VIDEO SYNCHRONIZATION VIA SPACE-TIME INTEREST POINT DISTRIBUTION

Jingyu Yan and Marc Pollefeys

{yan,marc}@cs.unc.edu
The University of North Carolina at Chapel Hill
Department of Computer Science
Chapel Hill, USA

ABSTRACT

We propose a novel algorithm to synchronize video recording the same scene from different viewpoints. Our method relies on correlating space-time interest point distribution in time between videos. Space-time interest points represent events in video that have high variation in both space and time. These events are unique in time and may pronounce themselves in videos from different viewpoints. We show that by detecting, selecting space-time interest points and correlating their distribution, videos from different viewpoints can be automatically synchronized.

1. INTRODUCTION

Dynamic scene reconstruction keeps attracting interest. To reconstruct a dynamic scene, it is common practice to have more than one camera recording the scene. In a scenario the scene is recorded by handheld cameras, the first step toward reconstruction is to synchronize these videos. Although the synchronization may be done manually, an automatic algorithm is highly desirable or even necessary when the reconstruction algorithm does not have control on when and how the videos come in.

We propose a novel algorithm to synchronize video recording the same scene from different viewpoints. Our method relies on correlating space-time interest point distribution in time between videos. Space-time interest points represent events in video that have high variation in both space and time. These events are unique in time and may pronounce themselves in videos from different viewpoints. We show that by detecting, selecting space-time interest points and correlating their distribution, videos from different viewpoints can be automatically synchronized.

Our method first detects space-time interest points in video using scale-adapted techniques. Then by selecting the strongest interest points using a uniform search with uniform sampling algorithm in each video, it forms a distribution of space-time interest points. This distribution becomes a descriptor of time feature of the video. A correlation algorithm then tries to correlate these distributions and estimate temporal difference between videos.

2. PREVIOUS WORK

Multiple view reconstruction requires synchronization of videos from different viewpoints recording the same scene. Most of the early work does not address the problem of synchronization and synchronization is mostly done manually. Until recently, Wolf etc.[1] proposed a synchronization algorithm that tries to find the time shift by minimizing the rank of a matrix stacked with tracking point data from two cameras and Caspi etc.[2] proposed a synchronization method that is based on matching the trajectories of objects from different viewpoints.

Our method differs from the previous methods in that it does not need any correspondence between image features, not even the linear-combination relation between image features required by [1]. Instead, it exploits the correlation of space-time interest point distribution in different videos of the same scene. By correlating these distributions it achieves synchronization without any explicit image feature correspondence.

Section 3,4 discuss the basic idea of scale-adapted space-time interest point. Section 5,6 introduce the concept of distribution of space-time interest points in time and gives an algorithm to sample these distributions for synchronization purpose. Section 7 show our experimental results. Section 8 draws the conclusion and describe potential future work.

3. SCALE-ADAPTED SPACE-TIME INTEREST POINT

We first introduce image interest points and extend the concept naturally to space-time interest points.

Image interest points have been studied for a long time. The most widely used detector of these points is the Harris corner detector[3]. Suppose a gray image is represented as I: $R^2 \rightarrow R$, the convolved matrix η indicates the variation in image space at point (x, y):

$$\eta = g(\cdot, \cdot, \sigma) * \begin{pmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{pmatrix} \quad (1)$$

I_x and I_y are the derivatives of image intensity along image coordinates. $g(\cdot,\cdot,\sigma)$ is the gaussian kernel with variance σ

$$g(\cdot,\cdot,\sigma) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

Harris corner detector detects an interest point by finding the positive maxima of the corner function which is proposed by Harris and Stephens [3]

$$H = det(\eta) - k\, trace^2(\eta) \tag{3}$$

Scale adapted image interest point detection is proposed by Lindeberg [4][5]. The idea is to use a combination of normalized Gaussian derivatives as a gauge for scale selection. For example, a possible choice of gauge is a normalized Laplacian:

$$\nabla^2_{norm} = I_{xx,norm} + I_{yy,norm} \tag{4}$$

where $I_{xx} = \sigma^2 I_{xx}$ and $I_{yy} = \sigma^2 I_{yy}$. σ is the variance of the Gaussian kernel. Estimating the scale boils down to detecting the σ at which ∇^2_{norm} assumes local maxima.
Further, Lindeberg[6] explores temporal scales followed by more recent work of Laptev and Lindeberg [7][8] that generalizes the above ideas to the space-time domain. The interest point detection matrix becomes

$$\eta = g(\cdot,\cdot,\cdot,\sigma) * \begin{pmatrix} I_x^2 & I_x I_y & I_x I_t \\ I_y I_x & I_y^2 & I_y I_t \\ I_x I_t & I_y I_t & I_t^2 \end{pmatrix} \tag{5}$$

