



Aligning Non-Overlapping Sequences^{*,†}

YARON CASPI AND MICHAL IRANI

Department of Computer Science and Applied Math, The Weizmann Institute of Science, 76100 Rehovot, Israel

caspi@wisdom.weizmann.ac.il

irani@wisdom.weizmann.ac.il

Abstract. This paper shows how two image sequences that have no spatial overlap between their fields of view can be aligned both in time and in space. Such alignment is possible when the two cameras are attached closely together and are moved jointly in space. The common motion induces “similar” changes over time within the two sequences. This correlated temporal behavior, is used to recover the spatial and temporal transformations between the two sequences. The requirement of “consistent appearance” in standard image alignment techniques is therefore replaced by “consistent temporal behavior”, which is often easier to satisfy.

This approach to alignment can be used not only for aligning non-overlapping sequences, but also for handling other cases that are inherently difficult for standard image alignment techniques. We demonstrate applications of this approach to three real-world problems: (i) alignment of non-overlapping sequences for generating wide-screen movies, (ii) alignment of images (sequences) obtained at significantly different zooms, for surveillance applications, and, (iii) multi-sensor image alignment for multi-sensor fusion.

Keywords: spatio-temporal alignment, temporal synchronization, multi-sensor alignment, alignment for wide-screen movies, alignment across different zooms

1. Introduction

The problem of image alignment (or registration) has been extensively researched, and successful approaches have been developed for solving this problem. Some of these approaches are based on matching extracted local image features, other approaches are based on directly matching image intensities. A review of some of these methods can be found in Torr and Zisserman (1999) and Irani and Anandan (1999). However, all these approaches share one basic assumption: that there is sufficient overlap between the two images to allow extraction of common image properties, namely, that there is sufficient “similarity” between the two images (“Similarity” of images is used here in the broadest sense. It could range from

gray-level similarity, to feature similarity, to similarity of frequencies, and all the way to statistical similarity such as mutual information (Viola and Wells III, 1995).

In this paper the following question is addressed: *Can two images be aligned when there is very little similarity between them, or even more extremely, when there is no spatial overlap at all between the two images?* When dealing with individual images, the answer tends to be “No”. However, this is not the case when dealing with image sequences. An image sequence contains much more information than any individual frame does. In particular, temporal changes (such as dynamic changes in the scene, or the induced image motion) are encoded *between* video frames, but do not appear in any individual frame. Such information can form a powerful cue for alignment of two (or more) sequences. Caspi and Irani (2000) and Stein (1998) have illustrated an applicability of such an approach for aligning two sequences based on common dynamic scene information. However, they assumed that the same temporal changes in the scene (e.g., moving

^{*}A shorter version of this paper appeared in ICCV 2001 (Caspi and Irani, 2001).

[†]This work was supported by the Moross Laboratory for Vision and Motor Control.

objects) are visible to both video cameras, leading to the requirement that there must be significant overlap in the FOVs (fields-of-view) of the two cameras.

In this paper we show that when two cameras are attached closely to each other (so that their centers of projections are very close), and move jointly in space, then the induced frame-to-frame transformations *within* each sequence have correlated behavior *across* the two sequences. This is true even when the sequences have no spatial overlap. This correlated temporal behavior is used to recover both the spatial and temporal transformations between the two sequences.

Unlike carefully calibrated stereo-rigs (Slama, 1980), our approach does not require any prior internal or external camera calibration, nor any sophisticated hardware. Our approach bears resemblance to the approaches suggested by Demirdijian et al. (2000) Horaud and Csurka (1998) and Zisserman et al. (1995) for auto-calibration of stereo-rigs. But unlike these methods, we do not require that the two cameras observe and match the same scene features, nor that their FOVs will overlap.

The need for “consistent appearance”, which is a fundamental assumption in image alignment or calibration methods, is replaced here with the requirement of “consistent temporal behavior”. Consistent temporal behavior is often easier to satisfy (e.g., by moving the two cameras jointly in space). A similar idea was used for “hand-eye calibration” in robotics research (e.g., Tsai and Lenz (1989) and Horaud and Dornaika (1995)).

Our approach is useful not only in the case of non-overlapping sequences, but also in other cases where there is very little common appearance information between images, and are therefore inherently difficult for standard image alignment techniques. This gives rise to a variety of real-world applications, including: (i) Multi-sensor alignment for image fusion. This requires accurate alignment of images (sequences) obtained by sensors of different sensing modalities (such as Infra-Red and visible light). Such images differ significantly in their appearance due to different sensor properties (Viola and Wells III, 1995). (ii) Alignment of images (sequences) obtained at different zooms. The problem here is that different image features are prominent at different image resolutions (Dufournaud et al., 2000). Alignment of a wide-FOV sequence with a narrow-FOV sequence is useful for detecting small zoomed-in objects in (or outside) a zoomed-out view of the scene. This can be useful in surveillance appli-

cations. (iii) Generation of wide-screen movies from multiple non-overlapping narrow FOV movies (such as in IMAX movies).

Our approach can handle such cases. Results are demonstrated in the paper on complex real-world sequences, as well as on manipulated sequences with ground truth.

2. Problem Formulation

We examine the case when two video cameras having (approximately) the same center of projection but different 3D orientation, move jointly in space (see Fig. 1). The fields of view of the two cameras do not necessarily overlap. The internal parameters of the two cameras are different and unknown, but fixed along the sequences. The external parameters relating the two cameras (i.e., the relative 3D orientation) are also unknown but fixed. Let $S = I_1, \dots, I_{n+1}$ and $S' = I'_1, \dots, I'_{m+1}$ be the two sequences of images recorded by the two cameras.¹ When temporal synchronization (e.g., time stamps) is not available, then I_i and I'_i may not be corresponding frames in time. Our goal is to recover the transformation that aligns the two sequences both in time and in space. Note the term “alignment” here has a broader meaning than the usual one, as the sequences may not overlap in space, and may not be synchronized in time. Here we refer to alignment as displaying one sequence in the spatial coordinate system of the other sequence, and at the correct time shift, as if obtained by the other camera.

When the two cameras have the same center of projection (and differ only in their 3D orientation and their internal calibration parameters), then a simple fixed homography H (a 2D projective transformation) describes the *spatial* transformation between *temporally*

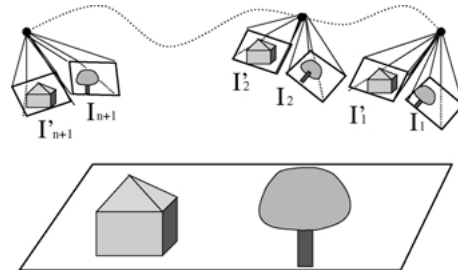


Figure 1. Two video cameras are attached to each other, so that they have the same center of projection, but non-overlapping fields-of-view. The two cameras are moved jointly in space, producing two separate video sequences I_1, \dots, I_{n+1} and I'_1, \dots, I'_{m+1} .

