15-24

15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Fine-grained event timing detection method using quasi-high frame generation for single camera image sequence

Ayumi MATSUMOTO
NTT Media Intelligence Laboratories

Dan MIKAMI
NTT Media Intelligence Laboratories

Hideaki KIMATA
NTT Media Intelligence Laboratories

Abstract

This paper describes a method we propose that obtains event timing at sub-frame level, which is more precise than frame level, on the basis of temporal super-resolution for sports videos. Since athletes move very quickly in sports situations the required time resolution for event detection for sports motion analysis is quite fine grained. Thus, we need to detect events that have not been recorded even if means having to check events at every frame. The proposed method is able to generate quasi-high frame rate videos, but does so with difficulty because the videos have too high a degree of freedom. We focus on the fact that repeated motions occur in many sports, and the projection of low-dimensional feature space is obtained on the basis of these repeated motions. Interpolations in the low-dimensional feature space make it easy to provide quasi-high frame rate videos that enable event detection at sub-frame level. Experiment results verified that our method detects ball release timing from 30 fps video with mean error of less than 0.02 seconds. Moreover, subjective evaluation experiments showed that the proposed method has accuracy applicable to the sports training VR system we developed.

1 Introduction

Recently, sports support methods using IT technology have become widely used [1]. The methods are diverse, involving factors such as tracking the sports players engage in [2], providing metadata to sports videos [3], and detecting events [4]. A specific example is the real-time tracking technology that the Hawk-eye system [5] provides, which has been officially adopted as a way to help tennis and soccer players make decisions.

With this background in mind, we have been doing research and development work aiming to support sports training using IT technology. We have proposed a method of dividing motion into multiple phases to evaluate proficiency of motion [12] and a method of presenting the motion image and the data acquired by the sensor in the VR space synchronously [11]. Through these studies, it has become clear that it is important to accurately detect instantaneous timing at which a specific event (impact of golf swing, release of pitch, etc.) occurs when analyzing and presenting sports motions. In this study, we suggest a framework that detects the timing of event that often occur in sports video. One point requires particular attention here. If the video frame rate is very high (e.g., 120 fps) and the event timing is recorded reliably, the timing detection is very easy. However, at the present time the video images generally used are 30 fps (in recent years, 60 fps has also become widely used), and it frequently happens that the actual moment of the ball's release is not recorded into the video data (Fig.1). Our research addresses this point. It can be interpreted as the event timing detection at the sub-frame level, which is not actually photographed because of the frame rate relationship.

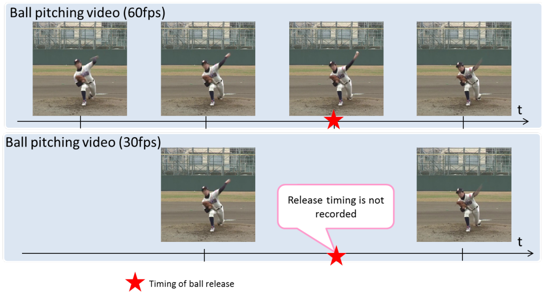
Ball pitching video (60fps)
t
Ball pitching video (30fps)
Release timing is not recorded
t
Timing of ball release

Fig. 1. Example of the difficulty of event timing detection in sports video. The same sports motion is recorded, but depending on the frame rate of the camera, the event to be detected may not be recorded.

To address this need, we have developed a new framework for detecting events with sub-frame accuracy by using a quasi-high frame rate. It is generally very difficult to create high frame rate (30 or more fps) videos with shots taken by a normal video camera. This problem is called temporal super-resolution and is one of the important issues in the field of image processing. For example, there is a method that captures the same scene with multiple cameras and performs temporal-spatial super-resolution by using a spatial alignment method [6]. Another method regards gait as quasi-periodic data series in performing temporal super-resolution [7]. In these methods, however, there are certain restrictions in capturing scenes. Concretely speaking, it is necessary to photograph the same scene with a multiple cameras, or it is necessary to assume that the target image is a quasi-periodic data series. In sports, there are many scenes where the same motion is repeated multiple times. Therefore, the method we propose uses a low-dimensional model to achieve quasi-high frame rate generation. Low dimension motion models can be made by mapping multiple videos of the same motion to low-dimensional space, which makes it possible to generate a quasi-high frame rate sequence in the motion model.