I_x and I_y are the derivatives of image intensity along image coordinates. I_t are the derivative along time.
The corner function that tells the strength of an interest point becomes

$$H = det(\eta) - k\, trace^3(\eta) \tag{6}$$

The normalized Laplacian for scale selection in space-time becomes:

$$\nabla^2_{norm} = I_{xx,norm} + I_{yy,norm} + I_{tt,norm} \tag{7}$$

$I_{xx,norm} = \sigma^{2a}\tau^{2b} I_{xx}$, $I_{yy,norm} = \sigma^{2a}\tau^{2b} I_{yy}$ and $I_{tt,norm} = \sigma^{2c}\tau^{2d} I_{tt}$. σ and τ are the variances of the Gaussian kernel in space and time. The parameters a,b,c and d are set as $a = 1, b = \frac{1}{4}, c = \frac{1}{2}$ and $d = \frac{3}{4}$ for a prototype space-time event of a gaussian blob[7][8]. These values can be used when no prior information of space-time events available and can be adjusted for more specific event detection.
We are going to adopt this space-time interest point detector in our synchronization algorithm.

4. SELECTION OF SPACE-TIME INTEREST POINTS

Space-time interest points represent events such as object appearing and disappearing, object breaking and merging, and velocity change[7][8]. We need a criterion to select space-time interest points in video.
First, space-time interest points are selected by their strength shown by the corner function (eq. 6). Toward this end, we need to manipulate the function. Let $H = 0$ in (eq. 6). We have

$$k = \frac{det(\eta)}{trace^3(\eta)} \tag{8}$$

where η is given by (eq. 5). k indicates the strength of a space-time interest point. The more variation in space and time the interest point has, the larger k is.
Another way to express k allows to observe the range of k:

$$k = \frac{\lambda_1\lambda_2\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)^3} \tag{9}$$

where λ_1, λ_2 and λ_3 are the eigenvalues of η given in (eq.5). Because we are only interested in positive k and in that case all eigenvalues must be greater than 0, the range of k is $(0, \frac{1}{27}]$.
We select those points whose k value is over some strength threshold S in $(0, \frac{1}{27}]$ as space-time interest points.
It could be the case that λ_1, λ_2 and λ_3 are small, which is an indication of weak interest points, but because they have similar values they result in large k, which is an indication of strong interest points. So we need one more criterion to exclude such a situation. We set a simple threshold T for the minimum of λ_1, λ_2 and λ_3.
Combining these two criteria, we select space-time interest points based on

- The minimum of eigenvalues is greater than threshold T
- The strength indicated by k computed using (eq.9) is greater than threshold S.

In practice, we are interested in finding a certain number, e.g. 200, of the strongest space-time interest points in a video. We do not set threshold S and T explicitly. Instead, our algorithm has two passes. In the first pass, we keep pruning interest points with smaller k until a certain number, e.g. 400, of interest points are left. In the second pass, out of the 400 interest points, we choose the 200 with larger minimum eigenvalues.

5. SPACE-TIME INTEREST POINTS DISTRIBUTION IN TIME

Because space-time interest points represent special events such as object appearing and disappearing, object breaking and merging, and velocity change[7][8],for two cameras recording the same scene, most of these events, if not all, can be expected to appear in both videos. The accumulation of space-time interest points between different

videos recording the same scene highly correlate. Space-time interest point distribution over time can serve as a descriptor of time feature of the video. By correlating these distributions, videos can be synchronized.

We need effective techniques to sample space-time interest point distribution, which is the topic of this section, and then to exploit their correlation, which is the topic of the next section.

An efficient algorithm to sample the distribution of space-time interest points is needed. The reasons why we need a sample rather than a complete distribution are

- It is computationally expensive to find all space-time interest points in video
- It is not necessary. Distribution formed by a subset of the strongest space-time interest points is usually all we need for correlation purpose.

We choose a uniform search with uniform sampling algorithm to form the distribution. We look at video data as a function $f(\vec{X})$ ($\vec{X} = [x, y, t]$, x, y are the image space coordinates and t is the time coordinate). We divide the domain of f into uniform regions, then uniformly sample within each region and find the strongest space-time interest point locally. Then we globally select a certain number of the strongest space-time interest points from regional ones, from which the distribution is formed.

