View-invariant Alignment and Matching of Video Sequences

Cen Rao, Alexei Gritai, Mubarak Shah

SEECS
University of Central Florida
Orlando, FL 32816
{rcen, agritsay, shah}@cs.ucf.edu

Tanveer Syeda-Mahmood

K57/B2
IBM Almaden Research Center
San Jose, CA 95120
stf@almaden.ibm.com

Abstract

In this paper, we propose a novel method to establish temporal correspondence between the frames of two videos. 3D epipolar geometry is used to eliminate the distortion generated by the projection from 3D to 2D. Although the fundamental matrix contains the extrinsic property of the projective geometry between views, it is sensitive to noise. Therefore, we propose the use of a rank constraint of corresponding points in two views to measure the similarity between trajectories. This rank constraint shows more robustness and avoids computation of the fundamental matrix. A dynamic programming approach using the similarity measurement is proposed to find the non-linear time-warping function for videos containing human activities. In this way, videos of different individuals taken at different times and from distinct viewpoints can be synchronized. A temporal pyramid of trajectories is applied to improve the accuracy of the view-invariant dynamic time-warping approach. We show various applications of this approach such as video synthesis, human action recognition, and computer aider training. Compared to state-of-the-art techniques, our method shows a great improvement.

1. Introduction

Video mosaicing, video retrieval, image based modelling and rendering, video synthesis, multi-sensor surveillance, and human action recognition require spatiotemporal alignment of video sequences. Methods that tackle this problem discover a correspondence between video sequences. Some methods assume the input video sequences are already synchronized, while others use optional built-in expensive hardware that provides synchronization. We present a novel approach of alignment and matching of video sequences and only assume that the given two video sequences are correlated due to the motion of objects. Based on this correlation we discover the correspondences (temporal alignment) between the frames of the two sequences.

When a feature point moves in a 3D space with respect to time, it generates a 3D trajectory: $\{(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots, (X_t, Y_t, Z_t)\}$, where t is the time stamp. This 3D trajectory is projected as a 2D trajectory in the image plane: $\{(u_1, v_1), (u_2, v_2), \ldots, (u_t, v_t)\}$. The relationship between a point (X_i, Y_i, Z_i) in 3D trajectory and its 2D projection (u_i, v_i) is defined as follows:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = P \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}, i = 1, 2, ..., t, \quad (1)$$

where P is the projection matrix (camera model).

Assume that the same motion is occurred with a different speed, then we obtain another 3D trajectory: $\{(X_{C(1)}, Y_{C(1)}, Z_{C(1)}), (X_{C(2)}, Y_{C(2)}, Z_{C(2)}), \cdots, (X_{C(t)}, Y_{C(t)}, Z_{C(t)})\}$, where $C(i)$ is a time-warping function such that

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} = \begin{bmatrix} X_{C(i)} \\ Y_{C(i)} \\ Z_{C(i)} \\ 1 \end{bmatrix}, i = 1, 2, ..., t.$$

Now assume that the viewpoint of the camera has also been changed. The projection of this 3D trajectory to a 2D trajectory, $\{(u'_1, v'_1), (u'_2, v'_2), \ldots, (u'_t, v'_t)\}$, is defined as:

$$\begin{bmatrix} u'_{C(i)} \\ v'_{C(i)} \\ 1 \end{bmatrix} = P' \begin{bmatrix} X_{C(i)} \\ Y_{C(i)} \\ Z_{C(i)} \\ 1 \end{bmatrix}, i = 1, 2, ..., t.$$

Therefore, the problem of aligning video sequences is to discover $C(i)$, for $i = 1, 2, \ldots, t$, using the information in two 2D trajectories, $\{(u_1, v_1), (u_2, v_2), \ldots, (u_t, v_t)\}$ and $\{(u'_{C(1)}, v'_{C(1)}), (u'_{C(2)}, v'_{C(2)}), \ldots, (u'_{C(t)}, v'_{C(t)})\}$.

There are two crucial considerations when exploring correspondences of video sequences. First, the 2D trajectory is highly dependent on the viewpoint. That is, the same 3D trajectory may look different in videos shot from different viewpoints. Second, the same motion may have different temporal extents. The second problem becomes more complicated when the motion changes dynamically, such that the indices of corresponding frames are not linearly related. This is very common in videos depicting human activities, since even the same person may perform the same activity with different speeds. We propose a novel approach for alignment and matching of videos, which is based on epipolar geometry, and we can discover the nonlinear time-warping function.

There are two main types of approaches for aligning sequences: sequence-to-sequence (direct) and trajectory-to-trajectory. The direct approach takes video frames as input

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

and applies the computation over all pixels in video frames. The trajectory-to-trajectory approach tracks the movement of feature points in the field of view, and the computation is based on the information from trajectories. Advantages of the direct approach include: it determines the spatial transformation between sequences more accurately than the trajectory-to-trajectory approach does, and it does not require explicit feature detection and tracking. On the contrary, since trajectories contain explicit geometric information, the trajectory-to-trajectory approach better determines large spatiotemporal misalignments, can align video sequences acquired by different sensors and is less affected by changes in background. The detailed comparison between these approaches is available in [11, 1]. Since video sequences in most applications contain a significant spatiotemporal variance, we choose the trajectory-to-trajectory approach. Because of its benefits, we can align video sequences where different people perform the same action.

