Feature-Based Sequence-to-Sequence Matching*

Yaron Caspi Denis Simakov Michal Irani
Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
76100 Rehovot, Israel

For a more detailed version of this paper see http://www.wisdom.weizmann.ac.il/∼vision/traj2traj.html

Image-to-image matching methods (e.g., [Faugeras et al. 2001; Hartley and Zisserman 2000; Xu and Zhang 1996; Bergen et al. 1992; Szeliski and Shum 1997; Zhang et al. 1995; Zoghlami et al. 1997]) are inherently restricted to the information contained in individual images, i.e., the spatial variations *within* image frames (which capture the scene appearance). But there are cases when there is not enough common spatial information within the two images to allow reliable image matching. One such example is illustrated in Fig. 1. The input images 1.a and 1.b contain a single object, but we want to match (or align) the entire frame. Alignment of image 1.a to image 1.b is not uniquely defined (see Fig. 1.c). However, a video sequence contains much more information than any individual frame does. In particular, a video sequence captures information about scene dynamics such as the trajectory of the moving object shown in Fig. 1.d and 1.e, which in this case provides enough information for unique alignment both in space and in time (see Fig. 1.f). The scene dynamics, exemplified here by trajectories of moving objects, is a property that is inherent to the scene, and is thus common to all sequences recording the same scene, even when taken from different video cameras. It therefore forms an *additional* or *alternative* powerful cue for matching video sequences.

The benefits of exploiting scene dynamics for matching sequences was noted before. Caspi and Irani [Caspi and Irani 2000] described a direct (intensity-based) sequence-to-sequence alignment method. Their method is based on finding the space-time transformation which minimizes the intensity differences (SSD) between the two sequences, and was applied to cases where the spatial relation between the sequences could be modeled by a 2D parametric transformation (a homography). It was shown to be useful for addressing rigid as well as complex non-rigid changes in the scene (e.g., flowing water), and changes in illumination. However, that method does not apply when the two sequences have different appearance properties, such as with sensors of different sensing modalities, nor when the spatial transformation between the two sequences is very large, such as in wide base-line matching, or in large differences in zoom.

This paper illustrates a feature-based approach for space-time matching of video sequences. The "features" in our method are space-time trajectories constructed from moving objects. This approach can recover the 3D epipolar geometry between sequences recorded by widely separated video cameras, and can handle significant differences in appearance between the two sequences.

The advantage of this approach over using regular feature-based image-to-image matching is illustrated in Fig. 2. This figure shows two sequences recording several small moving objects. Each feature point in the image-frame of Fig. 2.a (denoted by A-E) can in principle be matched to any other feature point in the image-frame of Fig. 2.b. There is no sufficient information in any individual frame to uniquely resolve the point correspondences. Point trajectories, on the other hand, have additional shape properties which simplify the *trajectory* correspondence problem (i.e., which trajectory corresponds to which trajectory) across the two sequences, as shown in Fig. 2.c and 2.d.

Stein [Stein 1998] and Lee et.al. [Lee et al. 2000] described a method for estimating a time shift and a homography between two sequences based on alignment of centroids of moving objects. However, in [Stein 1998; Lee et al. 2000] the centroids were treated as an *unordered* collection of feature points and not as trajectories. In contrast, we enforce correspondences between *trajectories*, thus avoiding the combinatorial complexity of establishing point matches of all points in all frames, resolving ambiguities in point correspondences, and allowing for temporal correspondences at *sub-frame* accuracy. This is not possible when the points are treated independently (i.e., as a "cloud of points").

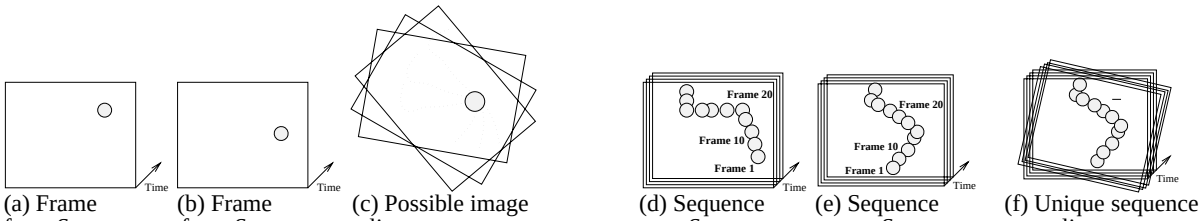
Our algorithm for recovering correspondences between trajectories across the two sequences is briefly described next. However, the ideas presented in this paper are not limited to this particular implementation.

