FRAME ALIGNMENT USING MULTIVIEW CONSTRAINTS

Sujit Kuthirummal, C. V. Jawahar, P. J. Narayanan

Centre for Visual Information Technology
International Institute of Information Technology
Gachibowli, Hyderabad 500 019
{sujit@gdit.,jawahar@,pjn@}iiit.net

ABSTRACT

Capturing an event using multiple cameras is gaining popularity today. Simultaneously, constraints underlying multiview geometry are also being explored. The video streams for each view have to be synchronized and aligned to a common time axis before the multiview constraints can be applied to them. Synchronization of the video frames necessarily needs a hardware based solution that is applied while capturing. However, the alignment problem between the frames of multiple views can be solved using the multiview algebraic constraints an aligned set of views must obey. In this paper, we formulate the multiview alignment problem and explore solutions for it. We also provide two specific algorithms for 3-view alignment and show their results on mapped image points.

1. INTRODUCTION

Multiview capture and analysis of dynamic events is possible today and is becoming increasingly popular. It has been recognized that the additional views can provide additional evidence that helps solve the problem of extracting meaning and structure from images of the scene. Mathematically, an image is the projection of the 3D scene onto a 2D plane of the camera. The projection results in the loss of information present in the third dimension, popularly referred to as the *depth* or the z dimension. It is easy to see that a plurality of projections can compensate for this loss more than a single view can. The multiview relations have found many applications in view generation [1], object recognition [2], video stabilization [3], etc.

The primary application of multiview constraints is in reconstructing the third dimension from a set of projections. The simplest example is classical stereo vision [4, 5]. The algebraic relations among the projections of a point onto multiple cameras have been studied extensively in Computer Vision [5, 6, 2, 7].

Multiple independent views of a dynamic event can be obtained using multiple video cameras. The multiview algebraic relations are then satisfied between the corresponding points of the views of the *same* time instant, provided the videos are synchronized to a common video signal. Using a still-camera analogy, synchronization ensures that the "shutters" of all cameras are opened at the same time instant. Thus, the visual world is sampled at the same time instants by all views. However, aligning the discretized time axes of each video to a common sequence so that the specific time instants in different views can be identified is a non-trivial task even for synchronized videos. We call this the *frame-alignment problem* for multiple views.

Hardware-based solutions to this problem are available. They involve a special equipment to insert a time-code into each video stream. The time-code can be read and compared accurately while processing the video. A common time-code signal is supplied to all videos so that the time-code is stamped on each frame of the video. The time-code usually consists of a frame number that advances with time depending on the sampling rate. The frames with identical time-codes from each camera correspond to the same time instant. Thus, the multiview relations hold for common

points of the corresponding frames.

We present a solution to the frame-alignment problem using the multiview constraints, exploiting the algebraic constraints satisfied by matched points of aligned views. We address two subproblems in this paper. The first deals with the situation when weak calibration among the views is known. Often, this can be computed as a first step from the video sequences themselves, provided there are a sufficient number of stationary points among the set of points identified in the different views. The alignment problem is solved in the Fourier domain by posing it as that of finding the shift that aligns two sequences. The second situation does not require any calibration between views. The frame-alignment problem is posed as a search in the possible space of shifts with the algebraic constraints used to verify the alignment.

We pose the problem in a multiview framework in Section 2. Section 3 describes the methodology adopted to solve the problem. The results of applying our algorithm is described in Section 4. Specific implementation details are also provided. Conclusions and future directions of research are described in Section 5.

2. PROBLEM FORMULATION

The multiview frame-alignment problem for synchronized videos that observe the same event can be defined as follows. Let the frames from n synchronized videos be written as $\ldots, f_j(-1),\ f_j(0),\ f_j(1),\ \ldots,\quad j = 1 \ldots n$ where $f_j(i)$ be the ith frame from the view j. The frames of each view are numbered consecutively, but are independent of other views. The matching problem reduces to identifying n integers $d_1 \ldots d_n$ such that the frames $f_j(i + d_j), j = 1 \ldots n$ correspond to the same time instant for all i. Without loss in generality, we can take $d_1 = 0$. The problem then reduces to finding $n-1$ integer offsets that align the frames of views $2 \ldots n$ to the frames of the first view. The multiview relations are satisfied by the set of n aligned frames since they contain n projections of the same scene. In this paper, we assume reliable correspondences for N points are available across the views. The computation of the correspondences is beyond the scope of this paper.