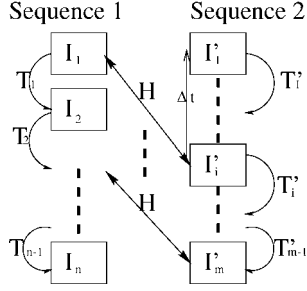


Figure 2. Problem formulation. The two sequences are spatially related by a fixed but unknown inter-camera homography H , and temporally related by a fixed and unknown time shift Δt . Given the frame-to-frame transformations T_1, \dots, T_n and T'_1, \dots, T'_m , we want to recover H and Δt .

corresponding pairs of frames across the two sequences (Hartley and Zisserman, 2000).

If there were enough common features (e.g., p and p') between temporally corresponding frames (e.g., I_i and I'_i), then it would be easy to recover the inter-camera homography H , as each such pair of corresponding image points would provide linear constraints on H : $p' \cong Hp$. This, in fact, is how most image alignment techniques work (Hartley and Zisserman, 2000). However, this is not the case here. The two sequences do not share common features, because there is no spatial overlap between the two sequences. Instead, the homography H is recovered from the induced frame-to-frame transformations *within* each sequence.

Let T_1, \dots, T_n and T'_1, \dots, T'_m be the sequences of frame-to-frame transformations within the video sequences S and S' , respectively. T_i is the transformation relating frame I_i to I_{i+1} . These transformations can be either 2D parametric transformations (e.g., homographies or affine transformations) or 3D transformations/relations (e.g., fundamental matrices). We next show how we can recover the spatial transformation H and the temporal shift Δt between the two video sequences directly from the two sequences of transformations T_1, \dots, T_n and T'_1, \dots, T'_m . The problem formulated above is illustrated in Fig. 2.

3. Recovering Spatial Alignment Between Sequences

Let us first assume that the temporal synchronization is known. Such information is often available (e.g., from time stamps encoded in each of the two sequences). Section 4 shows how we can recover the temporal shift between the two sequences when that informa-

tion is not available. Therefore, without loss of generality, it is assumed that I_i and I'_i are corresponding frames in time in sequences S and S' , respectively. Two cases are examined: (i) The case when the scene is planar or distant from the cameras. We refer to these scenes as “2D scenes”. In this case the frame-to-frame transformations T_i can be modeled by homographies (Section 3.1). (ii) The case of a non-planar scene. We refer to these scenes as “3D scenes”. In this case the frame-to-frame relations can be modeled by fundamental matrices (Section 3.2).

3.1. Planar or Distant (2D) Scenes

When the scene is planar or distant from the cameras, or when the joint 3D translations of the two cameras are negligible relative to the distance of the scene, then the induced image motions within each sequence (i.e., T_1, \dots, T_n and T'_1, \dots, T'_n) can be described by 2D parametric transformations (Hartley and Zisserman, 2000). T_i thus denotes the homography between frame I_i and I_{i+1} , represented by 3×3 non-singular matrix. We next show that temporally corresponding transformations T_i and T'_i are related by the same fixed inter-camera homography H (which relates frames I_i and I'_i).

Let P be a 3D point in the planar (or the remote) scene. Denote by p_i and p'_i its image coordinates in frames I_i and I'_i , respectively (the point P need not be visible in the two frames, i.e., P need not be within the FOV of the cameras). Let p_{i+1} and p'_{i+1} be its image coordinates in frames I_{i+1} and I'_{i+1} , respectively. Then, $p_{i+1} \cong T_i p_i$ and $p'_{i+1} \cong T'_i p'_i$. Because the coordinates of the video sequences S and S' are related by a fixed homography H , then: $p' \cong Hp$ and $p'_{i+1} \cong Hp_{i+1}$. Therefore:

$$HT_i p_i \cong Hp_{i+1} \cong p'_{i+1} \cong T'_i p'_i \cong T'_i H p_i \quad (1)$$

Each p_i could theoretically have a different scalar associated with the equality in Eq. (1). However, it is easy to show that because the relation in Eq. (1) holds for *all* points p_i , therefore all these scalars are equal, and hence:

$$HT_i \cong T'_i H. \quad (2)$$

Because H is non-singular we may write $T'_i \cong HT_i H^{-1}$, or

$$T'_i = s_i HT_i H^{-1} \quad (3)$$

where s_i is a (frame-dependent) scale factor. Equation (3) is true for all frames, i.e., for any pair of corresponding transformations T_i and T'_i ($i = 1..n$) there exists a scalar s_i such that $T'_i = s_i H T_i H^{-1}$. It shows that there is a *similarity relation*² (or a “conjugacy relation”) between the two matrices T_i and T'_i (up to a scale factor). A similar observation was made for case of hand-eye calibration (e.g., Tsai and Lenz (1989) and Horaud and Dornaika (1995)), and for auto-calibration of a stereo-rig (e.g. Zisserman et al. (1995)).

Denote by $eig(A) = [\lambda_1, \lambda_2, \lambda_3]^T$ a 3×1 vector containing the eigenvalues of a 3×3 matrix A (in decreasing order). Then it is known (Pearson (1983) p. 898.) that: (i) If A and B are similar (conjugate) matrices, then they have the same eigenvalues: $eig(A) = eig(B)$, and, (ii) The eigenvalues of a scaled matrix are scaled: $eig(sA) = s(eig(A))$. Using these two facts and Eq. (3) we obtain:

$$eig(T'_i) = s_i eig(T_i) \quad (4)$$

where s_i is the scale factor defined by Eq. (3). Equation (4) implies that the two vectors $eig(T_i)$ and $eig(T'_i)$ are “parallel”. This gives rise to a measure of similarity between two matrices T_i and T'_i :

$$sim(T_i, T'_i) = \frac{eig(T_i)^T eig(T'_i)}{\|eig(T_i)\| \|eig(T'_i)\|}, \quad (5)$$

where $\|\cdot\|$ is the vector norm. For real valued eigenvalues, Eq. (5) provides the cosine of the angle between the two vectors $eig(T_i)$ and $eig(T'_i)$. This property will be used later for obtaining the temporal synchronization between the two sequences (Section 4). This measure is also used for outlier rejection of bad frame-to-frame transformation pairs, T_i and T'_i (Section 6.3). The remainder of this section explains how the fixed inter-camera homography H is recovered from the list of frame-to-frame transformations T_1, \dots, T_n and T'_1, \dots, T'_n .