2 Proposed method

The new framework we have developed detects events from video sequences. As mentioned above, it is generally very difficult to create high frame rate (30 or more fps) videos with shots taken by a normal video camera. The method we propose enables low-dimensional space motion models to be taught. It does so by focusing on characteristics in the sports field, in which there are many situations where similar motions are repeatedly performed, and also by mapping multiple motion videos of the same scene in a low dimensional space. In low dimensional spaces, frame rates can be easily increased by using techniques such as linear interpolation and spline interpolation. Event detection is performed on a quasi-high frame rate image created in a low dimensional space to achieve event detection at a level higher than the frame rate of the original video, i.e., sub-frame precision

The proposed method performs the following two steps. First, it generates quasi-high frames by mapping video sequences to a low dimensional space. Next, it detects events in the quasi-high frame rate videos. We will describe these steps in the following sub-sections.

2.1 Quasi-high frame generation in low dimensional space

In this method, first, the video (Y) to be detected is input, then the time series data group is mapped to the low dimensional space and converted into a low dimensional variable X. Then, we create a quasi-high frame low-dimensional variable $X^{h\tilde{i}gh}$ by making a high frame rate in the low dimensional space. This concept is diagrammed in Fig. 2. At this time, selecting a method of maintaining time series information and mapping it to low dimensions makes it possible to easily achieve a quasi-high frame rate in a low dimensional space.

2.1.1 Mapping of low dimensional space

Our method uses the Gaussian Process Dynamical Model (GPDM)[8], which is a method capable of mapping to a low-dimensional space while keeping time series information. GPDM is a method that keeps the information of the time series data and can map it nonlinearly and stochastically to the low dimensional space. It is also used for human motion estimation [9]. It treats latent space dynamics with primary Markov chains, and compresses high dimensional time series $\mathbf{y}(t)$ in t at moments (D dimensional data) to latent variable $\mathbf{x}(t)$. The following expressions define this relationship.

$$\mathbf{y}(t) = g(\mathbf{x}(t); B) + \eta_y(t) \tag{1}$$

$$\mathbf{x}(t) = f(\mathbf{x}(t-1); A) + \eta_x(t) \tag{2}$$

In equation (2), g means the projection from the low dimensional latent space ($\mathbf{X}$) to the high dimensional data space ($\mathbf{Y}$), and A is a model parameter that defines the mapping function. In equation (1), f is a function that defines the dynamics in the low dimensional latent space and B is a model parameter that defines the dynamics function. The terms $\eta_x(t), \eta_y(t)$ in the equations are noise terms. These model parameters learn multiple inputs of the same kind of series data, thus enabling unknown sequence data to be mapped to a low-dimensional space. For example, as shown in Fig. 3, when mapping to a three-dimensional low dimensional space, the coordinate values at time n are $(x(n), y(n), z(n))$.

2.1.2 Quasi-high frame generation

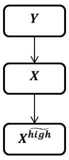

Fig. 2. Conceptual diagram of quasi-high frame generation in low dimensional space. $\mathbf{Y}$ is high dimensional data, i.e., videos. $\mathbf{X}$ is a mapped low dimensional variable. $\mathbf{X}^{\mathbf{h\tilde{i}gh}}$ is a quasi-high framed variable.

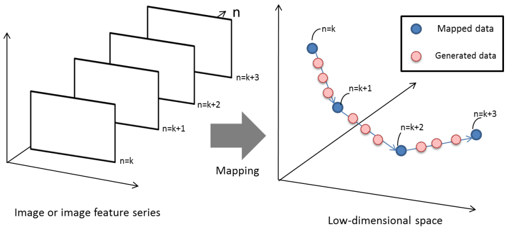

Fig. 3. Quasi-high frame generation in low dimensional space.

Next, it generates high frame rate data($X^{h\tilde{i}gh}$) in the low dimensional space on the space. Here, since the data string is mapped so as to change smoothly in the space, a pseudo high frame rate can be achieved by performing simple linear interpolation. For example, if it is necessary to increase the frame rate by a factor of K, $(x(n{+}1), y(n{+}1), andz(n{+}1))$ from the coordinates $(x(n), y(N+1))$ are divided into K, and the dividing point is set as a new coordinate value of $1 : KN$ frame. A conceptual diagram of this example is shown in Fig. 3. In this figure, the frame rate is increased fourfold.

2.2 Event timing detection in low dimensional space

On the basis of the coordinate value and the event time in the low dimensional space of the previously learned model, the probability P of occurrence of the target event in the low dimensional coordinate value of the test data is calculated. First, the distances between the coordinate value of the event timing in the low-dimensional space learned beforehand and each frame of the pseudo high frame rate converted test data are calculated. Then, with the hypothesis that there is a high possibility that the event timing exists in the frame as the distance becomes closer, the probability $P(k)$ at which the target event occurs at the time k is defined by the following expression. This concept is depicted in Fig. 4.

$$P(k) = exp(-\sum_{n}^{N} /(2\sigma^2)d^2(k)) \quad (3)$$

When learning is performed with multiple image sequences, the same calculation can be performed by accumulating the distances of all the image columns. Here, n represents the index of learning data and N represents the number of learning data elements used for calculation. Using a method of finding the time k that takes the maximum value and a method of finding the weighted average value makes it possible to determine the event timing based on equation (3). We use two methods when accumulating distances of multiple learning data. One is a method using all data used for learning in a low dimensional space (timing detection method (a)). The other is a method using partial data that is filtered by correlations between test data from learning data (timing detection method (b)). In method (b), only N' selected by the threshold value is used as learning data used for calculation in equation (3).

