Video Alignment

Literature Survey

Spring 2005
Prof. Brian Evans
Multidimensional Digital Signal Processing Project
The University of Texas at Austin

Omer Shakil

Abstract

This literature survey compares various methods to align two videos. The idea can be extended for alignment of multiple video sequences. The sequences are recorded by un-calibrated video cameras with fixed (but unknown) internal parameters. It will be shown that by using a combination of spatial information, temporal changes and the inbuilt frame-to-frame transformation of the video sequences, efficient video alignment can be achieved. The last method described in the survey demonstrates a novel approach to recover alignment of non-overlapping videos. Video Alignment gives rise to a wide variety of new applications that are not possible when only image-to-image alignment is used.

1. Introduction

The problem of image alignment has been extensively studied, and successful solutions have been suggested [4, 6]. The issue here is to estimate point correspondences between two images, i.e. given any pixel *(x, y)* in one image, find its corresponding pixel *(x', y')* in the other image, where $x' = x + u$, $y' = y + v$, and *(u, v)* is the calculated spatial displacement vector. The image alignment techniques can be divided into two broad categories: the first category includes feature-based approaches [7], in which common features are detected and matched across the two images; the second category comprises of methods that directly match image intensities [4]. As both of these categories rely on the spatial coherence of the given images, there must be a significant overlap between the two images for these techniques to work.

This literature survey evaluates methods to align two videos. Here, the problem is that the sequences may not be synchronized; hence alignment in time is required as well. Given points *(x, y, t)* from video sequence *V* and *(x', y', t')* from video sequences *V',* the task is to recover the transformation between them i.e. find *(u, v, w)* such that $(x', y', t') = (x+u, y+v, t+w)$. The temporal "redundancy" in successive frames of a video can be used to move a step beyond from the traditional image alignment techniques that exploit spatial coherence between the given images. Therefore, by using the temporal behavior of the videos (frame-to-frame transformations), alignment can be achieved even when the corresponding frames from each sequence have no spatial overlap between them [3].

A variety of applications, esp. in surveillance and security systems, can benefit heavily from alignment of multiple video sources. A primary application is monitoring multiple video inputs (high security areas) and the output being displayed in one composite video sequence, instead of various monitors dedicated for each individual input source. Other applications include generation of wide screen movies and super-resolution in time and space.

2. Overview

The problem of video alignment can be simplified to aligning corresponding frames using image alignment techniques, but there are cases when using only common spatial information is not enough to determine the transformation. One such example is illustrated below:

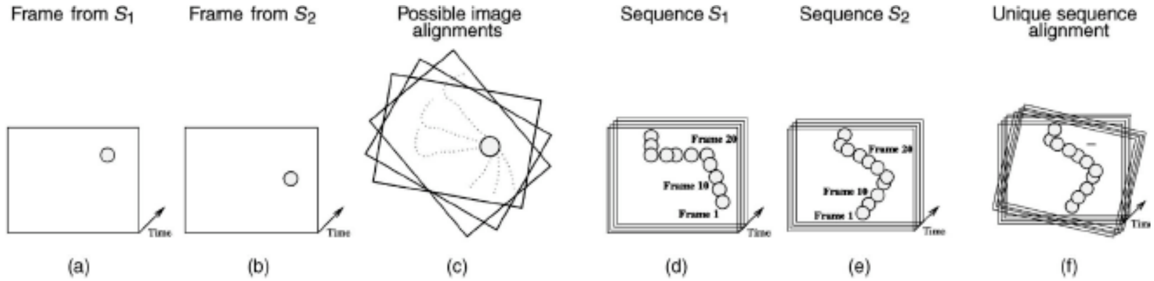


Figure 1: Spatial ambiguities in image-to-image alignment. (a) and (b) Show two corresponding frames in time from two different video sequences viewing the same moving ball. There are infinitely many valid image alignments between the two frames, some of them shown in (c). (d) and (e) Display the two sequences of the moving ball. There is only one valid alignment of the two trajectories of the ball. This uniquely defines the alignment both in time and in space between the two video sequences (f). [1]

This literature survey describes various approaches with different constraints that attempt to solve the problem of Video Alignment. As explained above, the simple image alignment techniques would not produce desirable results. In some of the successful methods, only the temporal variations (moving objects) between the sequences are used [1, 7, 9]. Other methods try to link both spatial and temporal information together for better results [1, 2, 3]. All of these methods impose some constraint on the movement of the video sequences. A table summarizing the advantages and disadvantages of all these techniques is also provided after discussing them in detail.

