VIDEO FRAME ALIGNMENT IN MULTIPLE VIEWS

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

Many events are captured using multiple cameras today. Frames of each video stream have to be synchronized and aligned to a common time axis before processing them. Synchronization of the video streams necessarily needs a hardware based solution that is applied while capturing. The alignment problem between the frames of multiple videos can be posed as a search using traditional measures for image similarity. Multiview relations and constraints developed in Computer Vision recently can provide more elegant solutions to this problem. In this paper, we provide two solutions for the video frame alignment problem using two view and three view constraints. We present solutions to this problem for the case when the videos are taken using affine cameras and for general projective cameras. Excellent experimental results are achieved by our algorithms.

1. INTRODUCTION

Imaging involves a projection of the 3D world onto a 2D image. 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. Multiple views can provide additional evidence for automated processing. The primary application of multiview constraints is to reconstruct the third dimension from a set of projections. The simplest example is classical stereo vision [1]. The algebraic relations among the projections of a point onto multiple cameras have been studied extensively in Computer Vision recently [1]. Multiview relations have found many applications in view generation, object recognition [2], video stabilization [1], etc.

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. Aligning the frames in this manner is the first step of all Computer Vision algorithms using multiple views. Figures 1,2, 3 show three situations of watching an event from different and wide-apart viewpoints.

Hardware-based solutions to this problem have been developed, involving a special equipment to insert a unique time-code into each video stream [3]. This 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 frames with identical time-codes from each camera correspond to the same time instant. This solution is cumbersome and is impractical when the cameras are physically distant. This is often the case in many application areas of Computer Vision involving multiple views. Examples include tracking a space launch vehicle or an aircraft from multiple locations, watching a sports event from different vantage points, surveillance of a large space, etc.

The frame alignment problem can be solved using an appropriate frame similarity measure used in conventional video processing. Colour histograms, shape features, etc., are popular frame similarity measures in applications like video segmentation, indexing, and content based retrieval [4]. The proper alignment of frames is an ordering of one with respect to the other that maximizes the similarity between corresponding pairs of frames. This technique can work if the features used for similarity are not too different in different views. Though this is often the case, the overall characteristics of the frame could be very different between views. An example would be the views of a football event from the sidelines and from a blimp high above.

We present a solution to the frame-alignment problem using the algebraic constraints satisfied by matched points

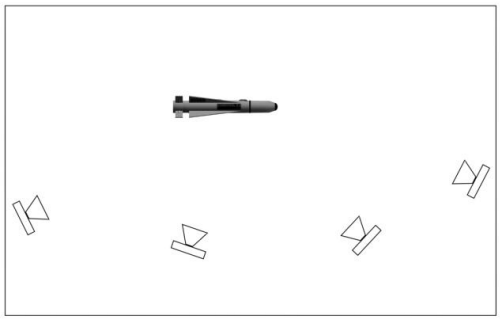

Fig. 1. A set of ground stations observing a ballistic motion

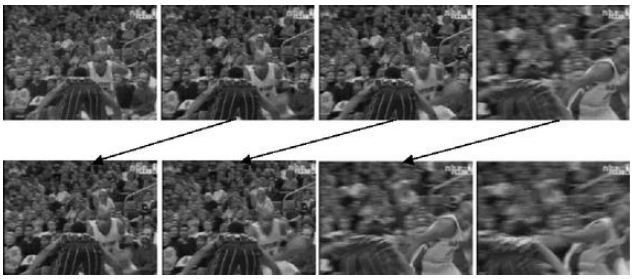

Fig. 2. Two videos of a sports event and their alignment

in aligned views. The constraints depend only on the underlying common geometry and are invariant to the viewing parameters. We present solutions for two cases in this paper. The first case is when the cameras used are affine (which is a generalization of orthographic cameras). Frame alignment between two videos can be solved using the linear constraint encoded by the Fundamental Matrix in this case. We show that a shift in the frames translates into a shift in phase in the Fourier domain. The second case deals with general projective cameras. We use the Fourier domain properties of the three view constraints encoded by the trilinear tensor in this case. This solution, however, needs to compute weak calibration in the form of the trilinear tensor before alignment. This can be computed easily using 7 or more stationary points in the views.

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 are 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 $\ldots, f_j(-1),\ f_j(0),\ f_j(1),\ \ldots,\qquad j = 1 \ldots n$ where $f_j(i)$ is the i^{th} frame from view j. The frames of each view are numbered consecutively, but are independent of other views. The alignment 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 that we are reliably able to track points across views.