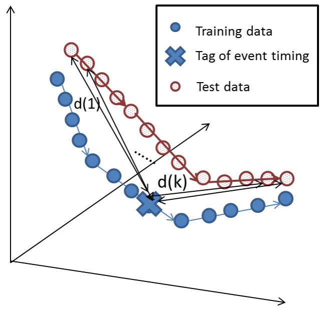

Fig. 4. Event timing detection in low dimensional space. 縲?First, the distance between the coordinate value of the event timing in the low-dimensional space learned beforehand and each frame of the pseudo high frame rate converted test data are calculated. Then, with the hypothesis that there is a high possibility that the event timing exists in the frame as the distance becomes closer, the probability $P(k)$ at which the target event occurs at the time k is defined.

3 Experiments

In order to demonstrate the effectiveness of the proposed method, we conducted an experiment to detect the timing of the ball release using baseball pitcher videos.

3.1 Preparation

In carrying out the experiment, Motion History Image (MHI) [10] was calculated from each image at each time and used as an image feature sequence. Further, before mapping to the low dimensional space, the peripheral time at which the event to be detected occurs from the image feature sequence was taken out as the event region. When learning a low-dimensional space, 10 frames before and after the frame closest to the time with the event tag were taken out and set as an event area, and a low-dimensional space model was created by mapping each of the pitcher videos for each pitcher. This enabled us to make pseudo high frame rates 4 times higher than those in low dimensional space.

The event region in the test data was extracted by template matching. The template matching processing was performed in the following manner. First, a template close to the event to be detected (Fig. 5 (1)) was detected. Next, the templates were compared in the time direction and the spatial direction of the video that shows the event whose detection is desired (Fig. 5 (2)). In addition, we determined the best matching frame and position from the comparison results (Fig. 5 (3)). This is the coordinate value of the central part of the event candidate region for each frame. On the basis of the evaluation value for each frame, we obtained the frame number N^* as the event detection result. Finally, several frames before and after the frame determined as described in (3) were collectively taken as the event area (Fig.5 (4)).

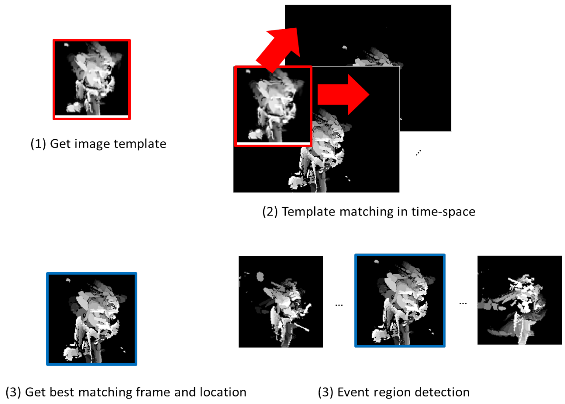
(1) Get image template
(2) Template matching in time-space
(3) Get best matching frame and location
(3) Event region detection

Fig. 5. Event volume detection with template matching.

3.2 Test of detection accuracy using baseball pitching data

We conducted experiments to evaluate the timing at which the release of the ball was detected, using baseball pitching videos for a game involving eight pitchers. The videos used were taken from the back net, and the viewpoint was fixed for the most part. Using the videos taken for each of the eight pitchers, we achieved event detection and quasi-high frame rate in low dimensional space. The total data used comprised 95 videos. And it was used seven balls for each pitcher to learn low-dimensional space. The number of images used for event detection varied from 3 to 7, depending on the pitcher. The frame rate of each image was 30 fps. For timing detection, the two methods described in 2.2 (timing detection methods (a) and (b)) were used. The absolute value of the event detection error for each player determined by experiment is shown in Fig. 6 and Fig. 7. For an image with a frame rate of 30 fps, the estimation error of one frame is about 0.033 seconds. However, according to Fig. 6, at least half of the pitchers get lower estimation accuracy, so it is possible to detect events in sub-frames. However, the estimation error increases depending on the pitcher. The basic idea of the proposed method is to use timing of similar multiple motions. However, since form varies significantly from pitcher to pitcher, we do not think that it can be suitably used to obtain learning data for timing detection. The obtained results lead us to expect that improvement in estimation error can be attained by selecting learning data to be used for timing detection with certain criteria.

