# ALIGNING SEQUENCES FROM MULTIPLE CAMERAS

*Thommen Korah, Christopher Rasmussen*

Department of Computer and Information Sciences
University of Delaware
Newark, DE 19711

## ABSTRACT

This paper studies the problem of aligning images from multiple cameras with minimally- or non-overlapping fields of view using frame-to-frame transformations calculated for sequences from each camera. We examine implementation issues for the algorithm of Caspi and Irani that performs the alignment for two cameras which are fixed relative to each other and have approximately the same center of projection. Furthermore, we extend it to compute the camera-to-camera homographies for an $N \geq 2$ multi-camera network by simultaneously solving for all parameters of the unknown transformations. This enforces tighter constraints on the solution than performing the alignment for each pair independently. We show the efficacy of the approach on both synthetic as well as real sequences captured using a polycamera built in our lab. The aligned images can be mosaiced together to obtain a wider field of view virtual camera for subsequent processing.

## 1. INTRODUCTION

Many vision applications today such as robotics, navigation, and surveillance may benefit from cameras with larger fields of view (FOV). One popular approach is catadioptric sensors, which combine lenses and curved mirrors [1]. However, the catadioptric approach to omnidirectional vision results in variable resolution images and can be expensive for small robots. Another promising means of obtaining omnidirectional vision is to tightly cluster together multiple cameras to build a polycamera [2] with increased FOV. Rather than process images from each camera individually, registration is carried out to mosaic temporally corresponding images to one common reference frame. Several approaches to do this can be found, usually relying on either feature matching [3] or directly minimizing the intensity disparities [4]. However, such appearance based matching will not succeed with non-overlapping FOV cameras or multiple sensor types.

While automatically registering two still images with no spatial overlap is impossible, Caspi and Irani showed in [5] that with image sequences captured from both cameras simultaneously, such alignment can be carried out by making use of the additional cues encoded within each sequence. When cameras with an unknown but fixed orientation are moved together, "similar" changes over time are induced within each sequence. The "coherent appearance" model for image alignment can be replaced by "coherent temporal behavior" for sequence alignment, a less stringent requirement when building polycameras. From a signal processing standpoint, this is akin to localizing multiple visual signals and aligning them in a common coordinate system. Other approaches have also used correlated temporal behavior for spatial and temporal alignment of sequences [6], but require the cameras to be viewing the same scene.

In this work, we generalize the method in [5] to align images from multiple ($N \geq 2$) cameras. Such a camera has been used for road following applications to increase the field of view. The camera-to-camera transformation can be described by a homography $H$ if the optical centers are approximately the same. Due to the minimal overlap between the images, classical image registration methods fail to recover $H$ accurately and tend to cause distortion in the peripheral regions without overlap. By formulating a relationship between the induced changes in each camera when subject to motion, $H$ can be recovered without placing constraints on the degree of overlap. The sequence alignment is carried out for all cameras simultaneously with respect to a reference camera. Crucial to the alignment is the selection of accurate frame-to-frame transformations and appropriate camera motion, and we detail the measures that can be used to guide this process.

## 2. RECOVERING INTER-CAMERA HOMOGRAPHY $H_{QR}$

Let $C_i | i = 1..N$ be the camera network with coincident camera centers, so that a homography $H$ fully describes the inter-camera transformation. $H_{qr}$ is a linear transformation of the projective plane and is represented by a 3x3 matrix such that $x^r = H_{qr} x^q$, where $x^r$ and $x^q$ are the homogeneous image coordinates in cameras $C_r$ and $C_q$ respectively

of the 3D point $X$. Note that $X$ need not be in the FOV of either camera. Given a reference camera $r$, the goal is to find $H_{ir}$ for all $C_i$.

Let $S^k = I^k_{1..m+1}$ be the sequence of $m+1$ frames captured by the camera $C_k$. We assume for now that the frames are synchronized, but the method can be easily adapted for unsynchronized sequences as well. Since classical image alignment techniques will not work well with non overlapping FOV, we wish to recover the inter-sequence transformations from the induced frame-to-frame transformations within each sequence. Let $T^k_1, .., T^k_m$ be the sequence of intra-sequence transformations within $S^k$ such that $T^k_i$ is the homography that relates frames $I^k_i$ and $I^k_{i+1}$. The assumption of frame-to-frame transformations being homographies holds true for distant or planar scenes.