Fig. 3. Observing an event using multiple cameras (Courtesy Keck Laboratory, University of Maryland)

The fundamental matrix [1] encodes a linear, epipolar constraint between projections of the same point in two views. Trilinear algebraic relations do the same for three views, constraining where the image of a point lies in a third view, given its position in two views. They are useful in solving a number of problems such as the recognition of an object from a new view point and synthesis of novel views [2]. The fundamental matrix is a rank 2 matrix that constrains the images 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.

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

where the βs are the elements of the fundamental matrix, defined only up to an unknown scale factor. Each point match gives one equation in terms of βs given by Equation 1. Therefore, eight point correspondences are necessary to estimate F since the fundamental matrix has 8 degrees of freedom. The number of unknowns reduces to 5 when the cameras are affine. Equation 1 can be written in this case as

$$ax^1 + by^1 + cx^2 + dy^2 + e = 0 \qquad (2)$$

The trifocal or trilinear tensor encapsulates the projective geometric constraints between three views that are in-

dependent of the scene structure. Let P be a point in 3D space that is projected onto 3 views with image points $p^1 = [x^1, y^1, 1]^\intercal$, $p^2 = [x^2, y^2, 1]^\intercal$, and $p^3 = [x^3, y^3, 1]^\intercal$, respectively. Then the trilinear relation between them can be expressed as [2, 1]

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

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 that the algebraic constraints such as those given in Equations 1 and 3 are satisfied by aligned frames of multiple videos. Equation 3 can also be used to predict the projection of a point in the third view given its projections in two views.

For the rest of the discussion, we focus on the problem of aligning two or three views. Let **A**, **B**, and **C** be the video sequences from 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 λ. In other words, the triplet $A(i), B(i),$ and $C(i+\lambda)$ are aligned for every i. We need to compute the λ that aligns **C** with the others.

3. METHODOLOGY

In this section, we describe our approach to align frames using multiview constraints. We consider two cases separately. The first case uses affine cameras. We show that the frames of **C** can be aligned in this case using only two views. In the second case, general projective cameras are used, a scenario that requires three views for alignment. In addition, weak calibration is assumed to be either given or computable from the video sequences themselves. In the 3-view situation that we are concerned with, this means that the trilinear tensor for the views **A, B,** and **C** is known before computing alignment.

Affine Camera Case

Let (x_i^j, y_i^j) be the image coordinates of a point in frame i of view j. Given two views **B** and **C**, Equation 2 is satisfied for each frame i if the videos are frame aligned. If we consider the sequence over time of the pairs of views, Equation 2 holds for each time instant or frame i independently. We can take the Fourier Transforms of the sequences, to get

$$aX^1(\omega) + bY^1(\omega) + cX^2(\omega) + dY^2(\omega) + e\delta(0) = 0. \quad (4)$$

If the sequences are not aligned but have a shift of λ between them, we have (using the Fourier shift theorem and ignoring the DC component)

$$aX^1(\omega) + bY^1(\omega) + c'(\omega)X^2(\omega) + d'(\omega)Y^2(\omega) = 0, \quad (5)$$

where $c'(\omega) = ce^{j2\pi\lambda\omega/n}$ and $d'(\omega) = de^{j2\pi\lambda\omega/n}$. The sequences X^1, Y^1, X^2, and Y^2 can be computed as the Fourier transforms of the corresponding sequences of image coordinates. The unknowns $a, b, c'(\omega)$, and $d'(\omega)$ can be solved for each ω using the sequences obtained by tracking a number of points across views and Equation 5. The c' values have the same magnitude for all ω and their phases are $\phi, 2\phi, \ldots, (n-1)\phi$ where $\phi = 2\pi\lambda/n$. The inverse Fourier transform of the sequence $c'(\omega)$ will have a distinct peak corresponding to the shift value of λ.

Algorithm FAlignA

1. Identify the set of tracked points in two views.
2. Compute the Fourier transform $X^j(\omega)$ and $Y^j(\omega)$ of the sequence of coordinates x_i^j and y_i^j across the frames.
3. Solve Equation 5 for a, b, c', and d' for each ω independently.
4. Compute the Inverse Fourier Transform of sequences c' and d'. Both should have strong peaks at the correct value of λ that align the sequence of frames.