For each pair of temporally corresponding transformations T_i and T'_i in sequences S and S' , we first compute their eigenvalues $eig(T_i)$ and $eig(T'_i)$. The scale factor s_i which relates them is then estimated from Eq. (4) using least squares minimization (three equations, one unknown).³ Once s_i is estimated, Eq. (3) (or Eq. (2)) can be rewritten as:

$$s_i H T_i - T'_i H = 0 \quad (6)$$

Eq. (6) is linear in the unknown components of H . Rearranging the components of H in a 9×1 column vector

$\vec{h} = [H_{11} H_{12} H_{13} H_{21} H_{22} H_{23} H_{31} H_{32} H_{33}]^T$, Eq. (6) can be rewritten as a set of linear equations in \vec{h} :

$$M_i \vec{h} = \vec{0} \quad (7)$$

where M_i is a 9×9 matrix defined by T_i , T'_i and s_i :

$$M_i = \begin{bmatrix} s_i T_i^t - T'_{i11} I & -T'_{i12} I & -T'_{i13} I \\ -T'_{i21} I & s_i T_i^t - T'_{i22} I & -T'_{i23} I \\ -T'_{i31} I & -T'_{i32} I & s_i T_i^t - T'_{i33} I \end{bmatrix}_{9 \times 9}$$

where I is the 3×3 identity matrix. Equation (7) implies that each pair of corresponding transformations T_i and T'_i contributes 9 linear constraints in the unknown homography H (i.e., \vec{h}), out of which at most 6 constraints are linearly independent (see Section 6). Therefore, in theory, at least two such pairs of independent transformations are needed to uniquely determine the homography H (up to a scale factor). In practice, we use all available constraints from all pairs of transformations to compute H . The constraints from all the transformations T_1, \dots, T_n and T'_1, \dots, T'_n can be combined into a single set of linear equations in \vec{h} :

$$A \vec{h} = \vec{0} \quad (8)$$

where A is a $9n \times 9$ matrix: $A = \begin{bmatrix} M_1 \\ \vdots \\ M_n \end{bmatrix}$. Equation (8) is a homogeneous set of linear equations in \vec{h} , that can be solved in a variety of ways (Bjorck, 1996). In particular, \vec{h} may be recovered up to scale by computing the eigenvector which corresponds to the smallest eigenvalue of the matrix $A^T A$.

3.2. 3D Scenes

When the scene is neither planar nor distant, the relation between two consecutive frames of an uncalibrated camera is described by the fundamental matrix (Hartley and Zisserman, 2000). In this case the input to our algorithm is two sequences of fundamental matrices between successive frames, denoted by F_1, \dots, F_n and F'_1, \dots, F'_n . Namely, if $p_i \in I_i$ and $p_{i+1} \in I_{i+1}$ are corresponding image points, then: $p_{i+1}^T F_i p_i = 0$. Although the relations *within* each sequence are characterized by fundamental matrices, the inter-camera transformation remains a homography H . This is because the two cameras still share the same center of projection (Section 2).

Each fundamental matrix F_i can be decomposed into a homography + epipole (Hartley and Zisserman,

2000) as follows:

$$F_i = [e_i]_{\times} T_i$$

where e_i is the epipole relating frames I_i and I_{i+1} , the matrix T_i is the induced homography from I_i to I_{i+1} via any plane (real or virtual). $[\cdot]_{\times}$ is the cross product matrix ($[v]_{\times} \vec{w} = \vec{v} \times \vec{w}$).

The homographies, T_1, \dots, T_n and T'_1, \dots, T'_n , and the epipoles e_1, \dots, e_n and e'_1, \dots, e'_n , impose separate constraints on the inter-camera homography H . These constraints can be used separately or jointly to recover H .

3.2.1. Homography-Based Constraints. The homographies T_1, \dots, T_n and T'_1, \dots, T'_n (extracted from the fundamental matrices F_1, \dots, F_n and F'_1, \dots, F'_n , respectively), may correspond to different 3D planes. In order to apply the algorithm of Section 3.1 using these homographies, we need to impose plane-consistency across the two sequences (to guarantee that temporally corresponding homographies correspond to the same plane in the 3D world). One possible way for imposing plane-consistency across (and within) the two sequences is by using the “Plane + Parallax” approach (Kumar et al., 1994; Irani et al., 1998; Shashua and Navab, 1994; Sawhney, 1994). However, this approach requires that a real physical planar surface be visible in *all* video frames. Alternatively, the “threading” method of Avidan and Shashua (1998) or other methods for computing consistent set of camera matrices (e.g., Beardsley et al., 1998), can impose plane-consistency within each sequence, even if no real physical plane is visible in any of the frames. Plane consistency *across* the two sequences can be obtained, e.g., if Avidan and Shashua (1998) is initiated at frames which are known to simultaneously view the same real plane in both sequences. This can be done even if the two cameras see different portions of the plane (allowing for non-overlapping FOVs), and do not see that plane at any of the other frames. This approach is therefore less restrictive than the Plane + Parallax approach.

3.2.2. Epipole-Based Constraints. The fundamental matrices $F_1 \dots F_n$ and $F'_1 \dots F'_n$ also provide a list of epipoles e_1, \dots, e_n and e'_1, \dots, e'_n . These epipoles are uniquely defined (there is no issue of plane consistency here).

Since the two cameras have the same center of projection, then for any frame i : $e'_i \cong H e_i$, or more

specifically:

$$(e'_i)_x = \frac{[h_1 h_2 h_3] e_i}{[h_7 h_8 h_9] e_i} \quad (e'_i)_y = \frac{[h_4 h_5 h_6] e_i}{[h_7 h_8 h_9] e_i} \quad (9)$$

Multiplying by the dominator and rearranging terms yields two new linear constraints on H for every pair of corresponding epipoles e_i and e'_i :

$$\begin{bmatrix} e'_i & \vec{0}^t - (e'_i)_x e'_i \\ \vec{0}^t & e'_i - (e'_i)_y e'_i \end{bmatrix}_{2 \times 9} \vec{h} = 0 \quad (10)$$

where $\vec{0}^t = [0, 0, 0]$. Every pair of temporally corresponding epipoles, e_i and e'_i , thus imposes two linear constraints on H . These $2n$ constraints ($i = 1, \dots, n$) can be added to the set of linear equations in Eq. (8) which are imposed by the homographies. Alternatively, the epipole-related constraints can be used *alone* to solve for H , thus avoiding the need to enforce plane-consistency on the homographies. Theoretically, four pairs of corresponding epipoles e_i and e'_i in general position (no 3 on the same line) are sufficient.

