of methods, spatial correspondences between views are exploited either to estimate the fundamental matrix [4] or to derive rank constraints on observation matrices [18]. A few other methods, however, use image-based temporal features without correspondences as in [17], where authors investigate a temporal descriptor of image sequence based on co-occurences of appearance changes. Synchronization of image sequences from moving cameras is dealt by some authors but, to the best of our knowledge, none of them addresses automatic synchronization. For example in [15], authors choose manually the 5 independently moving points that must be tracked successfully along both sequences.

All approaches above deal with the synchronization of two sequences of the same dynamic scene. Two methods of video synchronization [11, 17] have also been proposed in order to temporally align a same action performed by different persons. The first one evaluates temporal alignment by using dynamic time warping and rank constraints on observation matrices and relies on spatial correspondences between image sequences which are hard to obtain in practice. The second approach estimates the space-time transformation based on maximizing local space-time correlations for image sequences captured by stationary cameras.

1.2 Our approach

In this paper, we propose an approach to automatically synchronize either human actions or videos of the same dynamic event, recorded from substantially different viewpoints. Our method combines a temporal description of image sequences, which does not require point correspondences between views, with Dynamic Time Warping (DTW). This procedure allows us to deal with arbitrary non-linear time warps (up to monotonicity constraint), which is especially important for action synchronization. We only assume, for the time being, that the two sequences at hand show two viewpoints of the same dynamic event or of the same class of human actions with sufficient time overlap.

In contrast to the majority of existing methods, we do not impose assumptions as sufficient background information, point correspondences or linear modeling of the temporal misalignment. We use the self-similarity matrix (SSM) as a temporal "signature" of image sequences. This matrix, recently introduced in [6] for action recognition, is fairly stable under view changes and characterizes the dynamics of the scene. Indeed, similar dynamic events produce similar SSM structures as shown in Fig 1 for bending action performed by different persons and seen from different views. Time-dependent descriptors derived from these two matrices can be matched up to a time warping which we estimate using DTW.

In [6], Junejo *et al.* propose to describe the SSM with a temporal HoG-based local descriptors computed on log-polar supports of fixed size. Keeping fixed the size of these SSM portions is especially problematic when it comes to compare scenes with substantial dynamical differences (different frame rates in both synchronization problems we consider and/or different action speeds in case of action synchronization). To deal with this issue, we

[1] Emilie Dexter is supported by a grant from DGA/CNRS .

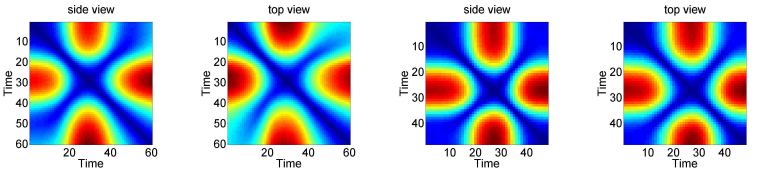


Figure 1: Self-similarity matrix illustration for two different performances of bending action, each seen from side and top views. These matrices are computed from joint trajectories of Motion Capture dataset (*mocap.cs.cmu.edu*). 3*D* data are projected to simulate both views. (a-b) SSMs of the first performance (c-d) SSMs of the second performance. The matrix structures are similar despite viewpoint differences and inter-performance variability.

introduce variable-size descriptors inspired from the scale-invariant descriptors used in visual matching. As another contribution, we also show how this synchronization framework can effectively handle cameras moving arbitrarily, provided that the dominant motion in the scene is estimated and compensated. To this end, we build SSMs exclusively on point trajectories that are automatically extracted. Note that, although point trajectories were introduced in [6] as a possible type of data for defining SSMs, optical flow was the main feature. Indeed, trajectories were only used on MoCap data, or on videos [5] where trajectories are extracted semi-automatically by tracking of body joints [1].

The remainder of this paper is organized as follows: Section 2 presents the proposed method of synchronization. Section 3 focuses on comparison results between fixed-size and time-adaptive descriptors. Section 4 and 5 are devoted to experimental results, respectively for dynamic scenes and for human actions. In Section 6, we conclude and propose future research directions.