Projective Camera Case

The projective camera is a generalization of the affine camera. The epipolar equation for projective cameras is given by Eqn 1. The cross-terms involving x and y from different cameras pose difficulties in extending the results of the affine case to the projective case. However, we outline a solution for aligning a video to a pair of aligned videos using the 3-view relations given in Equation 3. We need to know the tensor τ_k^{ij}, but this can be computed from the image itself if we can identify at least 7 static points in the video sequences. It may not be unreasonable to assume that seven stationary points exist in the tracked frames. The tensor can then be used to compute the coordinates in the view **C** of any point in **A** and **B** using Equation 3. Let $\mathbf{c}'$ be the sequence of the positions of a specific non-stationary point computed using the trilinear tensor and let **c** be the sequence of observed projections of the same 3D point in view **C**. The following relation holds between **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 $\mathbf{C}'(\omega) = e^{j\omega d}\mathbf{C}(\omega)$ for some constant d. The cross power spectrum of **C** and $\mathbf{C}'$ can be computed as

$$\frac{\mathbf{C}(\omega)\mathbf{C}'^*(\omega)}{|\mathbf{C}(\omega)\mathbf{C}'(\omega)|} = e^{-j\omega d} \quad (6)$$

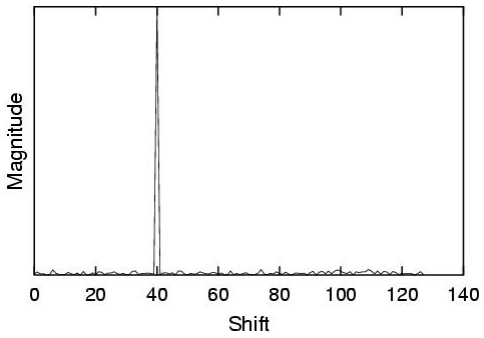

Fig. 4. The IDFT of the sequence c' for affine cameras for a shift of 40 frames

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 $\mathbf{C}$ with the other views. We present the frame-alignment algorithm briefly.

Algorithm FAlignB

1. Identify a subset of correspondent stationary points in all the views. If fewer than 7 stationary points are available, this algorithm cannot be used for alignment.
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 Transforms of $\mathbf{c}$ and $\mathbf{c}'$ and their cross power spectrum.
5. Find the highest 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, while its distinctness would reflect the quality of the alignment;

4. IMPLEMENTATION AND RESULTS

We tested our algorithm on a number of scenes. the dynamic scene is captured in two or three views. Moving points are tracked over the views and the proposed algorithms were tested on them.

Affine Camera Case: Algorithm FAlignA was used for estimating the constants $a, b, c'(\omega)$, and $d'(\omega)$ for each ω. Fifty points were tracked across 128 frames and two views for the experiments. The frames of the second view were shifted by different amounts and the algorithm was applied. The results were excellent and always estimated the correct shift. The plot of the IDFT of c' from one of the experiments is shown in Figure 4. The peak is sharp and very distinct.

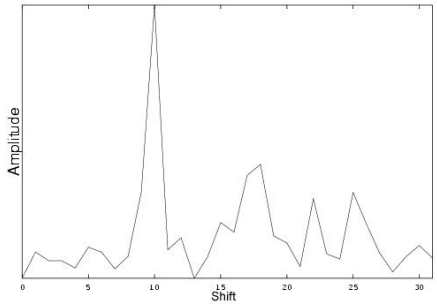

Fig. 5. The IDFT of the cross power spectrum for projective cameras for a shift of 10 frames.

Projective Camera Case: The third video 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 $\mathbf{c}$ and $\mathbf{c}'$ sequences for the algorithm. A plot of the IDFT of the cross power spectrum (Equation 6) of the x-coordinate of the $\mathbf{c}$ and $\mathbf{c}'$ sequences is shown in Figure 5. This gives a good peak at the correct shift value.

5. CONCLUSIONS

In this paper, we defined the frame-alignment problem for multiple views and presented algorithms to solve it when using affine cameras and projective cameras. The algorithms produced excellent results on a large number of experiments. The algorithms can easily be extended to aligning the frames of arbitrary number of views. We are currently working on a solution for projective cameras that does not require the knowledge of the tensor.

6. REFERENCES

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

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

[3] P. J. Narayanan, P. W. Rander, and T. Kanade, "'Synchronizing and capturing every frame from multiple cameras," *Tech. Rep. CMU-RI-TR-95-25, Carnegie Mellon University Robotics Institute*, 1995.

[4] Murat A. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.