Previously, researchers have tried using calibrated/uncalibrated stereo-rigs [6, 8] to recover the projection relationships among the videos. In these approaches, the fundamental matrix is used to find the spatial relationship between trajectories [3, 13]. However, due to the instability of the reconstruction process, those approaches can only be applied to some limited video sequences (e.g. simultaneously shot videos). Therefore, there is no previous method to synchronize two videos of different people performing the same 3D activity at different times employing the fundamental matrix.

This paper propose a method, which is based on the epipolar constraint, but does not need explicit reconstruction of the 3D relationships. This method can align videos containing 3D actions with large spatiotemporal variation. Since it is a well-studied problem to reconstruct the spatial alignment of video sequences given the correspondent frames, we do not discuss spatial registration in this paper. The experimental results show that our method is much more stable than previous approaches, and can be used in many applications.

1.1. Previous Work

Stein [10] achieved the alignment of tracking data obtained from multiple camera assuming a homography relationship between cameras. Stein did not use the trajectory information, but discovered the temporal alignment using exhaustive search among different intervals between video sequences. Due to this, his method is computationally quite expensive, and it can only align videos with a constant time shift.

Giese and Poggio [7] proposed a method to find the spatiotemporal alignment of two video sequences using the dynamic shift of the time stamp of the spatial information. They assumed that a 2D action trajectory can be represented as a linear combination of prototypical views, and the effect of viewpoint changes can be expressed by varying the coefficients of the linear combination. Since they did not use the 3D information, this method can only align some simple motion patterns.

Caspi and Irani [1] proposed a direct approach to align two surveillance videos by finding the spatiotemporal transformation that minimizes the sum of squares differences (SSD) between sequences. They extended the direct approach to the alignment of non-overlapping sequences captured by a stereo rig [2]. In these video sequences, the same motion induces "similar" changes in time. This correlated temporal behavior was used to recover the spatiotemporal transformation between sequences. They also proposed a trajectory-to-trajectory approach for alignment of sequences captured by cameras with significantly different viewpoints [3]. In this method the alignment of trajectories is based on computation of the fundamental matrix. Their approaches can only be used for applications, in which the time shift between video sequences is constant or is a linear function. Therefore, their method may fail for videos with a dynamic time shift.

Wolf and Zomet [13] proposed a method for self calibrating a moving rig. During movement, the viewing angles between cameras and the time shift are fixed, but internal camera parameters are allowed to change.

Extensive research has been done to find the period of cyclic motion. The repeating pose of the human body is discovered by measuring the similarity between video frames [9], or computing peaks in the Fourier transform [12] of trajectories. However, only Seitz and Dyer [9] used a view-invariant approach, which employed the affine camera model.

From these reviews, we can conclude existent methods are not successful in aligning video sequences containing different instances of human activities.

2. View-invariant Alignment of Video

We propose a dynamic computation of the time-warping function, and a novel measure of similarity that is based on epipolar geometry.

2.1. View-invariant Measure

First, let us consider measuring similarity between 2D trajectories, which are represented as $\{(u_1, v_1), (u_2, v_2), \ldots, (u_t, v_t)\}$ and $\{(u'_{C(1)}, v'_{C(1)}), (u'_{C(2)}, v'_{C(2)}), \ldots, (u'_{C(t)}, v'_{C(t)})\}$.

In Eq.1, the general camera projection can be modelled using the following perspective matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{23} & p_{33} & p_{34} \end{bmatrix}.$$

Readers can reference to any computer vision textbook to find properties of this projection matrix. We focus on the epipolar geometry, which represents the extrinsic projective geometry between views.

For the perspective model, the fundamental matrix, $\mathbf{F}$, is defined by the equation

$$s(i) = \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}^T \mathbf{F} \begin{bmatrix} u'_{C(i)} \\ v'_{C(i)} \\ 1 \end{bmatrix} = 0, \qquad (2)$$

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

for a pair of matching points $(u_i, v_i) \leftrightarrow \left(u'_{C(i)}, v'_{C(i)}\right)$ in two trajectories. Therefore, given a fundamental matrix, we can use Eq.2 to measure the similarity between trajectories, such that the summation of $s(i)$ for all points is minimized.

It is a well known fact computation of the fundamental matrix is not robust. The variation in motion can further worsen stability. For instance, video sequences containing human activities captured at different times may vary significantly. If a person performs the same movement differently, previous approaches [3, 13] will fail to synchronize these two video sequences. Therefore, we propose a novel approach, which avoids computation of the fundamental matrix.