Implementation:
Our current implementation is an extension of standard feature-based image matching methods (see examples of RANSAC/LMS-based methods in [Hartley and Zisserman 2000; Xu and Zhang 1996]). The first (and crucial) difference is that we use *trajectories* instead of *points* as our features. Since one trajectory consists of many points, therefore a single trajectory match induces multiple point matches (consequently, reducing the complexity of matching and increasing robustness in presence of errors – see "Benefits of the Approach").

A matching pair of 2D trajectory-features should correspond to projections of the same 3D trajectory of some 3D point. This 3D point need not be visible in the images (it can be real or virtual). For example, in our experiments we tracked moving objects (using background subtraction method) and extracted specific points on their blobs (e.g., the object centroid, or the highest point on the object silhouette, etc). The accuracy of approximating (real or virtual) 3D points from such 2D points on silhouettes is discussed in [Lee et al. 2000] and [Wong and Cipolla 2001].

The second difference (from standard feature based image matching implementations) is that we also deal with the temporal dimension to recover temporal matching as well. Schematically, the algorithm operates as follows: it searches in the space of possible trajectory correspondences (by a robust method, such as RANSAC or LMS). Each candidate trajectory correspondence is used for estimating spatial (homography H or fundamental matrix F) and temporal (Δt) parameters by iterating the following two steps:
(i) Fix Δt and approximate H (or F) using standard methods.
(ii) Fix H (or F) and refine Δt by fitting the best linear interpolation

*This work was partially supported by the European Commission (VIBES Project IST-2000-26001).



(a) Frame from S_1 (b) Frame from S_2 (c) Possible image alignments (d) Sequence S_1 (e) Sequence S_2 (f) Unique sequence alignment

Figure 1: **Spatial ambiguities in image-to-image alignment** (a) and (b) show two temporally corresponding frames 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).

(a)Frame from S_1 (b) Frame from S_2 (c) Sequence S_1 (d) Sequence S_2

Figure 2: **Point correspondences vs. trajectory correspondences.** (a) and (b) display two frames out of two sequences recording five small moving objects (marked by A,B,C,D,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, which simplify the correspondence problem across two sequences (both in space and in time).

value (we allow for sub-frame time shifts).
We then choose the spatial (H or F) and temporal (Δt) candidate parameters which minimize the overall error. For more details see the long paper.

Benefits of the Approach:
(i) Trajectory matching requires only a single correct "feature" (i.e., trajectory) correspondence, as opposed to 8 feature (point) correspondences as in regular image-to-image matching (for estimating the fundamental matrix). This provides a significant benefit in RANSAC-like matching algorithms when the probability to select at random a sample of eight correct point correspondences is low. Such cases occur in wide-baseline scenarios where the range of valid disparities is very large. A complete analysis of the complexity reduction due to the smaller number of required "feature" (trajectory) matches may be found in the longer version of this paper.
(ii) Since trajectory-features can be constructed from "virtual 3D points" our method can address cases where the cameras never image the same scene points (e.g., when the cameras are on opposite sides of the scene, such as in Fig. 5).
(iii) Often corresponding feature points do not have similar appearance properties across cameras such as in the case of multi-sensor modalities (e.g., Fig 3), or in significantly different zooms (Fig. 4). Yet, their trajectories share common geometric/shape properties that facilitate the matching (e.g., see Fig. 2) even when the appearance properties are different .
(iv) Unsynchronized video sequences can be temporally matched (synchronized) at *sub-frame* accuracy. Such sub-frame synchronization gives rise to new video applications including super-resolution in time [Shechtman et al. 2002].
(v) Sub-frame temporal alignment also provides higher accuracy in the spatial matching. Image-to-image matching is restricted to matching of existing physical image frames. However, when "corresponding" frames in time across the two sequences have not been recorded at exactly the same time (due to a *sub-frame* temporal misalignment between the two sequences), this leads to inaccuracies in the spatial matching (fundamental matrix or homography). Sequence-to-sequence matching, on the other hand, is not restricted to physical ("integer") image frames.