Distribution of space-time interest points is presented as a histogram. The x-axis represents time and y-axis, the total number of space-time interest points found accumulated at that time.

6. CORRELATION OF SPACE-TIME INTEREST POINT DISTRIBUTION AND ESTIMATION OF TEMPORAL DIFFERENCE

The correlation between space-time interest point distribution is a good measurement for synchronization. Suppose the distribution in the first video is represented as $V1 = [\eta_1, \eta_2, ..., \eta_m]$ and the distribution in the second video is represented as $V2 = [\zeta_1, \zeta_2, ..., \zeta_n]$, the correlation function is defined as

$$C(t) = \sum_{j=max(-t,1),k=max(t,1)}^{j=min(m,n)-max(-t,1),k=min(m,n)-max(t,1)} \eta_j \ \zeta_k$$

t is the possible frame(time) offset. t is an integer within $[-m + 1, n]$. t is such defined that when it is positive the first video is lagging behind, negative, running ahead. The t at which $C(t)$ reaches its maximum is the estimation of the temporal difference.

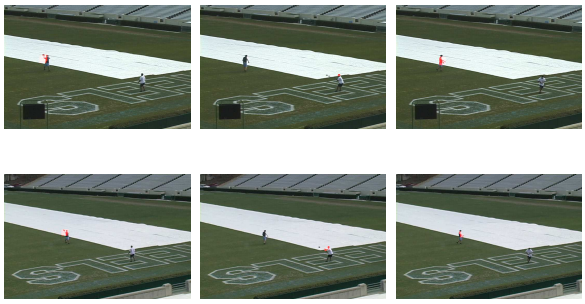

Figure 1: *(top) sample frames of the first football field sequence with space-time interest points denoted using transparent red squares (bottom)sample frames of the second football field sequence with space-time interest points denoted using transparent red squares.*

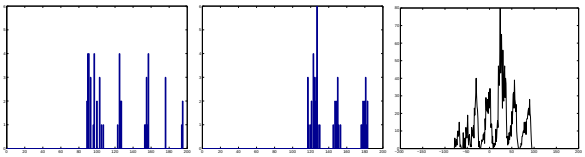

Figure 2: *(left) the distribution of space-time interest points for the first football field sequence (middle) the distribution of space-time interest points for the second football field sequence (right) the correlation of the two distributions. The correlation reaches its maximum for a frame offset of +24 which is the estimation of the temporal difference between two videos.*

7. EXPERIMENTS

We carried out experiments to test our synchronization method. Most of the tests turned out positively. There was one case where our method failed and we will give out the reason and, more importantly, a method which can inform us when our method possibly fails.

The first test case(Figure 1,2) is a remote scene with two cameras recording a field where two persons are practicing football. The videos that we use for synchronization are of 200 frames running about 6 seconds. The total number of strong space-time interest points in each video, detected, selected and used to build the distribution by our algorithm, is around 200. The temporal difference estimated by our method is 24 frames. The result corresponds to the value derived by human inspection.

The second test case(Figure 3,4) is an indoor scene with two cameras recording the same person talking and gesturing. The videos that we use are of 100 frames running about 3 seconds and well synchronized in a controlled lab environment. We offset one video by 20 frames. The total number of strong space-time interest points in each video, detected, selected and used to build the distribution by our

Figure 3: *(top) sample frames of the first indoor sequence with space-time interest points denoted using transparent blue squares (bottom) sample frames of the second indoor sequence with space-time interest points denoted using transparent blue squares.*

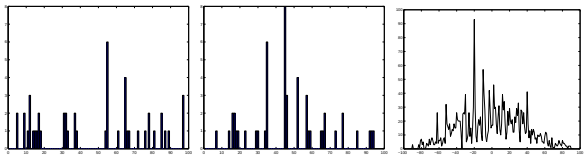

Figure 4: *(left) the distribution of space-time interest points for the first indoor sequence (middle) the distribution of space-time interest points for the second indoor sequence (right)the correlation of the two distributions. The correlation reaches its maximum for a frame offset of -20 which is the estimation of the temporal difference between two videos.*

Figure 5: *(top) sample frames of the third indoor sequence with space-time interest points denoted using transparent blue squares (bottom) sample frames of the second indoor sequence with space-time interest points denoted using transparent blue squares.*

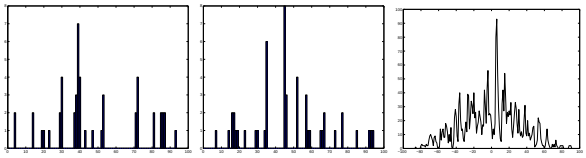

Figure 6: *(left) the distribution of space-time interest points for the third indoor sequence (middle) the distribution of space-time interest points for the second indoor sequence (right)the correlation of the two distributions. The correlation reaches its maximum for a frame offset of +6 which is the estimation of the temporal difference between two videos.*

algorithm, is around 200. The temporal difference estimated by our method is 20 frames. The result corresponds to ground truth.