4. Recovering Temporal Synchronization Between Sequences

So far we have assumed that the temporal synchronization between the two sequences is known and given. Namely, that frame I_i in sequence S corresponds to frame I'_i in sequence S' , and therefore the transformation T_i corresponds to transformation T'_i . Such information is often available from time stamps. However, when such synchronization is not available, we can recover it. Given two unsynchronized sequences of transformations T_1, \dots, T_n and T'_1, \dots, T'_m , we wish to recover the unknown temporal shift Δt between them. Let T_i and $T'_{i+\Delta t}$ be temporally corresponding transformations (namely, they occurred at the same time instance). Then from Eq. (4) we know that they should satisfy $\text{eig}(T_i) \parallel \text{eig}(T'_{i+\Delta t})$ (i.e., the 3×1 vectors of eigenvalues should be parallel). In other words, the similarity measure $\text{sim}(T_i, T'_{i+\Delta t})$ of Eq. (5) should equal 1 (corresponding to $\cos(0)$, i.e., an angle of 0° between the two vectors). All pairs of corresponding transformations T_i and $T'_{i+\Delta t}$ must simultaneously satisfy this constraint for the correct time shift Δt . Therefore, we recover the unknown temporal time shift Δt by maximizing the following objective function:

$$\text{SIM}(\Delta t) = \sum_i \text{sim}(T_i, T'_{i+\Delta t})^2 \quad (11)$$

The maximization is currently performed by an exhaustive search over a finite range of valid time shifts Δt . To address larger temporal shifts, we apply a hierarchical search. Coarser temporal levels are constructed by composing transformations to obtain fewer transformation between more distant frames.

The objective function of Eq. (11) can be generalized to handle sequences of different frame rates, such as sequences obtained by NTSC cameras (30 frame/sec) vs. PAL cameras (25 frames/sec). The ratio between frames corresponding to equal time steps in the two sequences is $25:30 = 5:6$. Therefore, the objective function that should be maximized for an NTSC-PAL pair of sequences is:

$$SIM(\Delta t) = \sum_i \text{sim} \left(T_{5i}^{s(i+1)}, T_{6i+\Delta t}^{6(i+1)+\Delta t} \right)^2 \quad (12)$$

Where T_i^j is the transformation from frame I_i to frame I_j . In our experiments, all sequences were obtained by PAL video cameras. Therefore only the case of equal frame-rate (Eq. (11)) was experimentally verified. We found this method to be very robust. It successfully recovered the temporal shift up to *field* (half-frame) accuracy. Sub-field accuracy may be further recovered by interpolating the values of $SIM(\Delta t)$ obtained at discrete time shifts.

5. Applications

This section illustrates the applicability of our method to solving some real-world problems, which are particularly difficult for standard image alignment techniques. These include: (i) Alignment of non-overlapping sequences for generation of wide-screen movies from multiple narrow-screen movies (such as in IMAX films), (ii) Alignment of sequences obtained at significantly different zooms (e.g., for surveillance applications), and (iii) Alignment of multi-sensor sequences for multi-sensor fusion. We show results of applying the method to complex real-world sequences. All sequences which we experimented with, were captured by “off-the-shelf” consumer CCD cameras. The cameras were attached to each other to minimize the distance between their centers of projections. The joint camera motion was performed manually (i.e., a person would manually hold and rotate the two attached cameras). No temporal synchronization tool was used.

The frame-to-frame input transformations within each sequence (homographies T_1, \dots, T_n and

T'_1, \dots, T'_n) were extracted using the method described in Irani et al. (1994). Inaccurate frame-to-frame transformations T_i are pruned out by using two outlier detection mechanisms (see Section 6.3). The input sequences were usually several seconds long to guaranty significant enough motion. The temporal time shift was recovered using the algorithm described in Section 4 up to field accuracy. Finally, the *best* thirty or so transformations were used in the estimation of the inter-camera homography H (using the algorithm described in Section 3.1).

5.1. Alignment of Non-Overlapping Sequences

Figure 3 shows an example of alignment of non-overlapping sequences. The left camera is zoomed-in and rotated relative to the right camera. The correct spatio-temporal alignment can be seen in Fig. 3(c). Note the accurate alignment of the running person both in time and in space.

Our approach to sequence alignment can be used to generate wide-screen movies from two (or more) narrow field-of-view movies (such as in IMAX movies). Such an example is shown in Fig. 4. To verify the accuracy of alignment (both in time and in space), we allowed for a very small overlap between the two sequences. However, this image region was *not* used in the estimation process, to imitate the case of truly *non-overlapping* sequences. The overlapping region was used only for display and verification purposes. Figure 4(c) shows the result of combining the two sequences (by averaging corresponding frames) after spatio-temporal alignment. Note the accurate spatial as well as temporal alignment of the soccer player in the averaged overlapping region.

5.2. Alignment of Sequences Obtained at Different Zooms

Often in surveillance applications two cameras are used, one with a wide FOV (field-of-view) for observing large scene regions, and the other camera with a narrow FOV (zoomed-in) for detecting small objects. Matching two such images obtained at significantly different zooms is a difficult problem for standard image alignment methods, since the two images display different features which are prominent at the different resolutions. Our sequence alignment approach may be used for such scenarios. Figure 5 shows three such



Figure 3. Alignment of non-overlapping sequences. (a) and (b) are temporally corresponding frames from sequences S and S' . The correct time shift was automatically detected. (c) shows one frame in the combined sequence after spatio-temporal alignment. Note the accuracy of the spatial and temporal alignment of the running person. For color sequences see www.wisdom.weizmann.ac.il/NonOverlappingSeqs.

examples. The results of the spatio-temporal alignment (right column of Fig. 5) are displayed in the form of averaging temporally corresponding frames after alignment according to the computed homography and the computed time shift. In the first example (top row of Fig. 5) the zoom difference between the two cameras was approximately 1:3. In the second example (second row) it was $\approx 1:4$, and in the third example (bottom row) it was $\approx 1:8$. Note the small red flowers in the zoomed view (Fig. 5.2.(b)), that can barely be seen in the corresponding low resolution wide-view frame (Fig. 5.2 (a)). The same holds for the Pagoda in Fig. 5.3 (b).

5.3. Multi-Sensor Alignment

Images obtained by sensors of different modalities, e.g., IR (Infra-Red) and visible light, can vary significantly in their appearance. Features appearing in one image may not appear in the other, and vice versa. This poses a problem for image alignment methods. Our se-

quence alignment approach, however, does not require consistent appearance between the two sequences, and can therefore be applied to solve the problem. Figure 6 shows an example of two such sequences, one captured by a near IR camera, while the other by a regular video (visible-light) camera. The scene was shot in twilight. In the sequence obtained by the regular video camera (Fig. 6(a)), the outdoor scene is barely visible, while the inside of the building is clearly visible. The IR camera, on the other hand, captures the outdoor scene in great detail, while the indoor part (illuminated by “cold” neon light) was invisible to the IR camera (Fig. 6(b)). The result of the spatio-temporal alignment is illustrated by fusing temporally corresponding frames. The IR camera provides only intensity information, and was therefore fused only with the intensity (Y) component of the visible-light camera (using the image-fusion method of Burt and Kolczynski (1993)). The chrome components (I and Q) of the visible-light camera supply the color information.

The reader is encouraged to view color sequences at www.wisdom.weizmann.ac.il/NonOverlappingSeqs.