Examples:
(i) Multi-sensor alignment: Fig. 3 shows results of aligning sequences obtained by two cameras of different sensing modalities. Fig. 3.(a) and 3.(b) display representative frames from a PAL *visible light* sequence and an NTSC *Infra-Red* sequence, respectively. The scene contains several moving objects: 2 kites, 2 moving cars, and sea waves. The trajectories induced by tracking the moving objects are displayed in 3.(c). The two camera centers were close to each other, therefore the spatial transformation was modeled by a homography. The output after spatio-temporal alignment via trajectory matching is displayed in 3.(d). The recovered temporal misalignment was 1.31 sec. The results are displayed after fusing the two input sequences (using Burt's fusion algorithm [Burt and Kolczynski 1993]). We can now clearly observe spatial features from both sequences. In particular note the right kite which is more clearly visible in the visible-light sequence, and the left kite which is more clearly visible in the IR sequence (both marked by circles).
(ii) Matching across significant zoom differences: Fig. 4 shows an example of aligning sequences obtained at significantly different zooms. Fig. 4.(a) and 4.(b) display two representative frames from the reference sequence and second sequence, showing a ball thrown from side to side. The sequence in column 4.(a) was captured by a wide field-of-view camera, while the sequence in column 4.(b) was captured by a narrow field-of-view camera. The cameras where located next to each other (the spatial transformation was modeled by a homography) and the ratio in zooms was approximately $1:3$. The two sequences capture features at significantly different spatial resolutions, which makes the problem of inter-camera image-to-image alignment very difficult. The dynamic information (the trajectory of the ball's center of gravity), on the other hand, forms a powerful cue for alignment both in time and in space. Column 4.(c) displays superposition of corresponding frames *after* spatio-temporal alignment. The dark pink boundaries in

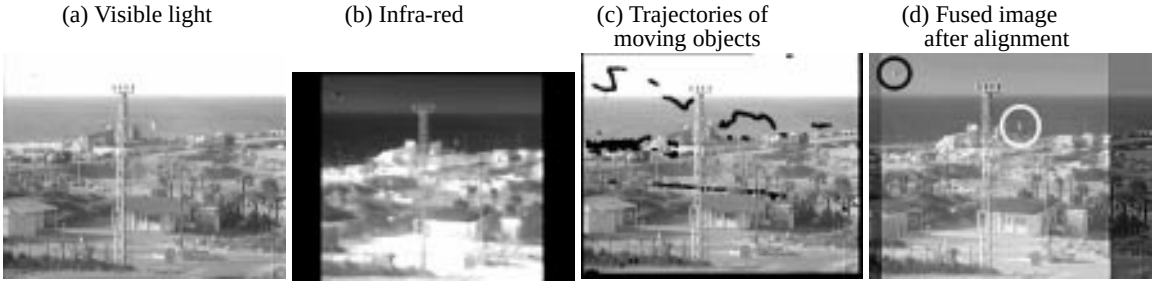


Figure 3: **Multi-Sensor Alignment** (see text).

4.(c) correspond to scene regions observed only by the reference (zoomed-out) camera.

(iii) Wide base-line matching: Fig. 5 shows an example of recovering the fundamental matrix using two cameras situated on opposite sides of the scene (i.e., the cameras are facing each other). Figs 5.(a) and 5.(b) display two representative frames from two sequences.Each camera is visible by the other camera and is circled and marked by a white arrow. Space-time trajectories induced by moving objects (ball and players) are displayed in 5.(c)-(d) in different colors for the different objects. The feature points that correspond to the current frame are marked in yellow. The recovered epipolar geometry is displayed in 5.(e) and 5.(f). Points and their epipolar lines are displayed in each image for verification. Note, that the only static objects that are visible in both views are the basket ring and the board. Accuracy of the recovered spatial alignment can be appreciated by the closeness of each point to the epipolar line of its corresponding point, as well as by comparing the intersection of epipolar lines with the ground truth epipole marked by a cross (which is the other camera). In this example the relative blob size of the moving objects was used to provide initial correspondence between the trajectories across the two sequences. Two trajectories (instead of one) were used on each RANSAC iteration, as most trajectories are planar. An initial temporal alignment with accuracy within one second (25 frames) was manually provided, and the final recovered temporal shift was 3.69 frames.

Summary:
We have shown that similar to [Caspi and Irani 2000] (where direct *intensity-based* image alignment was extended to *sequence alignment*), feature-based image matching can also be extended into trajectory-based sequence matching. This allows to address scenarios that are very difficult to solve otherwise.