Given a sufficient number of point matches, Eq.2 can be used to compute the unknown matrix $\mathbf{F}$ using the following equation:

$$\mathbf{Mf} = \begin{bmatrix} u'_{c(1)}u_1 & \cdots & u'_{c(t)}u_t \\ u'_{c(1)}v_1 & \cdots & u'_{c(t)}v_t \\ u'_{c(1)} & \cdots & u'_{c(t)} \\ v'_{c(1)}u_1 & \cdots & v'_{c(t)}u_t \\ v'_{c(1)}v_1 & \cdots & v'_{c(t)}v_t \\ v'_{c(1)} & \cdots & v'_{c(t)} \\ u_1 & \cdots & u_t \\ v_1 & \cdots & v_t \\ 1 & \cdots & 1 \end{bmatrix}^T \mathbf{f} = 0 \qquad (3)$$

where f is the elements of the fundamental matrix: $f = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{bmatrix}^T$. Let us denote $\mathbf{M}$ by the observation matrix, which is constructed using coordinates of points of two 2D trajectories. Since (3) is a homogenous equation, for a solution of $\mathbf{f}$ to exist, $\mathbf{M}$ must have rank at most eight. However, due to the noise or the matching error, the rank of $\mathbf{M}$ may not be exactly eight. In this case the 9^{th} singular value of $\mathbf{M}$, denote it as $dist$, estimates the necessary perturbation of coordinates of each point in matrix $\mathbf{M}$ to produce two projections of the same 3D trajectory. Therefore, we can use $dist$ to measure the matching of two trajectories. The smallest singular value of $\mathbf{M}$ corresponds to the best match of trajectories.

We generated two trajectories, selected nine points from each trajectory and put them into $\mathbf{M}$. The 9^{th} eigenvalue increases dramatically when there is a large change in x and y coordinates of one point, and it is close to zero only within a very small range. Therefore, if points are spread far enough from each other (points are not clustered in one specific location), by picking the nine corresponding points from each trajectory, we can decide whether two trajectories match or not. Since the trajectory contains temporal information, we can also use it to align trajectories. We discuss the use of temporal information for alignment in the section 2.2.

In some applications it is reasonable to assume that the time-warping function is linear, $C(i) = ai + b$. Then parameters of the time-warping function, a and b, can be found by using an exhaustive search and by minimizing the $dist$. To model more complicated time-warping functions, a higher order polynomial must be used. However, these types of time-warping function have very limited applications, such as synchronizing two video sequences that are captured simultaneously, or synchronizing stereo cameras. Generally, this approach fails to align video sequences shot at different times and containing human activities, since the time-warping function for human activities can not be easily modelled by a polynomial.

2.2. View-invariant Dynamic Time Warping

Dynamic Time Warping (DTW) is a widely used method for warping two temporal signals. The applications include speech recognition, gesture recognition, signature recognition [5]. It uses an optimum time expansion/compression function to perform nonlinear time alignment.For two signals I and J, a distance measure E is computed to measure the misalignment between temporal signals, where $E(i, j)$ represents the error of aligning signals up to the time instants t_i and t_j respectively. The error of alignment is computed incrementally using the formula:

$$E(i,j) = dist_{i,j} + e, and$$
$$e = \min\{E(i-1,j), E(i-1,j-1), E(i,j-1)\} \qquad (4)$$

Here $dist_{i,j}$ captures the cost of making time instants t_i and t_j correspondent to each other. The best alignment is then found by keeping track of the elements that contribute the minimal alignment error at each step and backward following a path from element $E(i, j)$ to $E(1, 1)$.

The above method can only align video sequences shot from the same viewpoint. To achieve view-invariance, we introduce 3D shape information into the analysis through $dist_{i,j}$. Based on the view-invariant similarity metric from Section 2.1, we propose a view-invariant DTW algorithm as follows:

(1) We specify eight corresponding points between first frames of two videos, and denote the image coordinates as $(x'_1, y'_1), ..., (x'_8, y'_8)$ and $(x_1, y_1), ..., (x_8, y_8)$.

(2) Track the feature points in two videos to acquire trajectories $I = \{(u'_1, v'_1),\ldots,(u'_n, v'_n)\}$ and $J = \{(u_1, v_1),\ldots,(u_m, v_m)\}$. In our experiments we used the mean-shift tracker [4].

(3) For each pair of corresponding points in trajectories, construct the 9×9 observation matrix:

$$\mathbf{M_O} = \begin{bmatrix} x'_1x_1 & \cdots & x'_8x_8 & u'_iu_j \\ x'_1y_1 & \cdots & x'_8y_8 & u'_iv_j \\ x'_1 & \cdots & x'_8 & u'_i \\ y'_1x_1 & \cdots & y'_8x_8 & v'_iu_j \\ y'_1y_1 & \cdots & y'_8y_8 & v'_iv_j \\ y'_1 & \cdots & y'_1 & v'_i \\ x_1 & \cdots & x_8 & u_j \\ y_1 & \cdots & y_8 & v_j \\ 1 & \cdots & 1 & 1 \end{bmatrix}^T . \qquad (5)$$

(4) Execute the DTW algorithm but using $dist_{i,j}$, which is the 9^{th} singular value of the matrix $\mathbf{M_O}$ in step 3.

