


# Video Synchronization with Trajectory Pulse

Xue Wang() and Qing Wang

School of Computer, Northwestern Polytechnical University,  
Xi'an 710072, People's Republic of China  
[xwang@mail.nwpu.edu.cn](mailto:xwang@mail.nwpu.edu.cn)

**Abstract.** This paper presents a method to temporally synchronize two independently moving cameras with overlapping views. Temporal variations between image frames (such as moving objects) are powerful cues for alignment. We first generate pulse images by tracking moving objects and examining the trajectories for changes in speed. We then integrate a rank-based constraint and the pulse-based matching, to derive a robust approximation of spatio-temporal alignment quality for all pairs of frames. By folding both spatial and temporal cues into a single alignment framework, finally, the nonlinear temporal mapping is found using a graph-based approach that supports partial temporal overlap between sequences. We verify the robustness and performance of the proposed approach on several challenging real video sequences. Compared to state-of-the-art techniques, our approach is robust to tracking error and can handle non-rigid scene alignment in complex dynamic scenes.

## 1 Introduction

Video synchronization is part of a more general video alignment problem which occurs in tasks such as human motion recognition, video retrieval, multi-view surveillance and 3D visualization. Videos must be aligned both spatially and temporally. Spatial alignment computes the geometrical transformation of 2D or 3D coordinate systems of temporally aligned frames, so that the object of interest is in correspondence. Temporal alignment computes 1D temporal transformation by synchronizing frames to achieve good spatial alignment.

Jointly reasoning about temporal and spatial alignment improves the robustness of the system. There are two main challenges. First, explicit 2D or 3D spatial alignment is very difficult to compute for moving cameras on dynamically changing scene with multiple moving objects. Second, due to non-predictable frame drops, temporal context constraints (i.e. continuity) can not be applied everywhere for temporal alignment.

The key insight is the spatio-temporal rhythm of movement of a human body. Both the geometrical configuration and the speed variations of body parts, are strong cues for alignment. Furthermore, the body configuration and movement are often coupled. Our method uses sparse space-time point trajectories as input.

---

This work was partially supported by National Natural Science Foundation of China (61272287, 61531014).

We introduce a temporal feature, *pulse*, along each trajectory by examining the changes in speed. We call this feature *pulse* as it reflects the rhythm of movement, also the peaks and troughs on a *pulse* image are often associated with the keyframes of body poses. With the *pulse* features, our objective is to measure the sequence-to-sequence alignment quality between pairs of frames with a gross approximation of synchronization (i.e., constant offset model). On the other hand, following traditional image-to-image alignment techniques, we measure the spatial configuration alignment between two camera frames using a rank constraint based on epipolar geometry. This implicitly considers 3D transformation without solving a hard reconstruction problem. Finally we fold both the *pulse* based matching and the rank constraint into a single alignment framework, and compute the globally optimal path that minimizes spatial and temporal misalignments.

## 2 Related Works

Most video alignment techniques assume stationary or rigidly fixed cameras, thus a fixed spatial transformation between corresponding frames is guaranteed and need not be re-estimated at runtime. Commonly exploited geometric constraints include plane-induced homography [1, 2], affine transformation [3], binocular epipolar geometry constraint [1, 4, 5], deficient rank conditions arose from special projection models [6–8] and so on. Anthony et al. [9] propose to synchronize stationary cameras using inflection points, which are found by examining the trajectories for changes in direction. Once an event has been identified in two such videos, a temporal mapping between the sequences can be globally described by simple parametric models, like constant offset model [2, 4, 6, 7] or 1D affine model [1, 5, 8]. Nonlinear temporal mapping is used to cope with free form of time correspondence [3, 10]. Assuming simultaneous recording, this kind of temporal rigidity is preserved even for independently moving cameras [11–14]. If related videos are captured at different points in time, previous work [15–18] assumes approximately coincident camera trajectories, to make sure that corresponding frames are captured from similar viewpoints.