For a more detailed version and example sequences see `www.wisdom.weizmann.ac.il/~vision/traj2traj.html`

References

BERGEN, J., ANANDAN, P., HANNA, K., AND HINGORANI, R. 1992. Hierarchical model-based motion estimation. In *European Conference on Computer Vision (ECCV)*, 237–252.

BURT, P., AND KOLCZYNSKI, R. 1993. Enhanced image capture through fusion. In *International Conference on Computer Vision (ICCV)*, 173–182.

CASPI, Y., AND IRANI, M. 2000. A step towards sequence-to-sequence alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 682–689.

FAUGERAS, O., LUONG, Q., AND PAPADOPOULO, T. 2001. *The Geometry of Multiple Images*. MIT Press.

HARTLEY, R., AND ZISSERMAN, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge university press, Cambridge.

LEE, L., ROMANO, R., AND STEIN, G. 2000. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 22*, Special Issue on Video Surveillance and Monitoring (August), 758–767.

SHECHTMAN, E., CASPI, Y., AND IRANI, M. 2002. Increasing video resolution in time and space. In *European Conference on Computer Vision (ECCV)*.

STEIN, G. P. 1998. Tracking from multiple view points: Self-calibration of space and time. In *DARPA IU Workshop*, 1037–1042.

SZELISKI, R., AND SHUM, H.-Y. 1997. Creating full view panoramic image mosaics and environment maps. In *Computer Graphics Proceedings, Annual Conference Series*, 251–258.

WONG, K.-Y. K., AND CIPOLLA, R. 2001. Structure and motion from silhouettes. In *International Conference on Computer Vision (ICCV)*, vol. II, 217–222.

XU, C., AND ZHANG, Z. 1996. *Epipolar Geometry in Stereo, Motion and Object Recognition*. Kluwer Academic Publishers, Dordecht, The Netherlands.

ZHANG, Z., DERICHE, R., FAUGERAS, O., AND LUONG, Q. 1995. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence 78*, 87–119.

ZOGHLAMI, I., FAUGERAS, O., AND DERICHE, R. 1997. Using geometric corners to build a 2d mosaic from a set of images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 420–425.