For the sake of completeness, we review the derivation in [5] to recover the alignment for two sequences. For two cameras $C_q$ and $C_r$, we assume that temporally corresponding frame-to-frame transformations are related by the fixed inter-camera homography $H_{qr}$. Let $x^q_i$ and $x^r_i$ be the image coordinates corresponding to world coordinate $X$ at frame $i$. In the next frame, these points are transformed to $x^q_{i+1}$ and $x^r_{i+1}$, such that $x^q_{i+1} \cong T^q_i x^q_i$ and $x^r_{i+1} \cong T^r_i x^r_i$, where $\cong$ denotes equality up to a scale factor. Because temporally corresponding frames are related by the fixed homography, $x^r_i \cong H_{qr} x^q_i$ and $x^r_{i+1} \cong H_{qr} x^q_{i+1}$. Given these relations we can conclude from Figure 1 that there are two equivalent paths that transform any point $x^q_i$ to $x^r_{i+1}$. Hence

$$H_{qr} T^q_i \cong T^r_i H_{qr}. \tag{1}$$

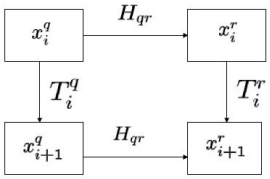


**Fig. 1**. *The two paths that transform $x^q_i$ to $x^r_{i+1}$ should be equivalent since they are true for all $x_i$*

Since the homographies are invertible we can write $T^q_i = s_i H_{qr} T^r_i H^{-1}_{qr}$ i.e the frame-to-frame transformations within the sequences are similar up to a scale factor. From the theory of matrices, the eigenvalues of similar matrices differ only by the scale factor $s_i$, which can be factored out by setting the determinants of both matrices to unity. In the relation

$$H_{qr} T^q_i - T^r_i H_{qr} = 0 \tag{2}$$

we rewrite $H_{qr}$ as $h_{qr}$, a column vector in row major order to get a set of linear equations in $h_{qr}$:

$$M^{qr}_i h_{qr} = 0$$

where $M^{qr}_i$ is a $9 \times 9$ matrix defined by

$$M^{qr}_i = \left[\begin{array}{c|c|c} {T^q_i}^T - T^r_{i11} I & -T^r_{i12} I & -T^r_{i13} I \\ \hline -T^r_{i21} I & {T^q_i}^T - T^r_{i22} I & -T^r_{i23} I \\ \hline -T^r_{i31} I & -T^r_{i32} I & {T^q_i}^T - T^r_{i33} I \end{array}\right] \tag{3}$$

Although each pair of transformations gives 9 equations, only 6 of them are linearly independent and at least two such pairs are needed to determine the unique homography relating the two sequences.

## 3. ALIGNING MULTIPLE SEQUENCES

With $N$ cameras, we would like to simultaneously solve for the $9(N-1)$ unknowns from each pair of homographies with respect to the reference camera $C_r$. Compared with the naive application of the method in [5] separately for all pairs of cameras, we claim that solving for multiple cameras simultaneously imposes tighter constraints on the final solution. The linear relation in (2) now becomes a function of all the pairwise homographies and can be expressed as

$$\sum_{q \neq r} H_{qr} T^q_i - T^r_i H_{qr} = 0 \tag{4}$$

This can be written as a system of linear equations in $h$:

$$M_i h = 0 \tag{5}$$

where $h$ is a column vector of the $9(N-1)$ unknowns and

$$M_i = \left[\begin{array}{c|c|c|c|c} M^{1r}_i & M^{2r}_i & ..... & M^{Nr}_i \end{array}\right] \tag{6}$$

a $9 \times 9(N-1)$ matrix. $M^{rr}_i$ is excluded since it is the identity. Since we have only 9 equations, and $9(N-1)$ unknowns we stack up $n$ reliable (with regard to the frame-to-frame transformations) $M_i$ matrices to form a $9n \times 9(N-1)$ matrix $A$ and solve for

$$Ah = 0.$$

Each homography has 8 degrees of freedom, while a pair of transformations provide only 6 linearly independent constraints on the homography relating them. Thus the minimum value of $n$ required to solve for all elements in $h$ is $\frac{4(N-1)}{3}$, but we use much more for greater stability. To recover all the camera-to-camera homographies, singular value decomposition is applied to compute $A = UDV^T$, and $h$ corresponds to the last column of $V$.