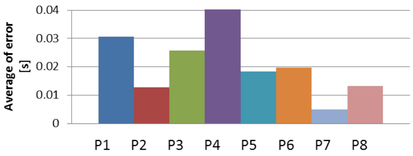

Fig. 6. Average detection error for each player with timing detection method (a) described in 2.2.

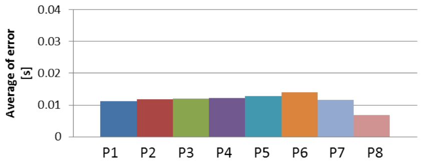

Fig. 7. Average detection error for each player with timing detection method (b) described in 2.2.

According to Fig. 7, by selecting the learning data to be used at the time of event detection, it is possible that most pitchers will be detected with accuracy of less than 0.02 seconds.

4 Discussion

We believe that it is necessary to detect pitcher release timing with sub-frame accuracy to support sports training. However, the tolerance limit of the accuracy of timing detection actually required for sports training is unknown.

Therefore, as an example of an application supporting sports training, we used a system that synchronously displays a pitching image and ball trajectory data obtained by a sensor, and verified the required accuracy of event timing detection. We carried out a subjective evaluation by deviating from manually attached synchronous timing in the same system in five stages ranging from (1. a sense of discomfort) to (5. a sense of naturalness). The subjects were ten ordinary men and women, and we initially presented the correct timing data twice to each of them and instructed them to evaluate it using 5 as a reference value. The stimuli comprised 15 pitchers, both right- and left-handed, for each of whom the amount of synchronization shift was changed. The average results for the subjects in the experiment are shown in Fig 8.

It is clear from this graph that the values from -0.01 to + 0.02 seconds are higher than the evaluation value of 4.5. This shows that the proposed method is applicable to the VR system we have developed.

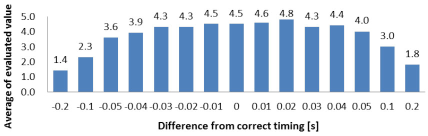

Fig. 8. Average evaluated values for differences from correct timing.

5 Conclusion

In this research, we are aiming at achieving instantaneous event detection in sub-frame units, especially for sports videos. For that purpose, we developed a method to achieve pseudo-high frame rates in low dimensional space and perform event detection in low dimensional space. Experiments conducted using baseball pitching videos confirmed that event detection is possible with an accuracy of 0.02 seconds or less, which exceeds human perception accuracy. In fact, we are developing a virtual reality training system that uses baseball pitching videos and ball trajectory data, and if it is able to achieve automatic synchronization based on learning data it will greatly contribute to provide added convenience to sports players using it. However, there are still problems involved in how players learn and how data to be used for learning should be selected. We are planning to consider these issues in the future by adopting a classification method and a framework for active learning. We have targeted sports videos in the method we developed, but it is not restricted to the types of motions made in sports but can be applied to other actions.

References

[1] Xinguo Yu and Drik Farin.: “Current and Emerging Topics in Sports Video Processing,” *ICME*, 2005.

[2] Gopal Sarma Pingali et.al.: “Gopal Sarma Pingali and Yves Jean and Ingrid Carlbom,” *CVPR*, 1998.

[3] Jurgen Assfalg et.al.: “Semantic Annotation of Sports Videos,” *IEEE Multimedia*, 2002.

[4] Ying Luo et.al.: “Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks,” *CVIU*, 2003.

[5] N. Owens et.al.: “Internationl Conference on Visual Information Engineering,” *Hawk-eye tennis system*, 2003.

[6] Eli Shechtman et.al.: “Space-time super-resolution,” *PAMI*, 2005.

[7] Yasushi Makihara et.al.: “Temporal Super Resolution from a Single Quasi-periodic Image Sequence Based on Phase Registration,” *ACCV*, 2010.

[8] Jack M. Wang et.al.: “Gaussian process dynamical models,” *NIPS*, 2005.

[9] Jack M. Wang et.al.: “Gaussian Process Dynamical Models for Human Motion,” *PAMI*, 2008.

[10] Davis, James W. et.al.: “The Representation and Recognition of Human Movement Using Temporal Templates,” *CVPR*, 1997.

[11] Dan MIKAMI, Mariko ISOGAWA, Kosuke TAKAHASHI, Hideaki TAKADA, and Akira KOJIMA, Immersive Previous Experience in VR for Sports Performance Enhancement, Proc. icSPORTS, 2016.

[12] Ayumi Matsumoto, Dan Mikami, Harumi Kawamura and Akira Kojima, Proficiency Estimation by Motion Variability obtained from Single Camera Input, Proc. icSPORTS, 2013.