2 Synchronization framework

In this section, we describe the common framework for synchronizing both human actions and dynamic scenes. First, we present the adaptive temporal descriptors of image sequences based on temporal self-similarities. Then, we describe the Dynamic Time Warping algorithm used in order to synchronize descriptor sequences.

2.1 Temporal descriptors

Computing temporal descriptors requires two steps: (i) building for each sequence a self-similarity matrix (SSM) which captures similarities and dissimilarities along the image sequence and (ii) computing a temporal descriptor which captures the main structures of the SSM.

2.1.1 Self-similarity matrices

Considering a sequence of images, denoted $I = \{I_1, I_2, \dots I_T\}$, the self-similarity matrix, $\mathscr{D}(I) = [d_{ij}]_{(i,j)=1\dots T}$, is a square symmetric matrix where each entry d_{ij} represents a distance between some features extracted from frames I_i and I_j. In this work, we use the Euclidean

distance on real point trajectories:

$$d_{ij} = \sum_{k=1}^{N_{ij}} \|\underline{x}_i^k - \underline{x}_j^k\|_2, \tag{1}$$

where $\underline{x}_i^k$ and $\underline{x}_j^k$ are the point positions at instants i and j of the k^{th} trajectory among the N_{ij} trajectories that span the complete time interval between frames I_i and I_j. Point trajectories are extracted with KLT tracker [13, 14]: all along the sequence, interest points (corners) are detected and tracked for some time. These trajectories can be short or long and correspond to static or moving points. Moreover we propose to overcome camera motions by estimating the dominant motion with a standard robust estimator [2, 9] and compensating point trajectories such that the point coordinates are expressed in the coordinate system of the first frame of the sequence.

2.1.2 Local SSM descriptors

To perform synchronization, we need to capture the structure of the SSM. Following [6], we consider the SSM as an "image" and extract a local descriptor around each diagonal entry of the matrix. We use a 8-bin normalized histogram of gradient directions for each of the 11 blocks of a log-polar structure. The descriptor vector of size 88, h_i corresponding to the frame I_i, is obtained by concatenating the normalized histograms as illustrated in Fig. 2. Finally, the temporal descriptor computed for an image sequence is the sequence $H = (h_1, ..., h_T)$ of such local descriptors.

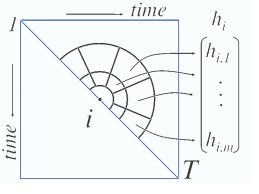


Figure 2: Local descriptors of an SSM are centered at every diagonal point $i = 1...T$. Histograms of gradient directions are computed separately for each block and concatenated to build the temporal descriptor h_i corresponding to time i.

Instead of using a fixed size for the log-polar structure, we propose to adapt its support to the temporal "scale" at the current instant. Indeed, for image sequences with different frame rates or actions performed at different speeds, corresponding SSM patterns exhibit different sizes due to these speed variations. Not taking into account this dependence of SSM patterns' size on speed fluctuations will affect in some cases the quality of the synchronization. To circumvent this problem, we propose to compute, for each diagonal point of the SSM, a temporal scale in a similar way as intrinsic scale is computed around interest points in images. At each diagonal point (i,i) of the SSM, we compute the normalized Laplacian

$$\triangle\mathscr{D}(i,i,\sigma) = \sigma^2(\mathscr{D} \star \partial_{xx}G_\sigma + \mathscr{D} \star \partial_{yy}G_\sigma)(i,i) \tag{2}$$

over a range of standard deviations σ, with G_σ denoting the isotropic Gaussian filter with variance σ^2. The best scale σ_i is the one maximizing the normalized Laplacian as proposed in [8]. The radius of the circular support that is used to compute descriptor h_i at time i is set to $2\sigma_i$.