We can also define a reliability measure that enforces constraints on all frame-to-frame transformations across the camera network. Because these transformations are *similar*, the vector composed of the eigenvalues of temporally corresponding $T_i$'s should all be parallel to each other, which

implies that their dot products should be unity. This can be expressed by the summation

$$REL_i = \sum_{p=1}^{N} \sum_{q=1}^{N} \hat{v}_i^p \hat{v}_i^q \tag{7}$$

where $\hat{v}_i^p$ and $\hat{v}_i^q$ are unit vectors composed of the 3 eigenvalues (in decreasing order) of $T_i^p$ and $T_i^q$ respectively. The higher this measure, the more parallel the vectors.

## 4. EXPERIMENTAL SETTING

**Fig. 2**. Polycamera consisting of 3 cameras

A reference sequence is synthesized by applying a series of random rotations, translations and scaling to a set of 250 feature points. Three other sequences are generated by applying a known camera-to-camera homography to the reference sequence. Two of these were to either side of the reference sequence with minimal overlap, and the third sequence had no overlap with any of the other sequences. Gaussian noise was also added to the feature points to assess the sensitivity of the algorithm.

Shown in Figure 2 is a polycamera with $N = 3$ cameras. Neither internal nor external calibration was done on the cameras, although the parameters are assumed fixed while in operation. We pointed the polycamera at a building (planar surface) and moved it around for about 5 seconds, capturing frames at 30 fps. We now document our experiences of various factors that could effect the accuracy of the final solution.

The algorithm is particularly sensitive to the type of camera motion, requiring both rotations and translations at a minimum to allow solving for all parameters of the unknown homographies. To guarantee significant motion and compensate for registration errors, the frame to frame transformations were computed between every fifth frame.

We first employed a Harris corner detector combined with RANSAC [3] to compute the homography between frames. However, their accuracy was highly dependent on the distribution of detected features, the result even changing slightly from run to run. The direct method in [4] used by Caspi gives sub-pixel accuracy but assumes roughly affine motion. Due to the non-negligible perspective effects in our

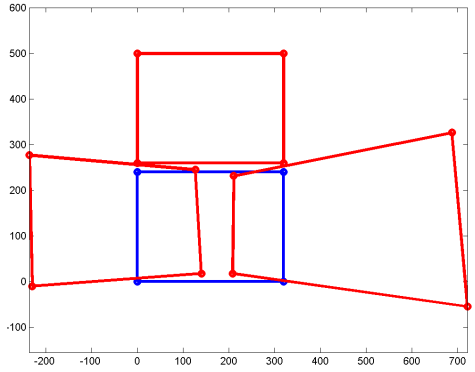


**Fig. 3**. *The recovered alignment for 4 synthesized sequences with the reference sequence shown in blue. Misalignment errors are close to zero as shown in the table below.*

building sequences, we used the slightly slower Levenberg-Marquardt style registration described by Szeliski [7], initializing it with the result of feature matching and RANSAC. The elements of the homography matrix that allow for perspective distortion are not as stable as the other elements in $H$, and this could potentially effect the accuracy of the final alignment between cameras.

We used both Sum of Squared Differences and the reprojection error to prune out the bad frame-to-frame homographies. For the computation of the inter-camera alignment, the set of homographies are sorted according to the $REL$ measure, and the top 15-20 sets are chosen. A more well-defined quality metric to choose the "best" sets of homographies remains the current area of focus. This is particularly important because a single set of bad homographies could drastically alter the computed alignment across cameras.

## 5. RESULTS

### 5.1. Synthesized sequences

Figure 3 shows the results of the algorithm on 25 frames of the 4 synthesized sequences. The reference sequence is shown in blue, and alignment of the 3 other sequences with respect to the reference is shown in red. The average pixel misalignment of the recovered homographies compared with ground truth is very close to 0, as can be seen in the table below. The error was raised to one pixel when all 250 feature points were subject to random perturbations of up to $10^{-3}$ pixels. Pure feature based matching cannot usually attain such accuracy, making the case for direct methods. We emphasize that the set of frame-to-frame transformations must include both rotations and translations, as either one alone is degenerate.

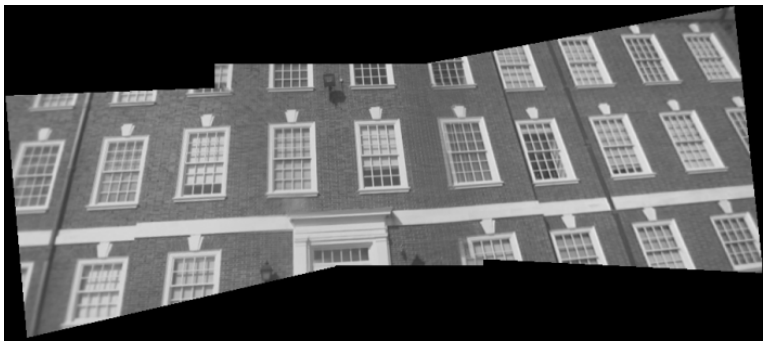

**Fig. 4**. *Alignment of images taken from 3 fixed cameras. For blending, pixels are weighted based on distance to centers*

| Sequence | Average pixel misalignment ($\times 10^{-7}$) |
|---|---|
| Left | 2.76 |
| Right | 7.76 |
| Top | 4.97 |

### 5.2. Real sequences

For real sequences, the frame-to-frame homographies were computed and the algorithm applied keeping the center camera as the reference. Notice the resulting increased FOV that could be especially beneficial for navigation and surveillance. No prior calibration is required and the technique is invariant to zoom or intensity differences across cameras. Classical image registration methods would fail to accurately recover the alignment due to insufficient overlap. While the Left Camera is almost perfectly aligned, one can notice artifacts caused at the edges of the Center and Right cameras. This occured due to appreciable vibrations of the right camera when moved. Experiments are currently being done on a newly built polycamera with even less overlap.

## 6. CONCLUSION

**Fig. 5**. *Polycamera mosaic obtained by manual correspondence for road following.*

We have described a new method to align images from multiple cameras with minimal or no overlap, based on the nature of the induced motion captured by the cameras. We evaluated a previous technique that handled only two cameras, discussing the various factors that can affect the accuracy of the technique. We extend that method to compute the alignment for all cameras simultaneously, claiming that it imposes tighter constraints. The resulting increased FOV can be useful for road following applications (Figure 5).

Future work involves defining a quality metric to select the best homographies that can recover the correct camera alignment. Another extension would be to use the property of temporal coherence for aligning 2D and 3D sensors.

## 7. REFERENCES


[1] S.K. Nayar, S. Baker, "Catadioptric Omnidirectional Camera," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[2] R. Swaminathan, S.K. Nayar, "Non-Metric Calibration of Wide-Angle Lenses and Polycameras," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[3] P.H.S. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in *Vision Algorithms Workshop*, Corfu, 1999.

[4] M. Irani, B. Rousso, and S. Peleg, "Computing occluding and transparent motions," *International Journal of Computer Vision,* 12(1):5-15, January 1994.

[5] Y. Caspi and M. Irani, "Alignment of non-overlapping sequences," in *International Conference on Computer Vision*, Vancouver 2001.

[6] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood, " View-invariant Alignment and Matching of Video Sequences," *IEEE International Conference on Computer Vision* ,Nice, France 2003.

[7] R. Szeliski, "Image mosaicing for tele-reality applications," in *IEEE Workshop on Applications of Computer Vision,* WACV 1994.