(5) Generate $C(i) = i, i = 1, \ldots, n$ by back tracing the path that minimizes the value of $E(i, j)$. If the cell $E(i, j)$ is on the warping path, it means i^{th} point of trajectory I corresponds to the j^{th} point from J.

Note that the DTW can establish the correspondence "on the fly", which means that it determines the best warping path

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

to element $E(i,j)$. To achieve more robust measure, we put previously found corresponding points up to i and j in $\mathbf{M}$, and update $\mathbf{M_O}$ in Eq.5. The matrix $\mathbf{M_R}$ is given as follows:

$$\mathbf{M_R} = \begin{bmatrix} \mathbf{M_P} \\ \hline \mathbf{M_O} \end{bmatrix}; \mathbf{M_P} = \begin{bmatrix} u'_1u_1 & \cdots & u'_{i-1}u_{j-1} \\ u'_1v_1 & \cdots & u'_{i-1}v_{j-1} \\ u'_1 & \cdots & u'_{i-1} \\ v'_1u_1 & \cdots & v'_{i-1}u_{j-1} \\ v'_1v_1 & \cdots & v'_{i-1}v_{j-1} \\ v'_1 & \cdots & v'_{i-1} \\ u_1 & \cdots & u_{j-1} \\ v_1 & \cdots & v_{j-1} \\ 1 & \cdots & 1 \end{bmatrix}^T \quad (6)$$

This algorithm is not affected by a change in the viewpoint, since the matching measure does not depend on the viewpoint, and it dynamically computes the nonlinear time-warping function between the two 2D trajectories.

2.3. Temporal Coarse-to-fine Refinement

As mentioned in Section 2.1, the matching measure does not require the explicit computation of the fundamental matrix, therefore the $rank(\mathbf{M}) = 8$ is only a necessary condition to determine whether or not the two points match. It can be noticed that the last singular value of the observation matrix shows an ambiguity if many points are very close to the correct one. Therefore, the matching algorithm might give wrong results due to the noise in the trajectory. The DTW is also sensitive to errors, such that if the warping function is incorrect at $E(i,j)$, then the error will be propagated to the rest of the warping path. To solve these problems we use temporal pyramids of trajectories.

In the temporal pyramid, a higher level has fewer points than the previous level, increasing the distance between consecutive points. The larger distance between points generates a larger change of the last singular value. Consequently, the significant variation of the last singular value determines matching points without ambiguity. Furthermore, the higher level of the pyramid provides a constraint for the lower level by propagating point correspondences. So by using the coarse-to-fine approach, we can prevent error propagation.

We propose a novel coarse-to-fine refinement for the view-invariant DTW algorithm:

(1) For the trajectory I use a spline to sub-sample the trajectory by factor of 2, such that $length(\mathbf{I}^k) = 0.5 * length(\mathbf{I}^{(k+1)})$ ($length$ is the total number of points in the trajectory), where k is the index of level of the pyramid and the highest level is labelled as $k = 0$. The same approach is applied for the trajectory J. The i^{th} point in trajectory I^k is represented as $((u'_i)^k, (v'_i)^k)$, and the j^{th} point in trajectory J^k is represented as $((u_j)^k, (v_j)^k)$.

(2) At the top level ($k = 0$), compute view-invariant DTW using I^0 and J^0.

(3) For level $k+1$, generate the observation matrix $\mathbf{M_R}$, whose first rows are the rows of observation matrix $\mathbf{M}$ from the k level.

$$\mathbf{M_R} = \begin{bmatrix} \mathbf{M_P} \\ \hline \mathbf{M_Q} \\ \hline \mathbf{M_O} \end{bmatrix}; \mathbf{M_P} = \begin{bmatrix} (u'_1)^k(u_1)^k & \cdots & (u'_{tn})^k(u_{tm})^k \\ (u'_1)^k(v_1)^k & \cdots & (u'_{tn})^k(v_{tm})^k \\ (u'_1)^k & \cdots & (u'_{tn})^k \\ (v'_1)^k(u_1)^k & \cdots & (v'_{tn})^k(u_{tm})^k \\ (v'_1)^k(v_1)^k & \cdots & (v'_{tn})^k(v_{tm})^k \\ (v'_1)^k & \cdots & (v'_{tn})^k \\ (u_1)^k & \cdots & (u_{tm})^k \\ (v_1)^k & \cdots & (v_{tm})^k \\ 1 & \cdots & 1 \end{bmatrix}^T$$