The third test case(Figure 5,6) is similar to the second one only that two cameras have a wider baseline. The temporal difference estimated by our method is 6 frames. The result corresponds to ground truth. However, it can be expected that as the baseline gets wider, it gets harder for our method to synchronize.

The fourth test case(Figure 7,8) shows a failure of our method. The scene is a jogger. This test case is challenging. First, two cameras have a very wide baseline. Second, the jogger is partially occluded in one viewpoint. Third, the jogging motion which incurs space-time interest points is repetitive. There are three high peaks apparently apart from each other corresponding to frame offsets of -11, -1 and 22 in the correlation graph(Figure 8). Although the maximum correlation is reached for a frame offset of -11, the actual temporal difference is 22 which corresponds to the third highest peak. Multiple peaks in the correlation graph indicate ambiguity of the estimation. It can be detected by the following method:

- Find all peaks that are above the value of 80% of the highest peak.
- If there is no other peak in that range except the highest one, the estimation is reliable.
- If there are multiple peaks that correspond to temporal differences apart from each other, more accurately, whose distances are beyond the accuracy tolerance of estimation, e.g. a frame offset of 5, the estimation is ambiguous; however, if they correspond to temporal differences whose distances are with the tolerance, the estimation from the highest peak is still reliable because the neighboring peaks are presumably side products of a good correlation.

8. CONCLUSION AND FUTURE WORK

By correlating space-time interest point distribution of videos recording the same scene, the temporal difference between them can be estimated. We demonstrate this conceptually, from the implication of space-time interest points and the

Figure 7: *top) sample frames of the first jogger sequence with space-time interest points denoted using transparent red squares (bottom)sample frames of the second jogger sequence with space-time interest points denoted using transparent red squares.*

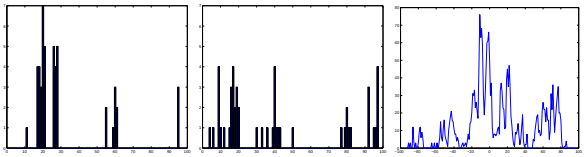

Figure 8: *(left) the distribution of space-time interest points for the first jogger sequence (middle)the distribution of space-time interest points for the second jogger sequence (right) the correlation of the two distributions. The correlation reaches its maximum for a frame offset of -11 which is the estimation of the temporal difference between two videos.*

correlation between their distributions, and factually, from different test cases. We also gives a method to decide whether ambiguity exists in the estimation.
Further investigation includes using space-time interest points to automatically calibrate cameras. This involves space-time interest point matching between videos and an effective algorithm to deal with outliers. We expect that successful application of it may require videos recording the same scene to satisfy certain requirements on baseline, object occlusion, etc.

9. REFERENCES

[1] L. Wolf and A. Zomet, "Correspondence-free synchronization and reconstruction in a non-rigid scene," in *Workshop on Vision and Modelling of Dynamic Scenes*, Copenhagen, May 2002.

[2] Denis Simakov Yaron Caspi and Michal Irani, "Feature-based sequence-to-sequence matching," in *Workshop on Vision and Modelling of Dynamic Scenes*, Copenhagen, May 2002.

[3] C. Harris and M.J. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988, pp. 147–152.

[4] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV, 30(2)*, pp. 77–116, 1998.

[5] Tony Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, 1994.

[6] T. Lindeberg, "On automatic selection of temporal scales in timecausal scale-space," in *AFPAC97: Algebraic Frames for the Perception-Action Cycle, volume 1315 of Lecture Notes in Computer Science*, Springer Verlag, Berlin, 1997, pp. 94–113.

[7] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV 2003*, Nice, France, 2003, pp. 432–439.

[8] I. Laptev and T. Lindeberg, "Interest point detection and scale selection in space-time," in *L.D. Griffin and M Lillholm, editors, Scale- Space03, volume 2695 of Lecture Notes in Computer Science*, Springer Verlag, Berlin, 2003, pp. 372–387.