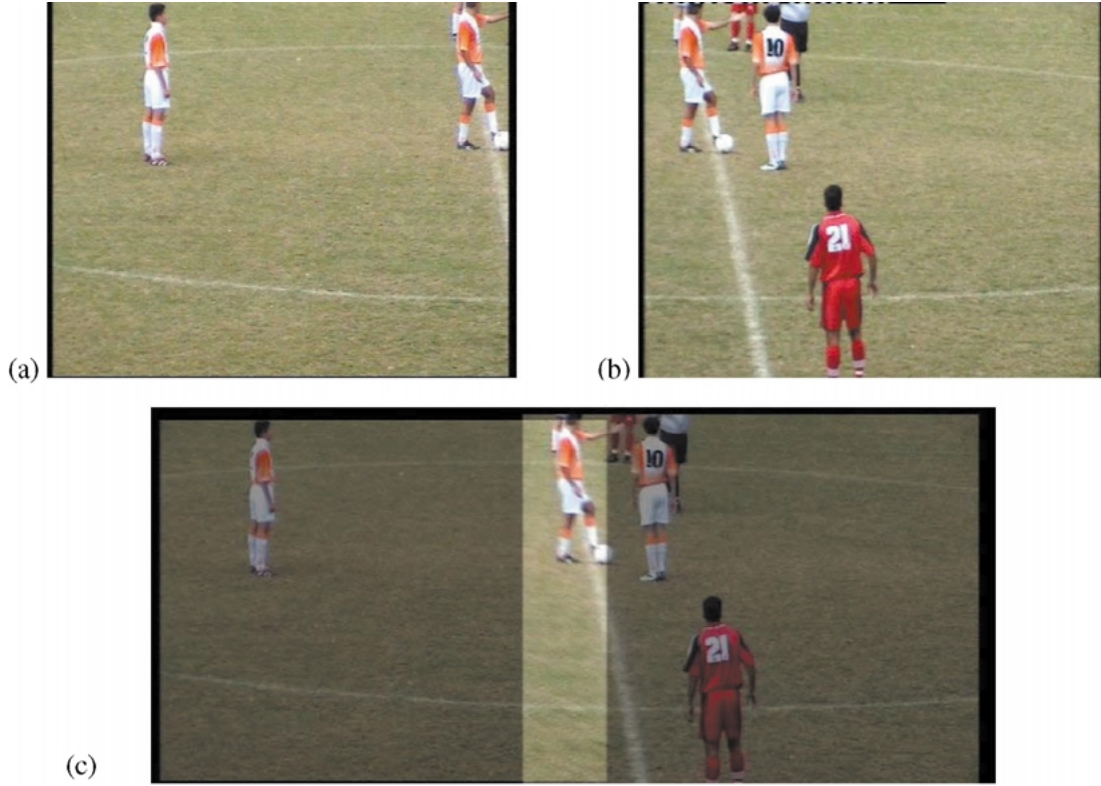


Figure 4. Wide-screen movies generation. (a) and (b) are temporally corresponding frames from sequences S and S' . The correct time shift was automatically detected. (c) shows one frame in the combined sequence. Corresponding video frames were averaged after spatio-temporal alignment. The small overlapping area was not used in the estimation process, but only for verification (see text). Note the accuracy of the spatial and temporal alignment of the soccer player in the overlapping region. For color sequences see www.wisdom.weizmann.ac.il/NonOverlappingSeqs.

6. Analysis

In this section we evaluated the effectiveness and stability, of the presented approach empirically (Section 6.1) and theoretically (Section 6.2) and numerically (Section 6.3).

6.1. Empirical Evaluation

In order to empirically verify the accuracy of our method, we took a real video sequence (see Fig. 7) and generated from it pairs of sequences with known (ground truth) spatial transformation H and temporal shift Δt . We then applied our algorithm and compared the recovered H and Δt with the ground truth.

For the case of non overlapping sequences, the real sequence of Fig. 7 was split in the middle, producing two non-overlapping sub-sequences of half-a-frame width each. The true (ground truth) homography H

therefore corresponds to a horizontal shift by the width of a halved frame (352 pixels), and Δt in this case is 0. The “inter-camera” homography H was recovered up to a misalignment error of less than 0.7 pixel over the entire image. The temporal shift ($\Delta t = 0$) was recovered accurately from the frame-to-frame transformations.

To empirically verify the accuracy of our method in the presence of large zooms and large rotations, we ran the algorithm on following three manipulated sequences with known (ground truth) manipulations: We warped the sequence of Fig. 7 (once by a zoom factor of 2, once by a zoom factor of 4, and once rotated it by 180°) to generate the second sequence.

The results are summarized in Table 1. The reported residual misalignment was measured as follows: The recovered homography was composed with the inverse of the ground-truth homography: $H_{true}^{-1}H_{recovered}$. Ideally, the composed homography should be the identity