$$\mathbf{M_Q} = \begin{bmatrix} (u'_1)^{k+1}(u_1)^{k+1} & \cdots & (u'_{i-1})^{k+1}(u_{j-1})^{k+1} \\ (u'_1)^{k+1}(v_1)^{k+1} & \cdots & (u'_{i-1})^{k+1}(v_{j-1})^{k+1} \\ (u'_1)^{k+1} & \cdots & (u'_{i-1})^{k+1} \\ (v'_1)^{k+1}(u_1)^{k+1} & \cdots & (v'_{i-1})^{k+1}(u_{j-1})^{k+1} \\ (v'_1)^{k+1}(v_1)^{k+1} & \cdots & (v'_{i-1})^{k+1}(v_{j-1})^{k+1} \\ (v'_1)^{k+1} & \cdots & (v'_{i-1})^{k+1} \\ (u_1)^{k+1} & \cdots & (u_{j-1})^{k+1} \\ (v_1)^{k+1} & \cdots & (v_{j-1})^{k+1} \\ 1 & \cdots & 1 \end{bmatrix}^T$$

$$\mathbf{M_O} = \begin{bmatrix} x'_1x_1 & \cdots & x'_8x_8 & (u'_i)^{k+1}(u_j)^{k+1} \\ x'_1y_1 & \cdots & x'_8y_8 & (u'_i)^{k+1}(v_j)^{k+1} \\ x'_1 & \cdots & x'_8 & (u'_i)^{k+1} \\ y'_1x_1 & \cdots & y'_8x_8 & (v'_i)^{k+1}(u_j)^{k+1} \\ y'_1y_1 & \cdots & y'_8y_8 & (v'_i)^{k+1}(v_j)^{k+1} \\ y'_1 & \cdots & y'_1 & (v'_i)^{k+1} \\ x_1 & \cdots & x_8 & (u_j)^{k+1} \\ y_1 & \cdots & y_8 & (v_j)^{k+1} \\ 1 & \cdots & 1 & 1 \end{bmatrix}^T .$$

(4) Compute the alignment of trajectories I^{k+1} and J^{k+1}.

(5) Repeat steps 3 and 4 till the lowest level is completed.

The correspondences of points from the upper level are smoothly transitioned to the lower level of the pyramid. The ambiguity is resolved and the error does not affect the rest of time-warping function.

3. Examples and Applications

We have applied our algorithm on various video sequences. First, we used synthetic trajectory data for an accurate evaluation of the proposed approach. Next, we applied our method to synchronize real videos. From Caspi and Irani's experiments [2], we chose sequences acquired by cameras with non-overlapping FOVs, and cameras with zoomed and non-zoomed overlapping FOV in order to show the view-invariance of the proposed approach. The alignment of videos, containing human activities captured by moving and stationary cameras, illustrates the robustness of the view-invariant measure used in DTW. The synchronization of videos of different dancers and matching results can be applied in training dancers.

3.1. Synthetic Examples

We generated a 3D sinusoidal curve, and projected it onto a 2D plane using different projection matrices. Fig.1(a) shows the synthetic 3D trajectory, and Fig.1(b) shows projected 2D trajectories.

First, we used (x, y) coordinates of trajectories for the general DTW algorithm. The DTW using Euclidian distance cannot match two trajectories, since the shape of two trajectories

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

	Perspective camera model with rank constraint similarity	Fundamental matrix based similarity
No noise	Fig.2(a): excellent results	Fig.2(a): excellent results
With noise	Same as Fig.2(a): excellent results	Fig.2(b): very bad results

Table 1: The performance evaluation for different model based approaches. Each approach was tested with perfect data and degenerated data.

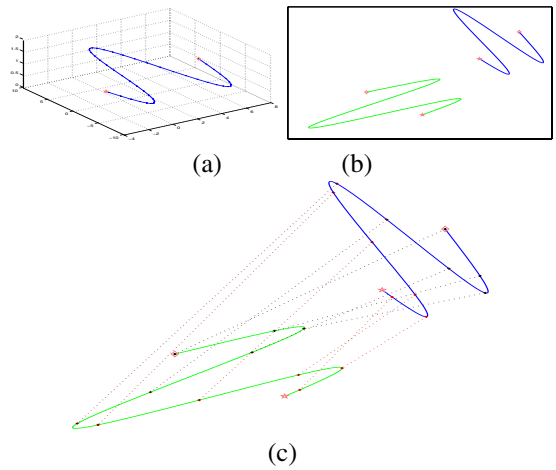
is significantly different due to projection effects. Second, we compared the view-invariant metric using the rank constraint and applied view-invariant DTW to obtain correspondence. Fig.1(c) shows results, in this figure the dotted lines connect corresponding points in each trajectory. Table 1 shows the error under different conditions.