The essential matrix [4] and its extension to the fundamental matrix [8] encode an epipolar constraint between two views and help reduce the search space for stereo matching [8]. Trilinear algebraic relations using point and line correspondences in three views were discovered recently [2, 6, 7]. These relationships constrain where the image of a point lies in a third view, given its position in two views. They are useful to a number of problems such as the recognition of an object from a new view point and synthesis of novel views [1, 2]. It is possible to extend such algebraic relationships to multiple views, though a recent result proves that greater than quadrilinear relationships do not add anything new [5].

The Fundamental Matrix is a rank 2 matrix that constrain the image of points in one view to lie on lines in the second. If $[x^1, y^1, 1]^{\mathrm{T}}$ and $[x^2, y^2, 1]^{\mathrm{T}}$ are corresponding points in two views, the fundamental matrix encodes the following constraint.

$$x^1x^2\beta_1 + y^1x^2\beta_2 + x^2\beta_3 + x^1y^2\beta_4 + y^1y^2\beta_5 + y^2\beta_6 + x^1\beta_7 + y^1\beta_8 + \beta_9 = 0 \quad (1)$$

where the βs are the elements of fundamental matrix, defined only up to an unknown scale factor. Each point match gives one equation in terms of βs given by Equation 1. Eight point correspondences are necessary to estimate F since the fundamental matrix has 8 degrees of freedom.

The trifocal or trilinear tensor encapsulates all the projective geometric constraints between three views that are independent of the scene structure. Let P be a point in 3D space that is projected onto 3 views with image points $p^1 = [x^1, y^1, 1]^{\mathrm{T}}$, $p^2 = [x^2, y^2, 1]^{\mathrm{T}}$, and $p^3 = [x^3, y^3, 1]^{\mathrm{T}}$, respectively. Then the trilinear relation between them can be expressed as [2, 7]

$$\begin{aligned}
x^3\mathcal{T}_i^{13}p^{1i} - x^3x^2\mathcal{T}_i^{33}p^{1i} + x^2\mathcal{T}_i^{31}p^{1i} - \mathcal{T}_i^{11}p^{1i} &= 0\\
y^3\mathcal{T}_i^{13}p^{1i} - y^3x^2\mathcal{T}_i^{33}p^{1i} + x^2\mathcal{T}_i^{32}p^{1i} - \mathcal{T}_i^{12}p^{1i} &= 0\\
x^3\mathcal{T}_i^{23}p^{1i} - x^3y^2\mathcal{T}_i^{33}p^{1i} + y^2\mathcal{T}_i^{31}p^{1i} - \mathcal{T}_i^{21}p^{1i} &= 0\\
y^3\mathcal{T}_i^{23}p^{1i} - y^3y^2\mathcal{T}_i^{33}p^{1i} + y^2\mathcal{T}_i^{32}p^{1i} - \mathcal{T}_i^{22}p^{1i} &= 0 \quad (2)
\end{aligned}$$

where $\mathcal{T}_k^{ij}$ is the trilinear tensor which has 27 elements. Since each corresponding triplet contributes

four linearly independent equations and the number of unknown entries of the tensor is 26, up to scale, at least seven corresponding triplets of points are necessary to compute the tensor.

The key idea behind our frame-alignment procedure is the use of algebraic constraints such as those given in Equations 1 and 2 to ascertain the quality of alignment of a given ordering of frames from multiple views. This is achieved in the Fourier domain when the weak calibration is either given or can be calculated for the views. When no calibration information is available, the measure of alignment given by the algebraic constraints is minimized over the possible range of shifts.

For the rest of the discussion, we focus on the problem of aligning three views, wherein one of the views is misaligned with the other two. Let **A**, **B**, and **C** be the three views and let the sequence of frames from **A** and **B** be already aligned. Frames from **C** can be aligned to them using an unknown shift d. In other words, the triplet $A(i), B(i),$ and $C(i+d)$ are aligned for every i.

3. METHODOLOGY

In this section, we describe our approach to align the frames based on multiview constraints. We consider two cases separately. In the first case, the weak calibration is either given or can be computed. In the 3-view situation we are concerned with, this means that the trilinear tensor for the views **A, B,** and **C** is known before computing alignment. The second case is when the only knowledge we have is that the frames are synchronized.