2.2 Descriptor alignment

This section addresses the problem of aligning descriptor sequences to perform synchronization. First, we describe, in Section 2.2.1, the method used for image sequences without prior on the time warping function. In Section 2.2.2, we propose a simple approach when the unknown warping function is a linear function with known frame rate ratio.

2.2.1 Synchronization without prior

Our goal is to align descriptor sequences extracted from two self-similarity matrices. This problem is similar to the one of warping two temporal signals as in speech recognition [10] for example. DTW is a classic tool to solve this problem. In our case, DTW is used to estimate the warping function w between the time axes of the two videos. The warping between frames i and j of sequences 1 and 2 respectively is expressed as $i = w(j)$. Considering two image sequences represented by their temporal descriptors denoted $H^1 = (h_1^1, ...h_i^1, ..., h_N^1)$ and $H^2 = (h_1^2, ...h_j^2, ..., h_M^2)$ respectively, we define the cost matrix $\mathscr{C}$ for a dissimilarity measure S (Euclidean distance for example) as

$$\mathscr{C} = [c_{ij}]_{i=1...N, j=1...M} = \left[S(h_i^1, h_j^2)\right]_{i=1...N, j=1...M}. \tag{3}$$

As a consequence, the best temporal alignment is expressed by the set of pairs $\{(i, j)\}$ that yields the global minimum of the cumulative dissimilarity measure, i.e.,

$$C_T = \min_w \sum_{j=1}^{M} S(h_{w(j)}^1, h_j^2). \tag{4}$$

We can solve this optimization problem recursively using dynamic programming. Considering three possible moves (horizontal, vertical and diagonal) for the admissible monotonic warps, the partial minimum accumulated cost, for each pair of frames (i, j), is

$$C_A(h_i^1, h_j^2) = c_{ij} + \min[C_A(h_{i-1}^1, h_j^2), C_A(h_{i-1}^1, h_{j-1}^2), C_A(h_i^1, h_{j-1}^2)]. \tag{5}$$

The final solution is by definition $C_T = C_A(h_N^1, h_M^2)$. Finally, the set of synchronized pairs $\{(i, j)\}$, is obtained by tracing back the optimal path in the accumulated cost matrix from the pair of frames (N, M) to the pair $(1, 1)$.

2.2.2 Synchronization for linear warping function

An alternative approach can be used when admissible warps are restricted to linear transformations where the frame rate ratio, a, between sequences is known. In that case, the synchronization problem boils down to estimating a simple time-shift. As before, we compute SSMs and descriptors for both image sequences and the corresponding cost matrix, $\mathscr{C}$, defined in (3). Then for each possible integer time-shift k, we compute the average cost

$$c_a(k) = \frac{\sum_{j=\max(1, \frac{1-k}{a})}^{\min(\frac{N-k}{a}, M)} S(h_{aj+k}^1, h_j^2)}{\min(\frac{N-k}{a}, M) - \max(1, \frac{1-k}{a}) + 1}. \tag{6}$$

The best linear warp is then $w(j) = aj + \arg\min_k c_a(k)$, the minimizer being typically taken over $k = \left[-\frac{M+N}{4}, \frac{M+N}{4}\right]$. For visualization purpose, we can plot this average cost, c_a, as a function of time-shift k, as we shall see in experimental sections.

3 Fixed vs. adaptive size of the log-polar structure

In this section, we propose to compare synchronization results obtained using the descriptors proposed in [6] and our time-adaptive descriptors for different simulated linear warps. First, we briefly describe the estimation of the alignment error for known linear ground truth, which will serve as the evaluation criteria in this section. Then we present quantitative evaluations on 3*D* Motion Capture (MoCap) data and on natural image sequences.

3.1 Error evaluation

Let us denote $\hat{w}$ the estimated time warp and (a,k), the frame rate ratio and the time-shift of the true linear warp. We define the estimation error as the average distance of points $\{(\hat{w}(j),j)\}_{j=1\ldots M}$ on the estimated path to the line with parameters (a,k):

$$dist = \frac{1}{M}\sum_{j=1}^{M}\frac{|\hat{w}(j)-aj-k|}{(1+a^2)^{\frac{1}{2}}}. \tag{7}$$

By constructing, DTW recovers the minimum cost path in the cost matrix between pairs $(1,1)$ and (N,M). As a consequence, errors are necessarily made at the beginning and at the end because DTW finds correspondences where they cannot exist. To limit the influence of these structural errors on the evaluation criteria, we limit the averaging in (7) to instants j such that the projection of $(\hat{w}(j),j)$ on the true line has coordinates within $[1,N]$ and $[1,M]$ respectively.

3.2 Results on MoCap data

We first carry out experiments on sequences with different frame rates from 3*D* MoCap data from CMU dataset (*http://mocap.cs.cmu.edu*). Considering different actions, we construct side and top "views" by appropriate projection on 2*D* planes and we temporally subsample trajectories of one sequence to simulate different frame rate changes. For a given time-shift, we compute and synchronize SSMs as described in Section 2.2.1. We estimate the mean synchronization error by the method presented in the previous section. We evaluate this error for three types of local SSM descriptors: (1) the size of log-polar structure is fixed over time and identical for both sequences as in [6]; (2) the size is fixed over time but manually tuned for each sequence (referred as [6]*) such that the ratio of support sizes is equal to the frame rate ratio; (3) the sizes are tuned automatically as described in 2.1.2. Table 1 summarizes some results for golf action sequence pairs with different simulated frame rate ratios.

Table 1: Mean synchronization error for MoCap data for different frame rate ratios, R_{FR}

Sequence name	$R_{FR}=2$			$R_{FR}=3$			$R_{FR}=4$		
	[6]	[6]*	proposed	[6]	[6]*	proposed	[6]	[6]*	proposed
golf seq1	1.43	1.24	1.08	1.92	5.30	0.70	3.02	2.15	0.77
golf seq2	0.98	1.46	0.78	2.10	3.70	0.74	2.83	3.16	0.69
golf seq3	0.98	0.97	0.85	1.51	3.74	0.90	3.97	2.85	0.63

In general, our time-adaptive descriptor gives lower mean error values for all simulated frame rate ratios. As we can observe, choosing manually the size of the circular windows

does not necessarily provide better results than the fixed and always less precise results compared to time-adaptive size descriptors. This confirms that intrinsic time scale is important for characterizing SSM structures.

3.3 Results on natural sequences

We also validate the performance of the automatic time-adaptive descriptor by simulating different frame rates for natural sequences originally captured with time-shift but identical frame rates. After subsampling temporally image sequences, we compute SSMs and the same three temporal descriptors as in the previous MoCap experiments. Results for some pairs of sequences are reported in Table 2. As for MoCap data, the time-adaptive descriptor yields lower synchronization error.

Table 2: Mean synchronization error for natural image sequences

Sequence name	$R_{FR}=2$			$R_{FR}=3$			$R_{FR}=4$		
	[6]	[6]*	proposed	[6]	[6]*	proposed	[6]	[6]*	proposed
seq1 (81)	3.72	4.67	4.55	3.73	8.75	3.66	8.6	6.41	2.17
seq2 (−9)	3.53	3.91	1.83	7.57	8.36	4.92	5.5	5.19	1.4
seq3 (−27)	2.93	15.42	2.48	5.77	23.13	4.77	16.56	17.79	2.14

4 Synchronization results for dynamic scenes

We now present various results of dynamic scene synchronization. The first experiments, proposed in Section 4.1, deal with image sequences from static cameras whereas the second ones, in Section 4.2, consider videos without constraints on camera motions.

4.1 Sequences for static viewpoints with linear time warping

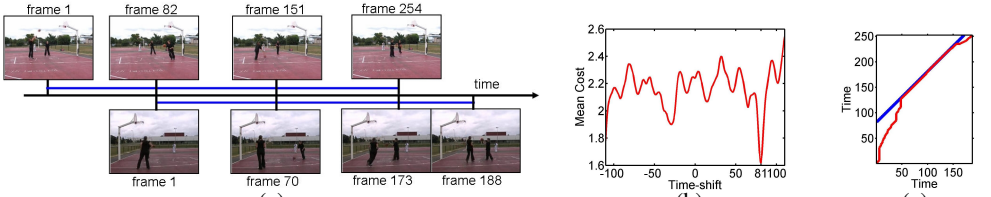
First, we validate the method for video synchronization by considering outdoor basketball videos captured by static cameras. Such a pair of image sequences is illustrated in Fig. 3. The viewpoints of the cameras are almost opposite which means that no point of the background is seen in both views and this also holds for most points on the players.

Considering that the time warp is affine with a known frame rate ratio ($a=1$), we can use the method proposed in Section 2.2.2 to estimate the unknown time-shift. Fig. 3(b) displays the mean alignment cost as a function of time-shift. We can observe a minimum for a time-shift value of 81 for both cases, which is the correct value. Assuming now that we have no prior on the time-warping function, we apply the DTW method. Extracted time warp, plotted as red curve in Fig. 3(c), almost recovers ground truth transformation (blue curve).

We can note that the dynamic scene content which is not shared by the two views can disturb the estimation of time warps. Indeed, a moving background object seen in only one view, induces trajectories which contribute to the SSM computation in one sequence and not in the second one. Consequently, SSM structures can differ and synchronization can fail although the sequences mostly show the same dynamic scene.

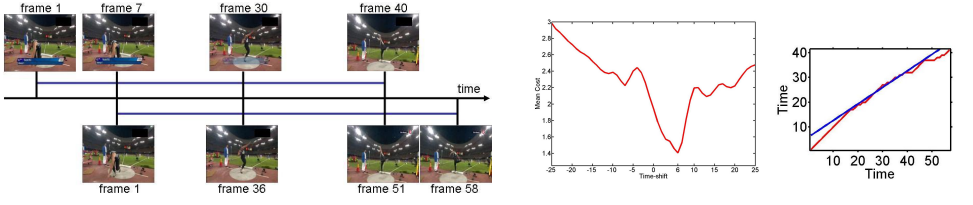
4.2 Unconstrained image sequences

Both types of warp estimation can be applied to image sequences captured by moving cameras. We detect and track KLT features whose trajectories are used in order to compute SSM after compensating the dominant motion. First, we present results on a pair of sequences



(a) (b) (c)

Figure 3: Synchronization results of basket ball sequences with opposite viewpoints and affine de-synchronization. (a) Snapshots of the two sequences showing two players. Their desynchronization simply amounts to a time-shift of 81 frames. (b) Mean cost as a function of time-shift, the global time-shift minimizer is 81, which is the correct value. (c) DTW estimates of the time warp (red) plotted with the ground truth (blue). Estimation almost recovers the ground truth with a mean estimation error of 7.29 frames.



(a) (b) (c)

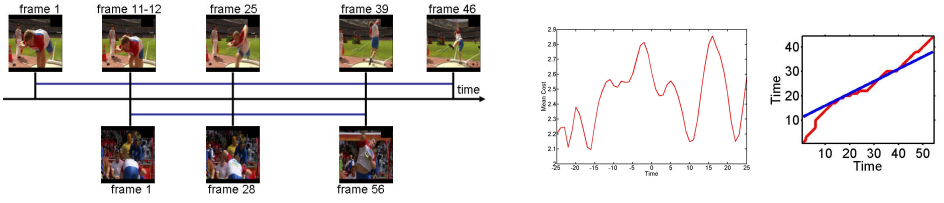
Figure 4: Synchronization results of sport sequence with its slow-motion replay. The warp to be recovered is affine with parameters a and k equal to $\frac{2}{3}$ and 6 respectively. (a) Snapshots of the two sequences displayed in a common time-line. (b) The time-warping estimation (red curve) recovers reasonably well the ground truth (blue curve) with a mean synchronization error equal to 1.16 frames.

extracted from sport broadcasts, with the second sequence being a slow-motion replay of the first sequence with linear time warp as illustrated in Fig. 4(a). Result of time-shift estimation for known frame rate ratio is proposed in Fig. 4(b) where we can observe that the estimated shift corresponds to the correct value of 6 frames. DTW estimation of synchronization recovers partially the ground truth as depicted in Fig. 4(c). Yet, the estimation is not perfect probably due to the limited moves allowed in the DTW approach. However the mean synchronization error is only equal to 1.16 frames with a standard deviation of 1.02.