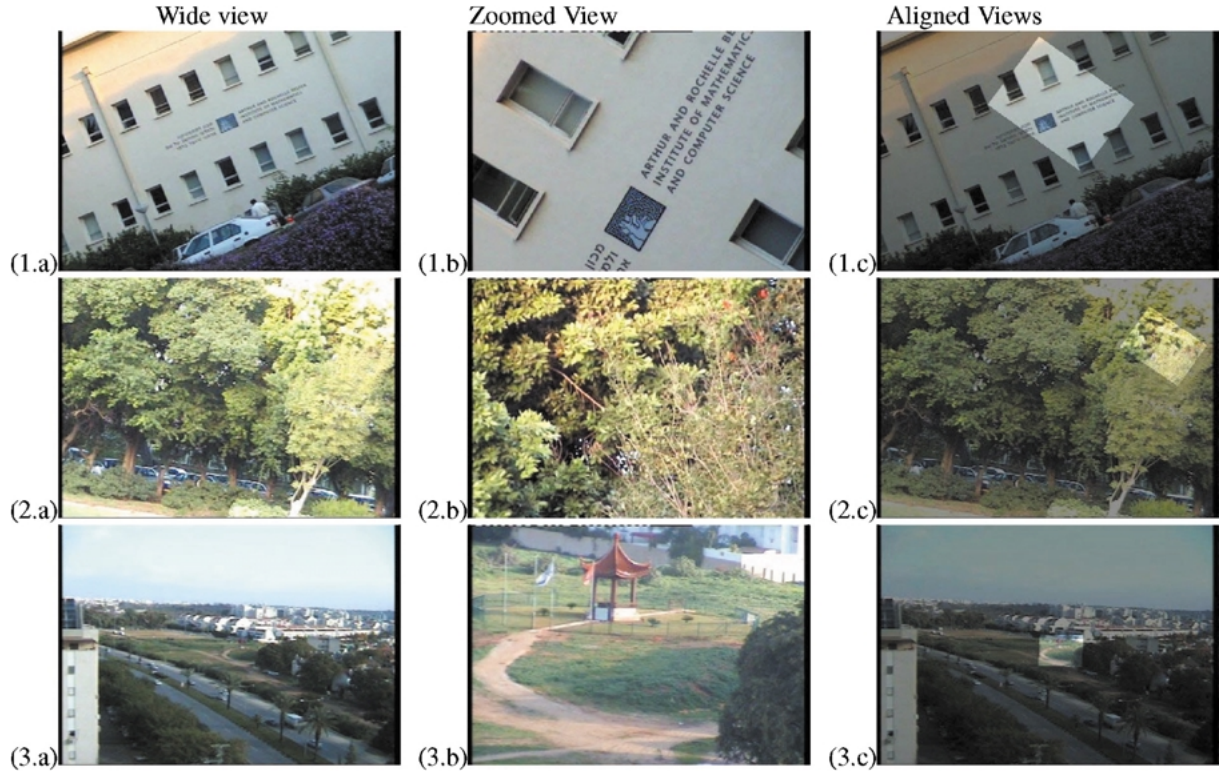


Figure 5. Finding zoomed region. This figure displays three different examples (one at each row), each one with different zoom factor. The left column (1.a, 2.a, 3.a) display one frame from each of the three wide-FOV sequences. The temporally corresponding frames from the corresponding narrow-FOV sequences are displayed in the center column (1.b, 2.b, 3.b). The correct time shift was automatically detected for each pair of narrow/wide FOV sequences. Each image on the right column shows super-position of corresponding frames of the two sequences after spatio-temporal alignment, displayed by color averaging (1.c, 2.c, 3.c). For color sequences see www.wisdom.weizmann.ac.il/NonOverlappingSeqs.

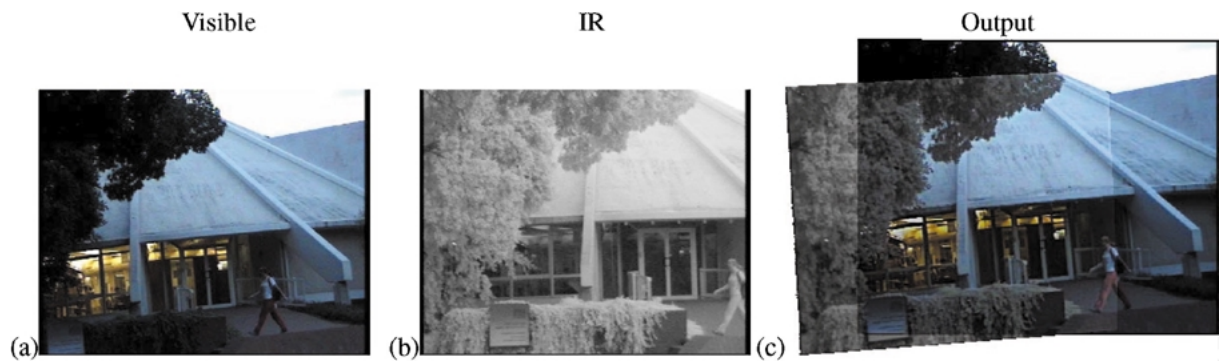


Figure 6. Multi-sensor alignment. (a) and (b) are temporally corresponding frames from the visible-light and near-IR sequences, respectively (the temporal alignment was automatically detected). The inside of the building is visible only in the visible-light sequence, while the IR sequence captures the details outdoors (e.g., the dark trees, the sign, the bush). (c) shows the results of fusing the two sequences after spatio-temporal alignment. The fused sequence preserves the details from both sequences. Note the high accuracy of alignment (both in time and in space) of the walking lady. For more details see text. For color sequences see www.wisdom.weizmann.ac.il/NonOverlappingSeqs.



Figure 7. The sequence used for empirical evaluation. (a, b, c) are three frames (0, 150, 300) out of the original 300 frames. This sequence was used as the base sequence for the quantitative experiments summarized in Table 1.

matrix. The errors reported in Table 1 are the *maximal* residual misalignment induced by the composed homography over the entire image. In all the cases the correct Δt was recovered (not shown in the table).

Table 1. Quantitative results.

Applied transformation	Recovered transformation	Max residual misalignment
Horizontal shift of 352 pixels	Horizontal shift of 351.6 pixels	0.7 pixels
Zoom factor = 2	Zoom factor = 1.9992	0.4 pixels
Zoom factor = 4	Zoom factor = 4.0048	0.4 pixels
Rotation by 180°	Rotation by 180.00°	0.01 pixels

This table summarizes the quantitative results with respect to ground truth. Each row corresponds to one experiment. In each experiment a real video sequence (Fig. 7) was warped (“manipulated”) by a known homography, to generate a second sequence. The left column describes the type of spatial transformation applied to the sequence, the center column describes the recovered transformation, and the right column describes the residual error between the ground-truth homography and the recovered homography (measured in maximal residual misalignment in the image space). In all 4 cases the correct temporal shift was recovered accurately. See text for further details.

6.2. Uniqueness of Solution

This section studies how many pairs of corresponding transformations T_i and T'_i are required in order to uniquely resolve the inter-camera homography H . To do so we examine the number of constraints imposed on H by a single pair of transformations via the similarity equation Eq. (3). Since we can extract the scale factor s_i directly from T_i and T'_i (see Section 3.1) we can omit the scale factor s_i and study the following question: How many constraints does an equation of the form $G = HBH^{-1}$ impose on H ? (e.g., $B = T_i$ and $G = T'_i$).⁴

The following notations are used: Denote by $\lambda_1, \lambda_2, \lambda_3$ the eigenvalues of the matrix B in decreasing order ($|\lambda_1| \geq |\lambda_2| \geq |\lambda_3|$). Denote by $\bar{\mathbf{u}}_{b_1}, \bar{\mathbf{u}}_{b_2}, \bar{\mathbf{u}}_{b_3}$ the corresponding eigenvectors with unit norm ($\|\bar{\mathbf{u}}_{b_1}\| = \|\bar{\mathbf{u}}_{b_2}\| = \|\bar{\mathbf{u}}_{b_3}\| = 1$). Denote by r_j the algebraic multiplicity⁵ of the eigenvalue λ_j , and denote by $V_j = \{\bar{\mathbf{v}} \in \mathbb{R}^n : B\bar{\mathbf{v}} = \lambda_j\bar{\mathbf{v}}\}$ the corresponding eigen subspace.

6.2.1. Basic Constraints. Similar (conjugate) matrices (e.g., B and G) have the same eigenvalues but different eigenvectors. Their eigenvectors are related by H . If \mathbf{u}_b is an eigenvector of B with corresponding eigenvalue λ , then $H\mathbf{u}_b$ is an eigenvector of G with the same eigenvalue λ : $G(H\mathbf{u}_b) = \lambda(H\mathbf{u}_b)$. The same holds for eigen subspaces. If V is an eigen subspace of B corresponding to an eigenvalue λ , then $H(V)$ is an eigen subspace of G with the same eigenvalue λ . We investigate the number of constraints imposed on H by B and G according to the dimensionality of their eigen subspaces. Let V be the eigen subspace corresponding to an eigenvector \mathbf{u}_b of B . We investigate three possible cases, one for each possible dimensionality of V , i.e., $\dim(V) = 1, 2, 3$.

Case I. $\dim(V) = 1$. This case mostly occurs when all three eigenvalues are distinct, but can also occur if some eigenvalues have algebraic multiplicity two or even three. In all these cases, V is spanned by the single eigenvector \mathbf{u}_b . Similarly $H(V)$ is spanned by the eigenvector \mathbf{u}_g of G . Therefore:

$$H\mathbf{u}_b = \alpha\mathbf{u}_g \quad (13)$$

with an unknown scale factor α . Equation (13) provides 3 linear equations in H and one new unknown α , thus in total it provides two new linearly independent constraints on H .

Case II. $\dim(V) = 2$. This occurs in one of the following two cases: (a) when there exists an eigenvalue with algebraic multiplicity two, or (b) when there is only one eigenvalue with algebraic multiplicity three, but the eigen subspace spanned by all eigenvectors has dimensionality of two.⁶ When $\dim(V) = 2$ then two eigenvectors span V (w.l.o.g., \mathbf{u}_{b1} and \mathbf{u}_{b2}). Then every linear combination of \mathbf{u}_{b1} and \mathbf{u}_{b2} is also an eigenvector of B with the same eigenvalue. Similarly, every linear combination of \mathbf{u}_{g1} and \mathbf{u}_{g2} is an eigenvector of G with the same eigenvalue. Therefore:

$$H\mathbf{u}_{bj} = \alpha_j \mathbf{u}_{g1} + \beta_j \mathbf{u}_{g2} \quad (14)$$

where α_j and β_j are unknown scalars ($j = 1, 2$). Hence, each of the two eigenvectors \mathbf{u}_{b1} and \mathbf{u}_{b2} provides 3 linear equations and 2 new unknowns. Therefore, in total, together they provide 2 new linear constraints on H .

Case III. $\dim(V) = 3$. In this case any vector is an eigenvector (all with the same eigenvalue λ). This is the case when $B \cong G \cong \lambda I$ are the identity transformation up to scale, i.e., no camera motion. In this case (as expected) B and G provide no additional constraints on H .

6.2.2. Counting Constrains. So far we counted the number of constraints imposed on H by a single eigen subspace. In order to count the total number of linear constraints that B and G impose on H , we analyze every possible combination of eigen subspaces according to the algebraic multiplicity of their eigenvalues:

1. $\lambda_i \neq \lambda_j \neq \lambda_k$. This implies $V_i \neq V_j \neq V_k$ and $\dim(V_i) = \dim(V_j) = \dim(V_k) = 1$.
2. $\lambda_i = \lambda_j \neq \lambda_k$ ($V_i = V_j \neq V_k$). There are two such cases:
 - (a) $\dim(V_i = V_j) = 2$, and $\dim(V_k) = 1$.
 - (b) $\dim(V_i = V_j) = 1$, and $\dim(V_k) = 1$.
3. $\lambda_i = \lambda_j = \lambda_k$. In this case there is only a single eigen subspace $V = V_i = V_j = V_k$. Its dimensionality may be 1, 2, or 3.

The following table summarizes the number of linearly independent constraints for each of the above cases:

Case	Eigenvalue algebraic multiplicity	Eigen subspace dimensionality	No. of linearly independent constraints
(1)	$\lambda_i \neq \lambda_j \neq \lambda_k$	$ V_i = V_j = V_k = 1$	6
(2.a)	$\lambda_i = \lambda_j \neq \lambda_k$	$ V_i = V_j = 2, V_k = 1$	4
(2.b)	$\lambda_i = \lambda_j \neq \lambda_k$	$ V_i = V_j = 1, V_k = 1$	4
(3.a)	$\lambda_i = \lambda_j = \lambda_k$	$ V_i = V_j = V_k = 1$	2
(3.b)	$\lambda_i = \lambda_j = \lambda_k$	$ V_i = V_j = V_k = 2$	2
(3.c)	$\lambda_i = \lambda_j = \lambda_k$	$ V_i = V_j = V_k = 3$	0

To summarize: When B (and G) have either two or three distinct eigenvalues (which is typical of general frame-to-frame transformations), then *two independent pairs of transformations suffice to uniquely determine H* . This is because each pair of transformations imposes 4 to 6 linearly independent constraints, and in theory 8 independent linear constraints suffice to uniquely resolve H (up to arbitrary scale factor).

6.3. Numerical Stability

The final step in our algorithm is to solve a homogeneous set of linear equations (Eq. (8)). Care has to be taken when solving this system. For example, inaccuracies in the estimated frame-to-frame transformations decrease the accuracy of the final output. The previous section showed that two independent pairs of transformations may suffice to uniquely determine H . In practice, however, to increase numerical stability, we use all available constraints from all pairs of *reliable* transformations after subsampling of the sequences, outlier rejection and normalization. These are explained next:

Temporal Subsampling. When the frame-to-frame transformations are too small, we often temporally subsample the sequences to obtain more significant transformations between successive frames. In our experiments where video clips were a couple of hundred frames long, we usually used 30 reliable transformations between distant (non-successive) frames. Such temporal subsampling should be done after recovering the temporal synchronization, to assure that it is done in a temporally synchronized manner across the two sequences.

Outlier Rejection. Inaccurate frame-to-frame transformations T_i are pruned out by using two outlier detection mechanisms:

- (i) The transformation between successive frames within each sequence are computed in both directions. Let T_i be the transformation from I_i to I_{i+1} , and $T_i^{Reverse}$ the transformation from I_{i+1} to I_i . Then we measure the deviation of the composed matrix $T_i T_i^{Reverse}$ from the identity matrix in terms of the maximal induced residual misalignment of pixels, i.e.,

$$Reliability(T_i) = \max_{p \in I_i} \|T_i T_i^{Reverse} p - p\| \quad (15)$$

- (ii) The similarity criterion of Eq. (5) can also be used to verify the degree of “similarity” between a pair of transformations T_i and T'_i . After Δt has been estimated and before H is estimated, an unreliable pair of transformations can be detected and pruned out by measuring the deviation of $Sim(T_i, T'_i)$ from 1. However, the first outlier criterion (that of Eq. (15)) proved to be more powerful.

Matrix Normalization. Using the heuristic provided in Golub and Van Loan (1989) (originally derived for Gaussian elimination) we normalize (scale) components of the input matrices T_i and T'_i in a way that the rows of the matrix A of Eq. (8) will have approximately the same norm. This is an equivalent step to the scaling proposed by Hartley (1997) for recovering fundamental matrices. This step indeed improve the results.