The noise with a normal distribution $\sigma = 0.00001$ and $mean = 0$ was added to (x, y) coordinates of 2D trajectories. Fig.2 shows the histogram of correspondence errors for different methods. The horizontal axis is the error, number of frames between correspondent frames, and the vertical axis is a total number of frames that have a certain error. There are total 183 points in the sequences. Rank based approaches are not affected by this small disturbance, however, the fundamental matrix based approach degraded dramatically. We used the toolbox provided by Torr (`http://research.microsoft.com/~philtorr/`) to compute the fundamental matrix and applied the linear and nonlinear approaches. We can conclude that the rank constraint based approach is much more stable than the fundamental matrix based approach.

3.2. Zoomed and Non-overlapping Sequences

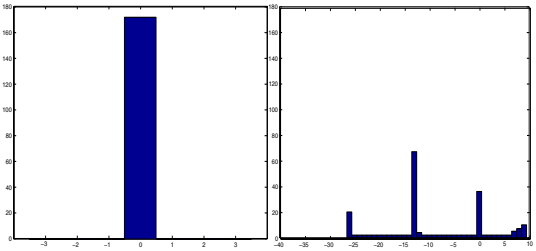
In [2], Caspi and Irani propose an attractive method to align two non-overlapping video sequences. Their approach is based on the computation of inter-frame transformations along each video sequence. This approach requires two fixed cameras installed on a common platform. In their experiments, the scene is static, but video cameras are moving. Although the field of views is non-overlapping, the spatial relationship (epipolar geometry) is still maintained.

We applied our method to sequences used in experiments in [2]. The first experiment contains one sequence captured by a camera with a wide FOV and the other captured by a camera with a zoomed FOV. The length of sequences is 300 frames. We tracked the lower left corner of the blue logo in both sequences to obtain trajectories. After alignment only nine frames had incorrect correspondences. Fig.3 shows results and the histogram of matching error. In the second experiment they used videos captured by moving cameras. There are 80 frames in each video. We tracked the right-upper corner of the gate in the right camera sequence and the left-upper corner of the gate in the left camera sequence. The view-invariant DTW discovered 71 correct correspondences, and eight frames with one frame shift. Fig.4 shows results and the histogram of matching error. In the third experiment they



(a) (b) (c)

Figure 1: (a) A synthetic trajectory in 3D space. (b) The two projected trajectories of (a) in 2D space. (b) The view-invariant dynamic time warping result, where the dot lines connect the corresponding points



(a) (b)

Figure 2: (a) The histogram of matching error using the rank constraint of the perspective camera with/without noise. (b) The histogram of matching error using fundamental matrix with very small noise in the data.

used non-overlapping sequences. The first half of the videos contains the building around the football stadium. We tracked one feature on the wall of the football stadium and the corner of the window. The view-invariant DTW discovered 151 correct correspondences, 21 frames with one frame shift, and 28 frames with two frames shift. Fig.5 shows results and the histogram of matching error. The error may due to the tracking error.

3.3. Alignment of Videos Containing Human Activities

From the previous experiments it is hard to evaluate the effectiveness of the DTW function. Video sequences were captured simultaneously so the trajectories do not contain the dynamic changes among the corresponding frames.

In the first experiment, we capture two students moving their hands up and down with different speeds. We recorded three videos using one camera. The first two videos were captured using a static camera from different viewpoints, while the third one was captured using a moving camera. The hands were tracked using the mean-shift tracker. We stabilize the frames of the third video by subtracting the image coordinates of a static point from the image coordinates of the hand. There was a time-shift of approximately half of the motion cy-

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

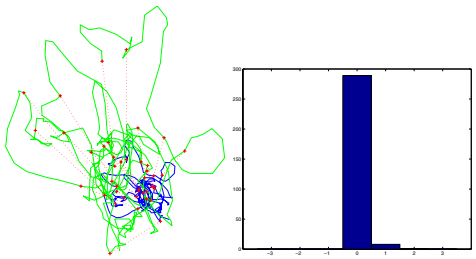


Figure 3: The correspondence results for the zoom sequences.

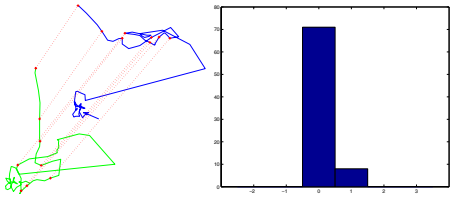


Figure 4: The view-invariant DTW correspondence result for jump sequences.

cle between videos. We used the perspective camera model in the rank-based approach to synchronize these videos. Despite the changes in the viewpoints and the nonlinear time shift, our method successfully established the correspondence between videos. Fig.6 shows input videos. Fig.7 shows results of the view-invariant DTW. Results are quite impressive, since large temporal variation has been dealt with.