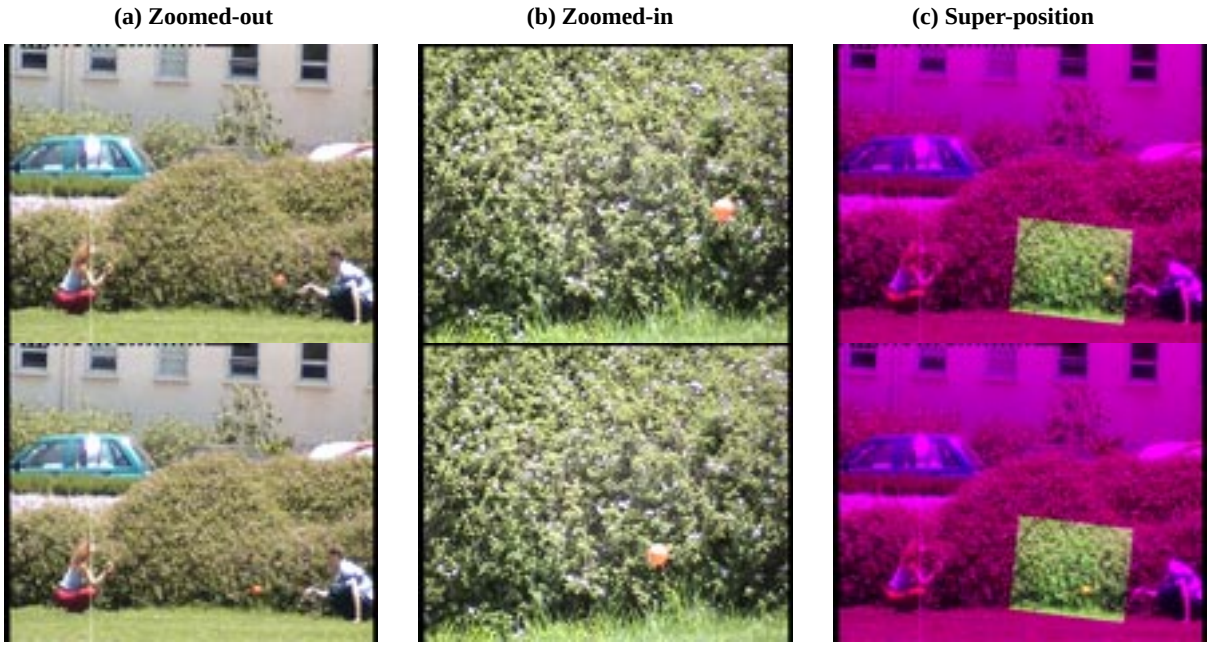


Figure 4: **Alignment of sequences obtained at different zooms** (see text).
For color sequences see `www.wisdom.weizmann.ac.il/~vision/traj2traj.html`

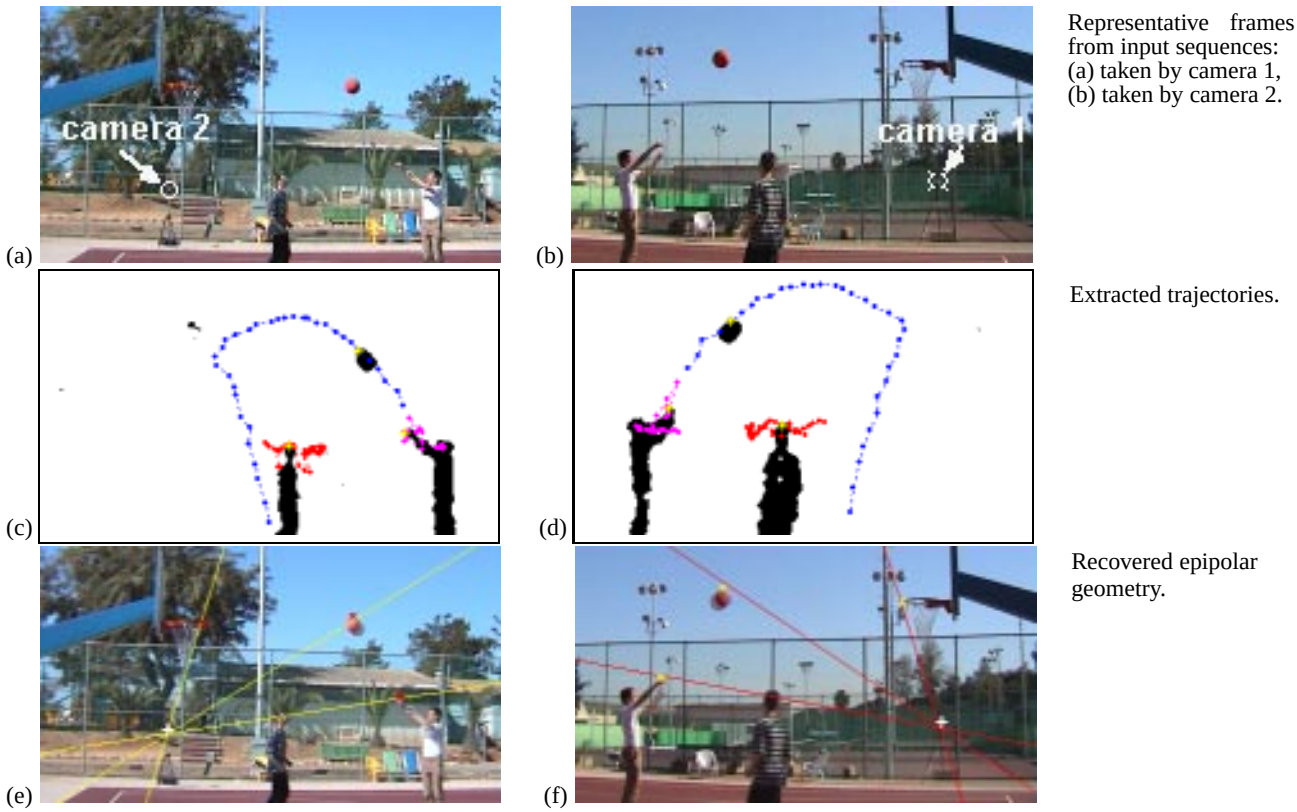


Figure 5: **Wide Base-Line Matching** (see text).
For color sequences see `www.wisdom.weizmann.ac.il/~vision/traj2traj.html`