3.1. Alignment with Weak Calibration

If we can identify at least 7 static points in the video sequences the trilinear tensor between the triplet of views can be computed first using any of the standard methods [7]. It is reasonable to assume seven such points will be among the identified across the views. The tensor can then be used to compute the coordinates in the view **C** of any point in **A** and **B** [3]. Let $\mathbf{c}'$ be the sequence of the positions of a specific non-stationary point computed using the trilinear tensor. Let $\mathbf{c}$ be the sequence of the same point in **C**. The following relation holds between $\mathbf{c}$ and $\mathbf{c}'$ since **C** has an unknown shift d with respect to **A** and **B**.

$$\mathbf{c}'(i) = \mathbf{c}(i+d)$$

Taking the Fourier Transform of the above two series and applying the time-shifting property of Fourier Transforms we get

$$\mathbf{C}'(\omega) = e^{j\omega d}\mathbf{C}(\omega) \tag{3}$$

for some constant d. The cross power spectrum of $\mathbf{C}$ and $\mathbf{C}'$ can be computed as

$$\frac{\mathbf{C}(\omega)\mathbf{C}'^{*}(\omega)}{|\mathbf{C}(\omega)\mathbf{C}'(\omega)|} = e^{-j\omega d} \tag{4}$$

The Inverse Fourier Transform of the cross power spectrum will have an impulse at d. The presence of a strong peak is an indication that the two sequences $\mathbf{c}$ and $\mathbf{c}'$ are shifted versions of each other. The position d of the peak gives the amount of shift that will align the view **C** with the other views. We present the frame-alignment algorithm briefly.

Algorithm FAlignA

1. Identify the subset of stationary points from the matched set of points. If fewer than 7 stationary points are available, this algorithm cannot be used for alignment. Use algorithm FAlignB.
2. Compute the trilinear tensor for the views **A**, **B**, and **C** using the stationary points.
3. Compute the sequence $\mathbf{c}'$ of image positions of a dynamic point in view **C** using its positions in the other views and the tensor computed above. This sequence is a version of the observed sequence $\mathbf{c}$ of the same point in view **C**.
4. Compute the Fourier Transform of $\mathbf{c}$ and $\mathbf{c}'$ and their cross power spectrum.
5. Find the sharp peak in the Inverse Fourier Transform of the cross power spectrum. Its location d gives the shift that would align the view **C** with the others.

3.2. Alignment without Calibration

We now study the solution to the frame alignment problem when the weak calibration is not available and cannot be computed as sufficient number of stationary points cannot be identified.

Given three frames f^a, f^b, and f^c from the three views, we can check if they correspond to the same time instant easily given point correspondences as follows. Use half of the points and compute the trilinear tensor for the three views using Equation 2 and a suitable method of solution [7]. Substitute the remaining point matches into the trilinear equations and compute the sum of their residues. If the views are aligned, the trilinear constraints will be valid and the residue will be zero or very low. This residual error can serve as a measure of alignment.

In most cases, the maximum possible shift to align the view **C** is known reasonably. Frame alignment can now be posed as a search problem over the range of possible shift values. The algorithm is described briefly below. Assume the views **A** and **B** each have n frames and the frame **C** has at least $n+2d_m$ frames where d_m is the maximum possible shift.

It is clear that the alignment problem can be posed as a search using two views, using the residual errors from the bilinear relation given in Equation 1. However, it is known that the the 2-view algebraic constraints are not as robust as the 3-view constraints [7]. We, therefore, use the trilinear equations to compute the quality of alignment.

Algorithm FAlignB

1. Compute the residual error by varying the shift d from $-d_m$ to d_m. Compute the residual error $E(d)$ using steps 2 to 4.
2. Consider all points from the block of n frames of **A** and **B**. Let **a** and **b** be an ordering of the points from respective views. Let **c** be the same ordering of points from the block of frames $C(d) \ldots C(d+n)$.
3. Compute the trilinear tensor using half the points of the ordered triplets **a**, **b**, and **c**.
4. Substitute the other half of points from the ordering in the trilinear equations and compute the sum of their residual errors.
5. The shift d for which the $E(d)$ is the minimum and is below a certain threshold will align the view **C** with the views **A** and **B**.

4. IMPLEMENTATION AND RESULTS

We tested our algorithm on a synthetic scene in which a large number of randomly placed points undergo linear motion in random directions. The 3D points are projected onto the three views with camera parameters containing translations and rotations. The image points of a video sequence are normalized across all frames of the video such that the centroid of the image positions in all frames of a view is the origin and the average distance of the points from the origin is $\sqrt{2}$. This improves the numerical performance of algorithms that estimate the tensor [7]. A least square error minimisation procedure is used to solve for the trilinear coefficients using the smallest singular values [7]. Although the point positions are generated synthetically, the image points were discretized to an integer grid. This can result in different 3D points being mapped to the same image points. The tensors computed from them are less accurate and the residual errors go up as a result.