Preferred Camera Motions. When acquiring the sequences of input transformations, we usually have control over the camera motion. In general, any type of camera motion provides a frame-to-frame transformation which induces constraints on the inter-camera homography H . However, some transformations provide more stable sets of equations than others. In particular, we would like to generate sequences of transformations which provide more reliable components in each column of the matrix A in Eq. (8). For example, image-plane rotations (i.e., rotations about the optical axis of one of the cameras) usually provide reliable entries in all columns of M_i (a block of A), thus impose stable constraints on H . To conclude, the camera rig can (and should) be moved freely, however, it is recommended that a few of the camera movements include non-negligible image-plane rotations.

7. Conclusion

This paper presents an approach for aligning two sequences (both in time and in space), even when there is no common spatial information between the sequences. This was made possible by replacing the need for “consistent appearance” (which is a fundamental requirement in standard images alignment techniques), with the requirement of “consistent temporal behavior”, which is often easier to satisfy. We demonstrated applications of this approach to real-world problems, which are inherently difficult for regular image alignment techniques.

Acknowledgment

The authors would like to thank R. Basri and L. Zelnik-Manor for their helpful comments.

Notes

1. The subscript i is used to represent the frame time index, and the superscript prime is used to distinguish between the two sequences S and S' .
2. A matrix A is said to be “similar” to a matrix B if there exists an invertible matrix M such that $A = MBM^{-1}$ (see Pearson (1983)). The term “conjugate matrices” is also often used.
3. Alternatively, the input homographies can be normalized to have determinant equal to 1, to avoid the need to compute s_i .
4. A general analysis of matrix equations of the form $GH = HB$ may be found in Gantmakher (1959).
5. If $\lambda_1 \neq \lambda_2 \neq \lambda_3$ then the algebraic multiplicity of all eigenvalues is 1 ($r_j = 1$). If $\lambda_1 = \lambda_2 \neq \lambda_3$ then the algebraic multiplicity of λ_1 and λ_2 is 2, and the algebraic multiplicity of λ_3 is 1 ($r_1 = r_2 = 2$ and $r_3 = 1$). If $\lambda_1 = \lambda_2 = \lambda_3$ then the algebraic multiplicity of λ_1, λ_2 , and λ_3 is 3 ($r_1 = r_2 = r_3 = 3$).
6. Eigenvalues with algebraic multiplicity 2 and 3 are not rare. For example a homography defined by pure shift $(\Delta x, \Delta y)$ has the form: $H = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix}$. This matrix has a single eigenvalue $\lambda_1 = \lambda_2 = \lambda_3 = 1$ with algebraic multiplicity three. The corresponding eigen subspace has dimensionality 2. It is spanned by two linearly independent eigenvectors $[1, 0, 0]^t$ and $[0, 1, 0]^t$.

References

- Avidan, S. and Shashua, A. 1998. Threading fundamental matrices. In *European Conference on Computer Vision*.
- Beardsley, P.A., Torr, P.H.S., and Zisserman, A. 1996. 3D model acquisition from extended image sequences. In *Proc. 4th European Conference on Computer Vision*, LNCS 1065, Cambridge, pp. 683–695.
- Bjorck, A. 1996. *Numerical Methodes for Least Squares Problems*. SIAM: Philadelphia.

- Burt, P.R. and Kolczynski, R.J. 1993. Enhanced image capture through fusion. In *International Conference on Computer Vision*, pp. 173–182.
- Caspi, Y. and Irani, M. 2000. A step towards sequence-to-sequence alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, June 2000, pp. 682–689.
- Caspi, Y. and Irani, M. 2001. Alignment of non-overlapping sequences. In *International Conference on Computer Vision*, vol. II, Vancouver, Canada, pp. 76–83.
- Demirdijian, D., Zisserman, A., and Horaud, R. 2000. Stereo auto-calibration from one plane. In *European Conference on Computer Vision*, pp. 625–639.
- Dufournaud, Y., Schmid, C., and Horaud, R. 2000. Matching images with different resolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, June 2000, pp. 612–619.
- Gantmakher, F.R. 1959. *The Theory of Matrices*. Chelsea Pub.: New York.
- Golub, Gene and Van Loan, Charles. 1989. *Matrix Computations*. The Johns Hopkins University Press: Baltimore and London.
- Hartley, R.I. 1997. In defence of the 8-point algorithm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(6):580–593.
- Hartley, R. and Zisserman, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press: Cambridge.
- Horaud, R. and Csorka, G. 1998. Reconstruction using motions of a stereo rig. In *International Conference on Computer Vision*, pp. 96–103.
- Horaud, R. and Dornaika, F. 1995. Hand-eye calibration. *International Journal of Robotics Research*, 14(3):195–210.
- Irani, M. and Anandan, P. 1999. About direct methods. In *Vision Algorithms Workshop*, Corfu, pp. 267–277.
- Irani, M., Anandan, P., and Weinshall, D. 1998. From reference frames to reference planes: Multi-view parallax geometry and applications. In *European Conference on Computer Vision*, Freiburg, June 1998, pp. 829–845.
- Irani, M., Rousso, B., and Peleg, S. 1994. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16.
- Kumar, R., Anandan, P., and Hanna, K. 1994. Direct recovery of shape from multiple views: Parallax based approach. In *International Conference on Pattern Recognition*, pp. 685–688.
- Pearson, C.E. (Ed.), 1983. *Handbook of Applied Mathematics*, 2nd edn. Van Nostrand Reinhold Company: New York.
- Sawhney, H. 1994. 3D geometry from planar parallax. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 1994, pp. 929–934.
- Shashua, A. and Navab, N. 1994. Relative affine structure: Theory and application to 3D reconstruction from perspective views. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994, pp. 483–489.
- Slama, C.C. 1980. *Manual of Photogrammetry*. American Society of Photogrammetry and Remote Sensing.
- Stein, G.P. 1998. Tracking from multiple view points: Self-calibration of space and time. In *DARPA IU Workshop*, Monterey CA, pp. 1037–1042.
- Torr, P.H.S. and Zisserman, A. 1999. Feature based methods for structure and motion estimation. In *Vision Algorithms Workshop*, Corfu, pp. 279–329.
- Tsai, R.Y. and Lenz, R.K. 1989. A new technique for full autonomous and efficient 3D robotics hand/eye calibration. *IEEE Journal of Robotics and Automation*, 5(3):345–358.
- Viola, P. and Wells III, W. 1995. Alignment by maximization of mutual information. In *International Conference on Computer Vision*, pp. 16–23.
- Zisserman, A., Beardsley, P.A., and Reid, I.D. 1995. Metric calibration of a stereo rig. In *Workshop on Representations of Visual Scenes*, pp. 93–100.