Our scenario is most closely related to the work in [11–14], which focuses on video alignment for independently moving cameras and non-rigid dynamic scenes. Tresadern and Reid [11] develop a unified rank constraint framework for homography, perspective and affine projection models. Tuytelaars and Gool [12] assume a scaled orthographic projection model and find corresponding frames use the line-to-line distance of the back-projection 3D lines of matching points. Lei and Yang [14] use the tri-ocular geometric constraint of point/line features to build the timeline maps for multiple sequences. These methods assume that the features are tracked successfully throughout each sequence and matched across sequences, which are hardly assured in wide baseline conditions. Also they try to recover a linear synchronization. Dexter et al. [13] propose a time-adaptive descriptor based on self-similarity matrices to perform nonlinear synchronization. However, they use static points in the background to estimate a dominant motion

to compensate modest camera motion, which only works for distant views or planar scenes.

### 3 Trajectory Pulse

Let  $F = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$  and  $F' = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_N\}$  denote two corresponding feature trajectories from video sequences  $M$  and  $N$  frames long respectively. Both sequences have a collection of feature trajectories  $\Gamma = (F^1, F^2, \dots, F^K)$  and  $\Gamma' = (F'^1, F'^2, \dots, F'^K)$ , with  $K$  the number of trajectories for both sequences. Let  $v_i(j)$  denote the  $j$ th frame of video  $i$ . Our goal is to find a nonlinear temporal mapping  $\mathbf{p} : \mathbb{N} \rightarrow \mathbb{N}$ , where  $\mathbf{p}(n) = m$  maps  $v_1(m)$  in the reference sequence to  $v_2(n)$  in the observed sequence. Considering the situation that the temporal displacements are not necessarily integer values, instead of a sub-frame accurate synchronization, we find the temporally closest frame.

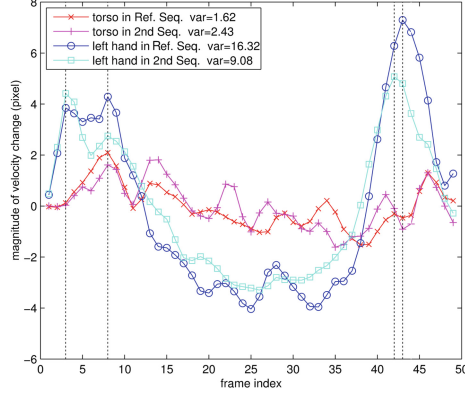
Given two sets of corresponding trajectories  $\Gamma$  and  $\Gamma'$ , there are two ways of looking at the spatio-temporal alignment. To represent the time varying structure in the trajectory space, between two corresponding point trajectories  $F$  and  $F'$ , a temporal trajectory affinity across views can be used for temporal synchronization. To represent the time varying structure in the shape space, between two instantaneous 2D point configurations  $(\mathbf{x}_m^1, \mathbf{x}_m^2, \dots, \mathbf{x}_m^K)$  and  $(\mathbf{x}_n'^1, \mathbf{x}_n'^2, \dots, \mathbf{x}_n'^K)$ , a spatial shape affinity across views can also be used. Thus a best match should have both high temporal trajectory affinity and high spatial shape affinity.

We generate pulse images by examining the trajectories for changes in speed, which reflects how fast the instantaneous point velocity changes. The frames where the speed changes drastically can be seen as the pulse feature for temporal alignment. Two examples of pulse images for corresponding trajectories are given in Fig. 1. Each pulse image has been normalized to zero mean. The trajectories of torso, whose variances are 1.62 and 2.43 respectively, lack distinctive pulse features so that multiple temporal mappings can align the two trajectories. While the trajectories of left hand, whose variances are 16.32 and 9.08 respectively, provide discriminative pulse features to determine a unique solution for temporal alignment.