Preceding example was based on two sequences with identical viewpoints. We now present another sport example with a drastic difference of viewpoints and different frame rates (Fig. 5(a)). Results are shown in Fig. 5(b) if we assume that the time-warping function is affine with unknown time-shift and in Fig. 5(c) when we assume no prior on the warping function. We can see that the DTW approach (red curve in Fig. 5(c)) recovers reasonably the ground truth (blue curve) as in the preceding example, whereas the time-shift estimated by the first method is not the correct value. However, we can observe that several local minima are close to the global one, including one reached for a time-shift of 10 frames, close to the correct value which is 11 frames.

5 Synchronization results for actions

In this section, we present various results on natural human action synchronization in different videos, including movies and sport broadcasts. Complete ground truth of synchronization



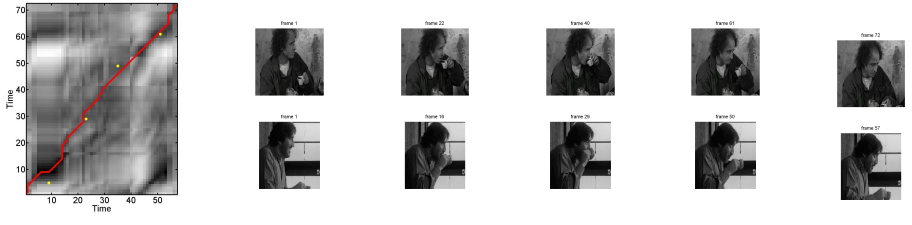
(a) (b) (c)

Figure 5: Synchronization results of sport sequence with its slow-motion replay. The warp to be recovered is affine with parameters a and k equal to $\frac{1}{2}$ and 11 respectively. (a) Sequences represented by several key-frames (b) Mean matching cost w.r.t. the time-shift. The minimum is reached for a time-shift of -16 frames which is not the correct value. (c) The time-warping estimation with DTW (red curve) recovers reasonably well the ground truth (blue curve) with a mean synchronization error equal to 1.77 frames.

are not available for these experiments. We use however some pairs of key-frames manually chosen to evaluate roughly the quality of estimated warps.

5.1 Drinking and smoking actions in a movie

In the movie *Coffee and Cigarettes*, a lot of drinking and smoking actions are performed by several actors and seen from different viewpoints. For both actions, we compute SSMs based on real point trajectories (or portions of trajectories) included in bounding boxes centered on the face of the actors (extracted from annotations [7]). An example of synchronization result for drinking action is illustrated in Fig. 6. The estimated warping function (red curve) is close to our partial ground truth time correspondences shown as yellow points, despite view and length differences between sequences.



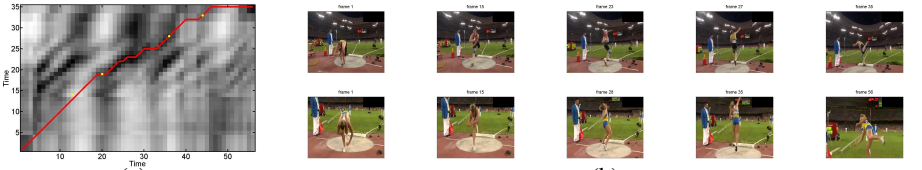
(a) (b)

Figure 6: Synchronization of two drinking actions. (a) Cost matrix with the estimated warping function (red curve) and some manually picked time correspondences (yellow points). (b) Some time correspondences from the estimated warping function illustrated by frame correspondences.

In this case, we use annotated bounding boxes. This could be automatized thanks to an appropriate object detectors (e.g. face detector). However this would probably be less robust for natural sequences. For example, classical face detection methods sometimes fail due to the face orientation of some characters in this movie.

5.2 Sport actions

Synchronizing actions can be particularly interesting for analyzing or comparing techniques of athletes. We propose synchronization results of such natural actions extracted from sport



(a) (b)

Figure 7: Synchronization of shot put action. (a) Cost matrix and estimated time warping function with some ground truth correspondences. (b) Some correspondence results extracted from the estimated time warp. For each correspondence, frames corresponds to similar athlete posture.

broadcasts. In this case, the views are really similar but the action in the second sequence is almost twice as long as the one in the first sequence. The synchronization result illustrated by the red curve in Fig. 7(a) is well aligned with the sparse ground truth correspondences (yellow points) that we picked by hand.

6 Conclusion

We have presented a general view-independent framework for synchronizing human actions and dynamic scenes. The approach is based on temporal speed-invariant description of image sequences from self-similarity matrices. To handle camera motions, we have proposed to compute these matrices from real point trajectories, previously compensated. As the structures of these matrices are stable while capturing discriminative dynamic patterns, they allow the definition of temporal descriptors adapted to the alignment problem in these two contexts. We have proposed to estimate a characteristic "time-scale" at each instant so that adapted descriptors better characterize the SSM structures. We have demonstrated that this automatic adaptation improves synchronization results especially for sequences captured with different frame rates. The main advantage of the whole framework is that we do not impose restrictive assumptions as sufficient background information or point correspondences between views. Due to the use of DTW, we can perform both tasks of synchronization even when temporal misalignment is not a simple shift, but an arbitrary warp. We have assessed the performance of our approach on challenging real image sequences captured by static or *moving* cameras.

In this work, we assume that the two image sequences correspond to a same/similar dynamic event(s). The method will be exploited in future work to address other tasks such as action clustering, action detection, video scene duplicate detection or video matching where such an assumption has to be relaxed.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proc. ICCV*, pages 1–8, 2007.

[2] M. J. Black and P.Chau Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, January 1996.

[3] R.L. Carceroni, F.L.C. Padua, G.A.M.R. Santos, and K.N. Kutulakos. Linear sequence-to-sequence alignment. In *Proc. CVPR*, pages I: 746–753, 2004.

[4] Y. Caspi and M. Irani. Spatio-temporal alignment of sequences. *IEEE T-PAMI*, 24(11): 1409–1424, November 2002.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE T-PAMI*, 29(12):2247–2253, December 2007.

[6] I.N. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *Proc. ECCV*, pages 293–306, 2008.

[7] I. Laptev and P. Pérez. Retrieving actions in movies. In *Proc. ICCV*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[8] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, November 1998.

[9] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.

[10] L. Rabiner, A. Rosenberg, and S. Levinson. Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 26(6):575– 582, 1978.

[11] A. Rao, C.and Gritai, M. Shah, and T. F. Syeda Mahmood. View-invariant alignment and matching of video sequences. In *Proc. ICIP*, pages 939–945, 2003.

[12] G.P. Stein. Tracking from multiple view points: Self-calibration of space and time. In *Proc. CVPR*, volume 1, pages 521–527, 1999.

[13] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University Technical Report CMU-CS-91-132, 1991.

[14] C. Tomasi and J. Shi. Good features to track. In *Proc. CVPR*, pages 593–600, 1994.

[15] T. Tuytelaars and L.J. Van Gool. Synchronizing video sequences. In *Proc. CVPR*, volume 1, pages 762–768, 2004.

[16] Y. Ukrainitz and M. Irani. Aligning sequences and actions by minimizing space-time correlations. In *Proc. ECCV*, 2006.

[17] M. Ushizaki, T. Okatani, and K. Deguchi. Video synchronization based on co-occurrence of appearance changes in video sequences. In *Proc. ICPR*, pages III: 71–74, 2006.

[18] L. Wolf and A. Zomet. Wide baseline matching between unsynchronized video sequences. *IJCV*, 68(1):43–52, June 2006.