3. Feature-Based Alignment

Feature-based image alignment can be generalized to video alignment by extending the notion of *feature points* into *feature trajectories*. The alignment between the given sequences can be recovered by establishing correspondences between trajectories. The advantage of using this method over simple image alignment is illustrated in figure 2. It shows two sequences with several small moving objects.

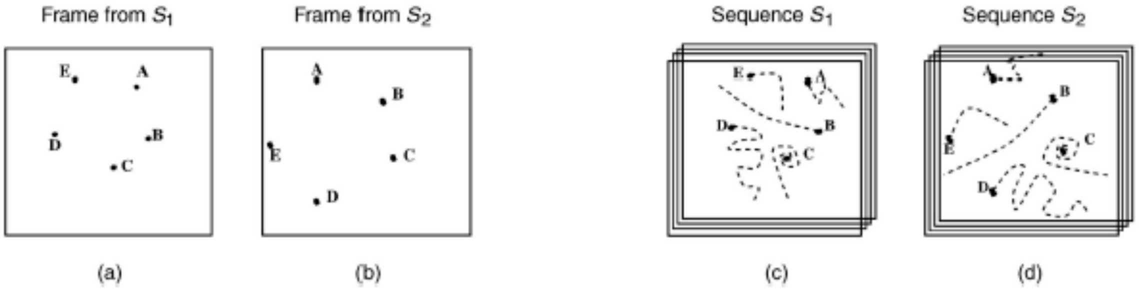


Figure 2: Point versus trajectory correspondences. (a) and (b) Display two frames out of two sequences recording five small moving objects (marked by A, B, C, D, and E). (c) and (d) Display the trajectories of these moving objects over time. When analyzing only single frames, it is difficult to determine the correct point correspondences across images. However, point trajectories have additional properties that simplify the correspondence problem across two sequences (both in space and time). [1]

Actually, a single pair of non-trivial (moving in a straight line can lead to ambiguity) corresponding trajectories is sufficient to find the alignment between the two image sequences. The main idea of the algorithm [1] is described below:

1. *Construct feature trajectories by tracking the centroids of moving objects.*

2. *For each possible pair of corresponding trajectories, find the transformation matrix.*

3. *Choose the transformation matrix that gives the least MSE after alignment.*

4. Repeat steps 2 and 3 N times, and choose the best transformation matrix.

Stein [7, 9] has proposed similar technique. However, he has treated the features as unordered collection of points and not as trajectories resulting in ambiguities. This is resolved by avoiding the combinatorial complexity of establishing point matches to all points in all frames [1]. It is also reported that using trajectories of moving objects centroids is better than using trajectories of intensity-based feature points. Each moving object is extracted as one unit and hence the output is improved. The algorithm works for only stationary, or jointly moving video sources.

4. Direct-Based Alignment

Image alignment techniques based on spatial brightness variations [4] can be generalized to recover the alignment parameters between two videos [1]. The coarse-to-fine estimation framework for image alignment is generalized here to handle both time and space.

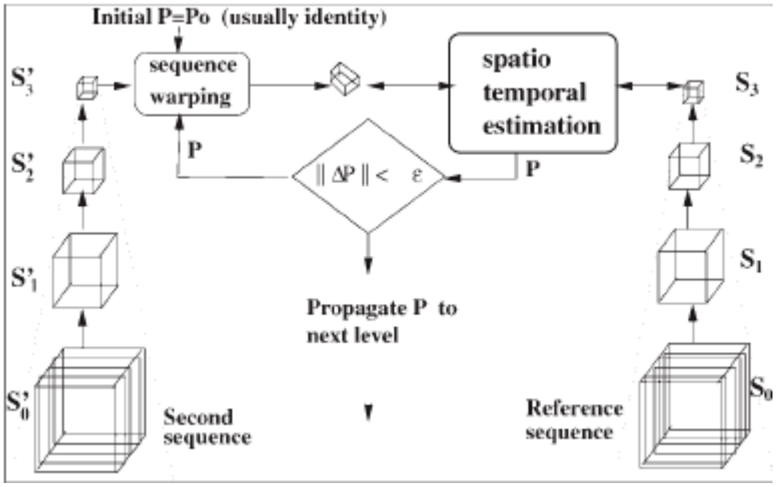

Initial P=Po (usually identity)
sequence warping
spatio temporal estimation
S'_3
S'_2
S'_1
S'_0
S_3
S_2
S_1
S_0
P
P
$\| \Delta P \| < \varepsilon$
Propagate P to next level
Second sequence
Reference sequence

Figure 3: Direct sequence-to-sequence alignment.

Direct-based alignment can handle very complex scenes, as it does not rely on feature extraction. This algorithm integrates all available spatio-temporal information within the sequence adding the benefits of both image-based and trajectory-based techniques [1]. The method works only for stationary or jointly moving video sources and is outlined below:

1. A spatio-temporal pyramid is constructed for each input sequence (left and right as in figure).

2. The spatio-temporal alignment estimator is applied iteratively at each level. It refines the approximation based on the residual misalignment between the reference sequence and warped version of the second sequence (warping in time and in space, marked by a skewed cube).

3. The output of the current level is propagated to the next level to be used as an initial estimate.

5. Video Matching

This algorithm describes a method that generates a new version of the secondary video that is spatially and temporally registered with the primary video. The algorithm [2] is described below:

1. Search for possible pairings between frames in the primary and secondary video using the robust image alignment algorithm to evaluate candidate frames

2. For each primary frame, warp back the found secondary frame into alignment with the primary frame.

The heart of the algorithm is the novel robust image alignment technique [2]. This method begins by selecting feature points followed by a search for the most likely match in the secondary image according to the corresponding weighting function (includes both pixel matching and motion consistency). Pixel matching basically compares every pixel from the primary images to a 3-by-3 neighborhood of pixels in the secondary image. The pixel receives a penalty if it lies outside the minimum and maximum images. Motion consistency is a measurement to determine how well the computed transformation for a particular correspondence agrees with its neighbors.

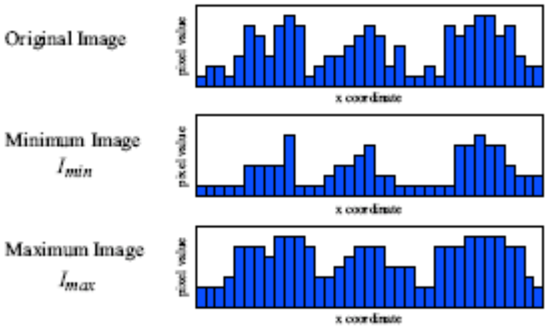
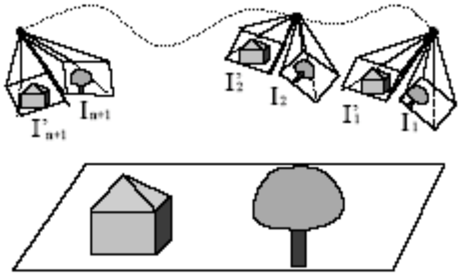


Figure 4: Each plot represents a cross section of a hypothetical image. The image is (non-linearly) filtered so each pixel becomes the minimum or maximum of its 3 by 3 neighborhood.

Video matching algorithm requires that each video sequence follow nearly the same trajectory through space and must contain sufficient texture. This is because the algorithm does not model discontinuities in the correspondence field.

6. Non-Overlapping Sequences



All of the different methods described above assume that there is a significant overlap in the field of view of the two image sequences. The algorithm described by Caspi and Irani [3, 8] outlines a simple approach to align non-overlapping sequences. This approach utilizes the similar transformations in the jointly moving camera sources to recover alignment both in time and in space. The main observation is that the frame-to-frame transformation matrices of the two image sequences are similar (conjugacy relation) and hence their eigen values are equal (up to a scale factor). The cosine of the angle between the two corresponding eigen vectors can be used to recover the temporal synchronization between the two sequences. The algorithm [3, 8] to determine spatial alignment for a planar scene is described as

1. For each pair of temporally corresponding transformations, find their eigen values.

2. Estimate the scaling factor using least squares minimization.

3. Use all the frame-to-frame transformations to form a homogenous set of linear equations and solve it to find the transformation matrix between the two sequences.

The approach for a non-planar scene is similar; transformation matrices and epipole constraints together define the complete 3D transformation. This approach works when the calculated transformations are extremely accurate. Spencer and Shah [10] have defined four new measures that are robust to noisy signals: Similarity, Roll motion, Translation magnitude and direction. This algorithm works only with stationary or jointly moving camera sources.

Technique	**Spatial**	**Temporal**	**Ambiguities**	**Overlap**	**Source**	**Extra Processing**
Image-Based	Yes	No	Many	Yes	Stationary	Coarse-to-fine
Feature-Based	No*	Yes	Few	Yes	Jointly Moving	Feature Extraction
Direct-Based	Yes	Yes	No	Yes	Jointly Moving	Coarse-to-fine
Image Matching	Yes	Yes	No	Yes	Similarly Moving	Statistical Methods
Non-Overlapping	Yes	Yes	No	No	Jointly Moving	Transformations

7. Conclusion

A summary of the advantages and disadvantages of all the explored techniques is shown in Table 1. All of the methods described do not allow free movement of cameras, which is undesirable and less representative of the real world.

The focus of the implementation would be on the alignment of synchronized non-overlapping video with planar scenes, based on the non-overlapping technique described by Caspi and Irani [3]. The idea will be extended to allow free movement of video sources. The approach can be divided into two parts: initially determining the intercamera transformation with jointly moving sources followed by utilizing the inter-frame relations to recover the later transformations. The transformations between the successive frames will be found using any image alignment technique [4, 6]. A reference frame can be chosen in the initial part of the video which will be used to find the overall transformation between any two points from the two image sequences.

8. References

[1] Y. Caspi and M. Irani, "Spatio-Temporal Alignment," *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1409-1424, November 2002.

[2] P. Sand and S. Teller, "Video Matching," *Proc. ACM Transactions on Graphics*, pp. 592-599, August 2004.

[3] Y. Caspi and M. Irani, "Aligning non-overlapping sequences," *Proc. International Journal of Computer Vision*, pp. 39-51, June 2002.

[4] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, "Hierarchical Model-Based Motion Estimation", *Proc. European Conference on Computer Vision (ECCV),* pp. 237–252, May 1992.

[5] Y. Caspi and M. Irani, "A step towards sequence-to-sequence alignment," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 682–689, June 2000.

[6] R. Szeliski. "Video mosaics for virtual environments", *IEEE Computer Graphics and Applications*, pp. 22-30, March 1996.

[7] G. P. Stein, "Tracking from multiple view points: Self-calibration of space and time," *Proc. DARPA IU Workshop*, pp. 1037–1042, 1998.

[8] Y. Caspi and M. Irani, "Alignment of non-overlapping sequences," *Proc. International Conference on Computer Vision*, pp. 76-83, July 2001.

[9] L. Lee, R. Romano and G. Stein, "Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame," *Proc. IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 758-767, August 2000.

[10] L. Spencer and M. Shah, "Temporal Synchronization from Camera Motion," *Proc. Asian Conference on Computer Vision (ACCV)*, Jan 2004.
http://www.cs.ucf.edu/~vision/projects/time_shift/time_shift.htm