The next experiment dealt with synchronizing of videos that contain more complicated human activities. We recorded three dancers performed the same dance. For each dancer we captured two video sequences from two significantly different view points. Fig.8 shows trajectories of the left feet of dancers in the six sequences. The differences between trajectories are due to variations in viewpoint, speed, and motion. We computed the temporal correspondence for each trajectory point with respect to the points in the other five trajectories. So there are total $C_2^6 = 15$ computations. Based on the pair-wise correspondences, we generated a video containing all six synchronized dance sequences, such that sequence $\#1$ is warped to sequence $\#6$ based on the warping function computed from trajectories $\#1$ and $\#6$, sequence $\#2$ is warped to sequence $\#6$, and so on. Note the large spatial difference between trajectories $\#3$ and $\#4$. Fig.9 shows one of the warping results, in which all sequences are warped to sequence $\#6$. Each row contains some key frames in the video, and corresponding frames are shown in each column. Please visit `http://cs.ucf.edu/~vision` for the full size input/output movies. Although videos contain large variations in motion, our algorithm successfully computed the correspondence from one frame to the frames in the other sequences.

3.4. Computer Aider Training

The time-warping function is a path that minimizes the alignment error at each step through the similarity measure $\mathbf{E}$. Each point from the path represents the correspondence between the i^{th} point in trajectory I and the j^{th} point in trajectory J. If many points in the trajectory I correspond to the same point in trajectory J, then it means that the movement of sequence I is slower than the movement of sequence J at that moment. This observation gives us a clue about the performance estimation. We considered sequence $\#6$ as a model and sequence $\#1$ as a test, and computed the warping path between them. Fig.10(a) shows results. From this figure we can notice dancer $\#1$ had a pause at around frame 150. Fig.10(b) shows the time-warping path between sequences $\#2$ and $\#6$. This figure shows dancer $\#2$ did not slow down at frame 80. This method shows where improvement is needed. Fig.11(a) shows the similarity measure along the time-warping path for sequences $\#1$ and $\#6$. We noticed dancer#1 did well overall, but she had a bad movement from frames 150 to 200. We checked the input sequence, and found that she lowered her leg from the upper most position around that time. Therefore, we concluded that she may need to improve that part. Fig.11(b) shows the similarity measure for sequences $\#2$ and $\#6$, we detected dancer $\#2$ had the same problem as dancer $\#1$.

With the help of view-invariant DTW, we can easily develop a self-training system, such that the users (dancers $\#1$ and $\#2$) record their performance, and compare them to the master's (dancer $\#3$) performance. Then the system can give

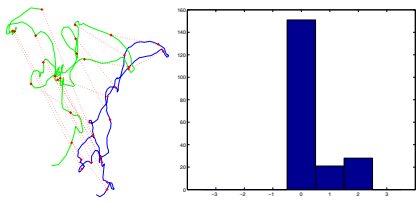


Figure 5: The view-invariant DTW correspondence result for football sequences.

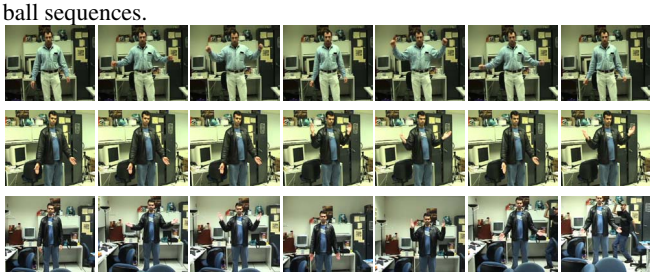


Figure 6: The human activity sequences. The first, second and third rows respectively shows the first, second and third input sequences, which are not synchronized. The columns are ordered as frame 0,20,40,60,80,100, and 120 for each sequence.

Figure 7: The output of the view invariant dynamic time warping. The columns represent the synchronized corresponding frames. Every 40th of the output frames are shown, they are 11,51,91,131,171,211,251,291.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

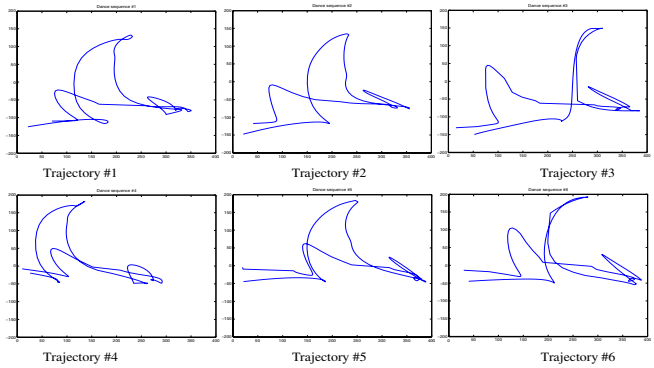


Figure 8: Trajectories of the right feet of dancers in 6 sequences. The 1^{st} row contains trajectories $\#1, \#2$ and $\#3$ corresponding to 1^{st}, 2^{nd} and 3^{rd} dancers respectively. The 2^{nd} row contains trajectories $\#4, \#5$ and $\#6$ corresponding to 1^{st}, 2^{nd} and 3^{rd} dancers respectively.