The conclusions accord with the general impression. A static object contributes nothing for video alignment. In general, greater temporal variations a dynamic scene the better chances of exact video synchronization. The pulse feature with obvious change in speed provides powerful alignment cues. Once we have generated the pulse images for the corresponding trajectories, for each feasible pair of frames, the synchronization is grossly determined by the frames using a constant offset model. Accordingly the pulse-based trajectory affinity  $\mathbf{A}_t$  for frame pair  $(m, n)$  is defined as follows,

$$\mathbf{A}_t(m, n) = \exp\left(-\frac{trj_{m,n}}{\sigma_t^2}\right), \quad (1)$$

where  $trj_{m,n}$  is the maximum pulse images difference for corresponding trajectories, and  $\sigma_t$  is a positive rate.



**Fig. 1.** Two examples of pulse images for corresponding trajectories. The discriminative pulse features are indicated in black dotted vertical lines.

## 4 Nonlinear Temporal Alignment

For perspective projection model, given  $K$  corresponding points, the unknown fundamental matrix  $\mathbf{F}$  can be computed using  $\mathbf{M}\mathbf{f} = 0$ , where  $\mathbf{M}$  is a  $K \times 9$  observation matrix of constraints defined by the image feature locations,  $\mathbf{f}$  is the elements of the fundamental matrix:  $\mathbf{f} = [f_1, \dots, f_8, 1]^T$  [11]. Since it is a homogenous equation, for a solution of  $\mathbf{f}$  to exist,  $\mathbf{M}$  must have rank at most eight. However, due to the noise or the tracking error, the rank of  $\mathbf{M}$  will almost always be of full rank. We examine the *effective rank*,  $\hat{n}$ , of the observation matrix [7]. Let  $\lambda_1, \dots, \lambda_h$  denote the singular values of  $\mathbf{M}$ . The sum of remaining singular values, denoted as  $dst = \sum_{k=\hat{n}+1}^h \lambda_k$ , can be used to measure the matching of two instantaneous 2D point configurations. The smallest  $dst$  of  $\mathbf{M}$  corresponds to the best match of frames. Finally, we transform  $dst$  to the shape affinity  $\mathbf{A}_s(m, n)$  as follows,

$$\mathbf{A}_s(m, n) = \exp\left(-\frac{dst_{m,n}}{\sigma_s^2}\right). \quad (2)$$

where  $\sigma_s$  is a positive rate. Thus we set the spatio-temporal affinity  $\mathbf{A}(m, n)$  for frame pair  $(m, n)$  by integrating the shape and trajectory affinity,

$$\mathbf{A}(m, n) = \exp\left[-\left(\frac{dst_{m,n}}{\sigma_s^2} + \frac{trj_{m,n}}{\sigma_t^2}\right)\right]. \quad (3)$$

where  $\sigma_t$  and  $\sigma_s$  control the rate of decay for trajectory and shape weights respectively. Finally, we transform the  $\mathbf{A}$  to obtain the cost matrix  $\mathbf{C}$  in which low values indicate frames that are likely to have a *good* match. The entries of the cost matrix  $\mathbf{C}$  are given by,

$$\mathbf{C}(m, n) = 1 - \frac{\mathbf{A}(m, n)}{\mathbf{A}_{max}}, \quad (4)$$

where  $\mathbf{A}_{max}$  is the maximum value of  $\mathbf{A}$ .

The nonlinear temporal mapping  $\mathbf{p} : \mathbb{N} \rightarrow \mathbb{N}$  is referred as a path through the cost matrix. We use the mapping computing algorithm described in [16] to find the optimal path.

## 5 Experiments

In this section, we evaluate the proposed alignment algorithm on several channeling real data. We focus on the alignment of sequences captured by independently moving cameras simultaneously from different viewpoints. The corresponding feature trajectories are the joint points in a human body labelled manually. When some points are occluded, we interpolate the missing locations between consecutive frames.

We perform an ablative analysis of our approach, by comparing to the following baselines: (1) using trajectory affinity  $\mathbf{A}_t$  alone of Eq. 1, and (2) using shape affinity  $\mathbf{A}_s$  alone of Eq. 2. We additionally compare our approach with three state-of-the-art synchronization algorithms for independently moving cameras [12, 13, 16], abbreviated as BPM, MFM and SMM, respectively.

Given the ground truth  $\{\hat{\mathbf{p}}(j), j\}_{j=1\dots M}$ , we use the average absolute temporal alignment error  $\varepsilon = \frac{1}{M} \sum_{j=1}^M |\hat{\mathbf{p}}(j) - \mathbf{p}(j)|$  as our basic accuracy metric.

### 5.1 Non-rigid Scene Alignment

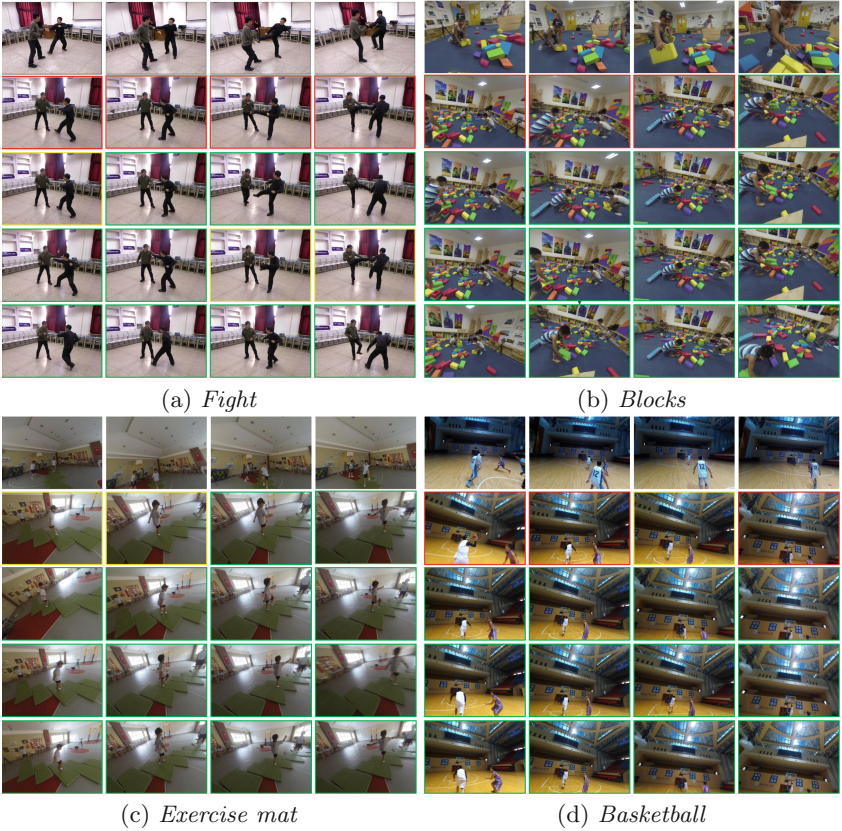
For evaluation with videos captured with independently moving simultaneously, we first use the *Fight* dataset provided by [19]. Further, we introduce a first-person dataset captured by head-mounted cameras, which consists of three social interaction scenes. The scenes, *Blocks* and *Exercise mat*, capture tetradic interactions between children aged 5–6. For the *Basketball* scene, the players strategically take advantage of team formation (5v5). Ground truth are obtained by manual synchronization. We take two clips with temporal overlapping for alignment. Within the observed sequence, we drop several frames randomly at a maximum rate 5%.

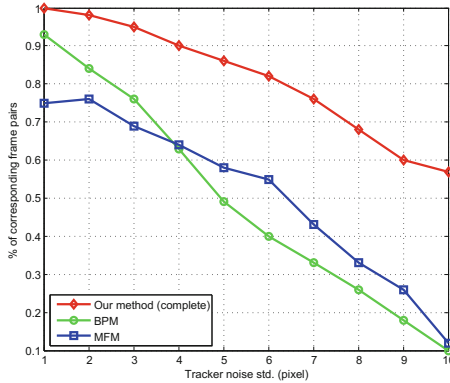
The average temporal alignment errors with respect to the ground truth are summarized in Table 1. The complete model of our approach outperforms other methods on the test sequences. The content-based snapping [16] assumes that two frames are more likely to be “alignable” if they contain a large number of similar features, and it is unable to accurately synchronize sequences in the wide baseline viewing condition.

Figure 2 shows the synchronization results for sample frames using the complete model of our approach, BPM, MFM and SMM on different scenes. Three alignment situations are defined according to the alignment accuracies. For a frame  $v_i(j)$ , its alignment error is defined as  $\varepsilon_j = |\hat{\mathbf{p}}(j) - \mathbf{p}(j)|$ . Thus, the frame  $v_i(j)$  is referred as a *matched*, *slightly mismatched* or *mismatched* frame when  $\varepsilon_j \leq 1$ ,  $\varepsilon_j \leq 2$  and  $\varepsilon_j \geq 3$ , respectively.

**Table 1.** Comparisons of alignment error for non-rigid scene alignment.

	<i>Fight</i>	<i>Blocks</i>	<i>Exercise mat</i>	<i>Basketball</i>
BPM [12]	12.6	24.7	15.1	22.3
MFM [13]	4.9	14.2	16.8	9.4
SMM [16]	18.7	138.5	106.9	19.1
Our method (shape)	8.1	17.0	30.7	15.3
Our method (trajectory)	2.5	13.6	8.4	3.7
Our method (complete)	<b>1.2</b>	<b>1.6</b>	<b>2.5</b>	<b>0.8</b>

**Fig. 2.** Synchronization results for sample frames on different scenes. *From top to bottom:* Sample frames from the reference sequence, corresponding frames from the observed sequence by the complete model of our method, BPM, MFM and SMM. The red, yellow and green rectangles around the frames indicate *matched*, *slightly mismatched* and *mismatched* frames, respectively. (Color figure online)



**Fig. 3.** Comparison of alignment error versus localization error on the *Blocks* scene. The alignment error bound is fixed at  $\zeta = 1$  frame.

## 5.2 Robustness Analysis

Feature-based methods rely on the point trajectories as input data for alignment. In a practical situation, the feature trajectory is usually imperfect and contains noise. A robust alignment algorithm should be tolerant to certain tracking errors. We evaluate the effect of noisy trajectories on the proposed approach using the robustness analysis [3, 5]. We consider the percentage of estimated corresponding frame pairs below a given bound  $\zeta$ , which allows us to assess the algorithm robustness to compute high-accurate timelines ( $\zeta \leq 1$  frame) as well as its behavior in a less challenging situation (e.g.,  $\zeta \leq 2$  frames or  $\zeta \leq 5$  frames).

Normally distributed and zero mean noise with various values of variance is added to the tracked feature trajectories. The original point locations are labelled manually. Then we can estimate the algorithm by computing the average temporal alignment error in a variety of settings. Figure 3 shows the impact of localization error on alignment accuracy for the complete model of our approach, BPM and MFM on the *Blocks* scene. As expected, the ability to achieve accurate alignments diminishes with increased noise levels. Our approach can align almost 80% of the total frames within a  $\pm 1$  frame offset with respect to the ground truth, even when the tracker noise variance reaches 6. Due to the sensitivity to tracking error, previous methods deteriorate at a faster rate as the tracking noise level increases comparing to ours.

## 6 Conclusion

We present a general framework for synchronizing dynamic scenes in the presence of independent camera motion. We demonstrate the feasibility of folding pulse-based trajectory affinity and rank-based shape affinity into a single alignment framework. Experiments conducted on several challenging video sequences show that the proposed approach outperforms the synchronization accuracy and the robustness w.r.t the state-of-the-art techniques.



## References

1. Caspi, Y., Simakov, D., Irani, M.: Feature-based sequence-to-sequence matching. *Int. J. Comput. Vis.* **68**, 53–64 (2006)
2. Dai, C., Zheng, Y., Li, X.: Accurate video alignment using phase correlation. *IEEE Signal Process. Lett.* **13**(12), 737–740 (2006)
3. Lu, C., Mandal, M.: A robust technique for motion-based video sequences temporal alignment. *IEEE Trans. Multimedia* **15**, 70–82 (2013)
4. Pundik, D., Moses, Y.: Video synchronization using temporal signals from epipolar lines. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010. LNCS*, vol. 6313, pp. 15–28. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15558-1\\_2](https://doi.org/10.1007/978-3-642-15558-1_2)
5. Pádua, F., Carceroni, R., Santos, G., Kutulakos, K.: Linear sequence-to-sequence alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 304–320 (2010)
6. Wolf, L., Zomet, A.: Correspondence-free synchronization and reconstruction in a non-rigid scene. In: *Workshop on Vision and Modelling of Dynamic Scenes* (2002)
7. Wolf, L., Zomet, A.: Wide baseline matching between unsynchronized video sequences. *Int. J. Comput. Vis.* **68**, 43–52 (2006)
8. Rao, C., Gritai, A., Shah, M., Syeda-Mahmood, T.: View-invariant alignment and matching of video sequences. In: *International Conference on Computer Vision* (2003)
9. Whitehead, A., Laganieri, R., Bose, P.: Temporal synchronization of video sequences in theory and in practice. In: *Applications of Computer Vision and the IEEE Workshop on Motion and Video Computing*, pp. 132–137(2005)
10. Singh, M., Cheng, I., Mandal, M., Basu, A.: Optimization of symmetric transfer error for sub-frame video synchronization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008. LNCS*, vol. 5303, pp. 554–567. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-88688-4\\_41](https://doi.org/10.1007/978-3-540-88688-4_41)
11. Tresadern, P.A., Reid, I.D.: Video synchronization from human motion using rank constraints. *Comput. Vis. Image Underst.* **113**, 891–906 (2009)
12. Tuytelaars, T., Gool, L.V.: Synchronizing video sequences. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2004)
13. Dexter, E., Pérez, P., Laptev, I.: Multi-view synchronization of human actions and dynamic scenes. In: *Proceedings of the British Machine Vision Conference* (2009)
14. Lei, C., Yang, Y.: Trifocal tensor-based multiple video synchronization with sub-frame optimization. *IEEE Trans. Image Process.* **15**, 2473–2480 (2006)
15. Evangelidis, G., Bauckhage, C.: Efficient subframe video alignment using short descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 2371–2386 (2013)
16. Wang, O., Schroers, C., Zimmer, H., Gross, M., Sorkine-Hornung, A.: Videosnapping: interactive synchronization of multiple videos. In: *SIGGRAPH* (2014)
17. Diego, F., Ponsa, D., Serrat, J., López, A.: Video alignment for change detection. *IEEE Trans. Image Process.* **20**, 1858–1869 (2011)
18. Diego, F., Serrat, J., López, A.: Joint spatio-temporal alignment of sequences. *IEEE Trans. Multimedia* **15**, 1377–1387 (2013)
19. Ye, G., Liu, Y., Hasler, N., Ji, X.: Performance capture of interacting characters with handheld kinects. In: *Proceedings of European Conference on Computer Vision* (2012)



Intelligent Visual Surveillance

4th Chinese Conference, IVS 2016, Beijing, China,

October 19, 2016, Proceedings

Zhang, Z.; Huang, K. (Eds.)

2016, XII, 163 p. 71 illus., Softcover

ISBN: 978-981-10-3475-6