4.1. With Weak Calibration

We applied the algorithm FAlignA given in Section 3.1 on the synthetic triplet of views. The third frame was shifted by a variable number of frames before applying the algorithm. We used 20 stationary points to compute the trilinear tensor. A non-stationary point was then tracked over 32 frames in three views to obtain the **c** and **c**$'$ sequences for the algorithm. A plot of the IDFT of the cross power spectrum (Equation 4) of the x-coordinate of the **c** and **c**$'$ sequences is shown in Figure 1. This gives a good peak at the correct shift value.

4.2. Without Calibration

We applied the algorithm FAlignB given in Section 3.2 on the synthetic triplet of views to estimate the shift of the view **C** for different values of the shift. The residual error $E(d)$ was high when

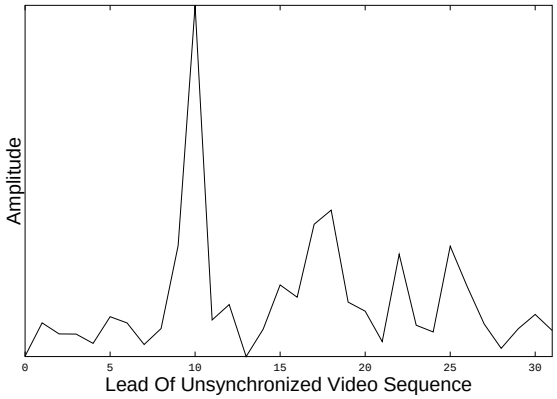

Figure 1: The IDFT of the cross power spectrum with a shift of 10.

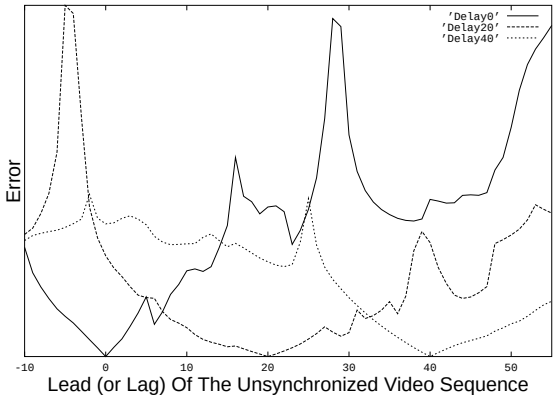

Figure 2: Plot of the error $E(d)$ against d for different values of the true shift.

the candidate shift was very different from the real shift. A plot of $E(d)$ against d for different values of the real shift is shown in Figure 2. The behaviour of the error curve is very good for the discretized views. The error $E(d)$ had a low and sharp minimum at the real shift value as can be seen from the figure.

5. CONCLUSIONS

In this paper, we defined the frame-alignment problem for multiple views and presented two algorithms to solve the same for the 3-view situation. One of them computed the weak calibration of the views from a number of stationary points and used this information to align the frames directly. The second algorithm gave a procedure to verify frame alignment for blocks of frames using the trilinear constraints among the views. Both algorithms performed excellently on synthetic experimental data consisting of three views and moving points in the presence of significant overlap among the views. That is, the shift in the frames is small compared to the length of the sequences for each view. Even though we addressed only the three-view alignment problem in this paper, it is easy to see how this can be extended to multiple views using the appropriate constraints.

We are currently extending this idea to arrive at an alignment procedure that does not involve searching for frame alignment without calibration. The alignment of multiple views is a first step towards accurate reconstruction of the underlying geometry.

REFERENCES

[1] Y. Wexler and A. Shashua, "On the synthesis of dynamic scenes from reference views," *IEEE Conf. Comp. Vision Patt. Rec.*, 2000.

[2] A. Shashua, "Algebraic functions for recognition," *IEEE Tran. Pattern Anal. Machine Intelligence*, vol. 16, pp. 778–790, 1995.

[3] A. Shashua, "Trilinear tensor:The fundamental construct of multipleview geometry and its applications," *Int. Worskshop on AFPAC*, 1997.

[4] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[5] O. Faugeras and Q. Luong, *The Geometry of Multiple Images*. USA: MIT Press, 2001.

[6] R. Hartley, "Lines and points in three views: An integrated approach," *Proc. ARPA Image Understanding Workshop*, 1994.

[7] R. Hartley and A. Zisserman, *Multiple View Gemoetry in Computer Vision*. Cambridge Univ. Press, 2000.

[8] O. Faugeras, *Three Dimensional Computer Vision*. USA: MIT Press, 1992.