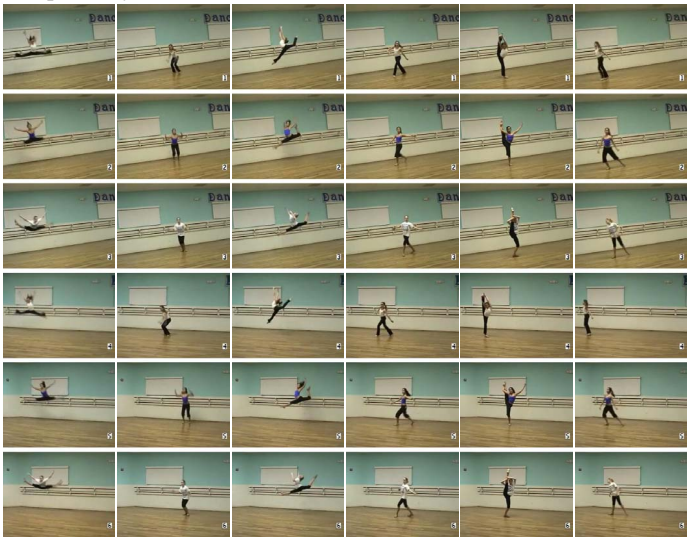


Figure 9: Key frames of output sequences (the frame index is shown at the bottom of figures). Sequences $\#1, \#2, \#3, \#4, \#5$ are warped to sequence $\#6$ and are shown according to rows. The 1^{st} and 4^{th} correspond to dancer $\#1$, the 2^{nd} and 5^{th} correspond to dancer $\#2$, and the 3^{rd} and 6^{th} correspond to dancer $\#3$.

suggestions about the speed and the extent of their movement. Note that the beginner's and master's camera viewpoints can be different. Therefore, this method has great potential.

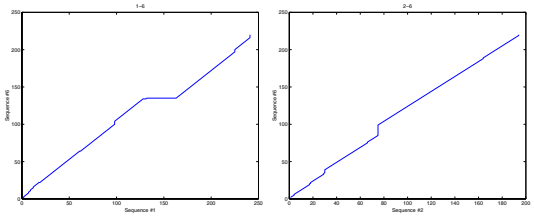
3.5. Conclusion

In this paper, we proposed the view-invariant DTW method to establish the temporal correspondence between frames of two videos. We demonstrated applications using this method, such as video synthesizing, computer aider training, and non-overlapping video sequences.

References

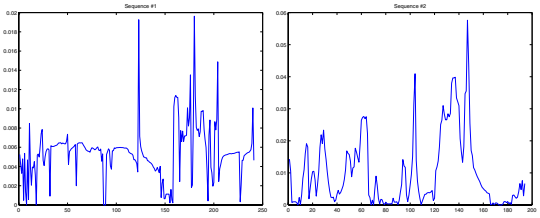
[1] Y. Caspi and M. Irani. A step towards sequence-to-sequence alignment. In *CVPR00*, pages 682–689, 2000.

[2] Y. Caspi and M. Irani. Alignment of Non-Overlapping sequences. In *ICCV'01*, pages 76–83, 2001.



(a) (b)

Figure 10: (a) The time-warping path between sequences $\#1$ and $\#6$. (b) The time-warping path between sequences $\#2$ and $\#6$.



(a) (b)

Figure 11: (a) The similarity measure between sequences $\#1$ and $\#6$; there is a large spatial difference from frame 150 to 200. (b) The similarity measurement between sequences $\#2$ and $\#6$; there is a large spatial difference from frame 120 to 160.

[3] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. In *VAMODS workshop with ECCV*, 2002.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000.

[5] T.J. Darrell, I.A. Essa, and A.P. Pentland. Task-specific gesture analysis in real-time using interpolated views. *IEEE, PAMI*, 1995.

[6] D. Demirdjian, A. Zisserman, and R. Horaud. Stereo autocalibration from one plane. In *ECCV*, pages 625–639, 2000.

[7] M. Giese and T. Poggio. Synthesis and recognition of biological motion patterns based on linear superposition of prototypical motion sequences. In *MVIEW 99*, 1999.

[8] R. Horaud and G. Csurka. Autocalibration and euclidean reconstruction using rigid motion of a stereo rig. In *ICCV*, 1998.

[9] S. M. Seitz and C. R. Dyer. View-invariant analysis of cyclic motion. *IJCV*, 25:1–25, 1997.

[10] G.P. Stein. Tracking from multiple view points: Self-calibration of space and time. In *DARPA IU Workshop*, pages 521–527, 1998.

[11] P.H.S. Torr and A Zisserman. Feature based methods for structure and motion estimation. In *International Workshop on Vision Algorithms*, 1999.

[12] P.S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection for motion based recognition. *Pattern Recognition*, 27(12), 1994.

[13] L. Wolf and A. Zomet. Sequence to sequence self-calibration. In *ECCV*, May 2002.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE