Multi-view video synchronization using motion rhythms of human joints

Siqi Sun†, Kosuke Takahashi†, Dan Mikami†, Mariko Isogawa†, Yoshinori Kusachi†

Abstract This paper presents a novel method of estimating temporal offsets between multi-view unsynchronized videos. When synchronizing multiple cameras scattered in a large area with a wide baseline (e.g., a sports stadium, an event hall, etc.), conventional epipolar-based approaches sometimes fail due to the difficulty of robust point correspondences. For such cases, 2D projections of human joints can be robustly associated with each other even in wide baseline videos and can be utilized as corresponding points. However, the detected 2D poses include detection errors in general that cause estimation failures. To address these problems, we introduce the motion rhythm of 2D human joints as a cue for synchronization. The proposed method detects motion rhythms from videos and estimates temporal offsets with the best harmonized motion rhythms. Moreover, we propose a hybrid synchronization algorithm to get sub-frame precision. We demonstrate our method’s performance with indoor and outdoor data.

Key words: Motion rhythms, Video synchronization, Wide baseline, 2D human joints.

1. Introduction

Applications such as athlete tracking, surveillance, and 3D pose estimation often employ multiple cameras to deal with occlusion and to improve area coverage1-5. In such applications, precise video synchronization is a fundamental step and is crucial for further processing.

A hardware-based approach is generally used for multi-view video synchronization6-8. In such approaches, multiple cameras are connected with sync cables and send trigger signals. Audio signals are also utilized in software-based approaches9-11. However, these approaches are not practical when the cameras are scattered in a large area such as a sports stadium or an event hall. In such cases, preparing for very long sync cables is inconvenient, and delays in sound arriving cannot be ignored.

While some video-based approaches12-21 are free from physical constraints or delays in sound arrival, there are some cases in which such video-based approaches are unstable in large area configurations. They need accurate point correspondences detected by local feature descriptors such as SIFT22; matching corresponding points sometimes fails when the cameras are set with wide baselines.

This paper addresses the problem with multi-view video synchronization with wide baselines. The key to this approach is to utilize 2D projections of human joints detected from multi-view videos as corresponding points since they can be robustly associated even in wide baseline cases. The problem is that the detected 2D poses include some detection errors that cause estimation failures of common video-based approaches using epipolar geometry. To address the problem, this paper introduces the motion rhythm of 2D human poses as a cue for synchronization. We define it as a sequence
of timing when each 2D human joint starts to move and stop, as illustrated in Figure 1, which is consistent even from wide baseline views and is robust against detection errors. When these cameras are synchronized, the motion rhythms should strictly harmonize with each other. Therefore, our algorithm first detects 2D poses in multi-view videos using 2D pose estimators, then extracts their motion rhythms. Finally, it outputs the temporal offset with the best harmonized motion rhythms. Our motion-rhythm-based approach is robust against detection errors with 2D poses and synchronizes the multi-view videos even in a frame precision with a large initial time offset. Since the epipolar-geometry-based approach can synchronize them in sub-frame precision with a small initial time offset, we combine these approaches to realize a hybrid algorithm that synchronizes in sub-frame precision even with noised 2D poses and a large initial time offset.

The contributions of this paper are as follows:
1) to introduce a novel idea, motion rhythm of 2D human poses, and to utilize it for video synchronization.
2) to propose a hybrid video synchronization algorithm that consists of a motion-rhythm-based approach and an epipolar-based approach.
3) to show that our method works with a large initial time offset and is robust against 2D pose detection errors.

Experiments demonstrate that our algorithm estimates accurate temporal offset more robustly than other video-based approaches do. This algorithm enables us to synchronize multi-view videos easily and stably even for practical and challenging scenarios such as sporting events or music concerts.

The remainder of this paper is organized as follows. Section 2 shows work related to video synchronization without hardware. Section 3 introduces our algorithm using the motion rhythm of 2D human joints and our hybrid approach. Section 5 reports the performance evaluations, and Section 6 discusses our method. Section 7 concludes the paper.

2. Related Work

This section describes related work that has been done on video synchronization techniques. Multi-view videos are generally synchronized by a hardware-based approach that connects the cameras with sync cables to receive a trigger signal from an external sensor. While the synchronization of this approach is quite accurate, preparing such cables is inconvenient when the cameras are scattered around a large area. To address this problem, we focus on software-based techniques that are categorized in two groups: (a) audio-based synchronization and (b) video-based synchronization.

(a) Audio-based synchronization Audio signals are a good cue for video synchronization. Nils et al. estimated the temporal offset by computing cross correlation between the audio signals of multi-view videos. Prarthana combined audio and visual features of each video. For stadium or concert situations, however, the delay of sound arriving cannot be ignored, and a lot of loud sound sources can degrade estimation precision.

(b) Video-based synchronization Video-based approaches commonly utilize an epipolar constraint, a geometric constraint satisfied by multiple views, to synchronize multi-view videos. Cenek et al. and Noguchi computed a homography or a fundamental matrix from point correspondences based on epipolar geometry and estimated the best temporal offset by minimizing error functions based on them. These approaches struggle to make reliable correspondences if the cameras are set with wide baselines, degrading precision and stability. For such cases, some studies utilized detected 2D human poses as corresponding points because they are robustly associated even in the aforementioned challenging configurations. However, these detected poses sometimes include detection errors that cause estimation failures.

Moreover, epipolar-based video synchronization approaches often run into difficulty in situations where there is a large temporal offset. We assume the existence of local minima can cause the accuracy to deteriorate, although it is known to be difficult to prove through experiments. Among the epipolar-based approaches, Albl et al.’s method is known as one of the best ones in the case of large temporal offset.

Our approach is video-based synchronization that utilizes 2D human poses as corresponding points. We propose a motion-rhythm-based approach that focuses on the temporal structure of 2D human poses because the temporal structure is robust against detection errors and achieves consistency in a wide baseline setup. We also propose a hybrid approach consists of the motion-rhythm-based approach and Albl et al.’s method to get precise sub-frame accuracy results.

3. Proposed Method

3.1 Problem Formulation

As Figure 2 illustrates, $N_c$ cameras are scattered with
wide baselines and capture a human body as a set of articulated joints. The \( j \) \(( j = 1, \cdots, N_j)\) th 3D joint position at time \( t \) \((t = 1, \cdots, N_t)\) and its 2D projection to the image plane of \( i \) \((i = 1, \cdots, N_c)\) th camera \( C_i \) are denoted as \( p_j^t \) and \( q_{ij}^t \), respectively. These 2D projections of human joints can be detected by 2D pose estimators\(^{23}\). In this paper, we use \( C_1 \) as the base camera without losing generality.

Let \( \Delta = \{ \delta_i \} \) denote the temporal differences in a frame scale compared with the base camera. This \( \delta_i \) satisfies

\[
\delta_i = t_i - t_1,
\]

where \( t_i \) is a frame index of the video captured by \( C_i \). The goal of our work here is to estimate appropriate temporal offset \( \delta_i \) from multi-view videos.

### 3.2 Key Idea: Motion Rhythm

Figure 3 shows the trajectories of one set of 2D joint coordinates and the L2-norm in each camera. Since the shapes of these trajectories depend on each camera’s pose and position, they do not match even with the appropriate temporal offset in wide baseline cases. However, the timings in starting and in stopping movement (shown respectively as red and blue circles in the Fig. 3) match even with challenging configurations with the appropriate temporal offset. We call this timing sequence as the motion rhythm of 2D human joints, and utilize it as a cue for synchronization.

Figure 4 illustrates the outline of our algorithm. First, the algorithm detects 2D poses from multi-view videos with 2D pose estimators such as Openpose\(^{23}\). Next, it extracts the motion rhythms of all the joints in each video. Finally, it computes the score of each \( \delta_i \) and outputs the best temporal offset of each camera with the highest score.

1. **Preprocessing**

   First, we apply a noise reduction filter to the 2D coordinates of each joint, then use the L2-norm to represent motion changes in timing sequences. In this study, we apply a median filter and a Savitzky-Golay filter to them.

2. **Extracting Motion Rhythms**

   The L2-norm of 2D coordinates in adjacent frames \( n_{ij}^t \) are used to represent motion changes in every human joint. The \( n_{ij}^t \) can be calculated with the 2D joint coordinates of \( x \) and \( y \) as follows,

\[
n_{ij}^t = \sqrt{\left| x_{ij}^{t+1} - x_{ij}^t \right|^2 + \left| y_{ij}^{t+1} - y_{ij}^t \right|^2}.
\]

As Figure 5 (a) illustrates, if the L2-norm of the \( j \)th joint in \( C_i \) satisfies the two conditions below, we define time \( t_m \) as the timing from “stop” to “move.” Let \( N_{move} \) denote the time window and let \( th_{move} \) denote the threshold for detecting the motion rhythms. We define a time \( t \) as \( t_m \) when the following conditions are satisfied:

**Condition 1:** Between frame \( t \) and \( t - N_{move} \), over \( \gamma \% \) of the L2-norm of \( q_{ij}^t \) is less than \( th_{move} \).

**Condition 2:** Between frame \( t \) and \( t + N_{move} \), over \( \gamma \% \) of the L2-norm of \( q_{ij}^t \) is greater than threshold \( th_{move} \).

Similarly, if the L2-norm of the \( j \)th joint in \( C_i \) satisfies the following two conditions in time \( t \), we define the \( t \) as \( t_s \) (that is, the timing from “move” to “stop”):

\[
\delta_i \]
Fig. 4: Outline of our method. In this example, the motion rhythm of the first joint has the highest score with $\delta_i = 2$ and we output it as the best temporal offset.

Fig. 5: (a) The timing of changing from the “stop” state to the “move” state. (b) The timing of changing from “move” to “stop.”

Fig. 6: First red point is detected as “stop” to “move” timings, black points are ignored.

Condition 1: Between frame $t$ and $t - N_{move}$, over $\gamma$% of the L2-norm $f q^t_{i,j}$ is greater than threshold $th_{move}$.

Condition 2: Between frame $t$ and $t + N_{move}$, over $\gamma$% of the L2-norm $f q^t_{i,j}$ is less than threshold $th_{move}$.

Fig. 7: An example of harmonizing motion rhythms (red point: “stop” to “move” state; blue point: “move” to “stop” state).

Note that when some sequential times (for example, $t, t + 1$) satisfy the conditions for $t_m$ or $t_s$ above, only the first timing $t$ is recognized as the desired timing of motion rhythm and the others are ignored. Figure 6 demonstrates the process in which we exclude black points and red points detected as “stop” to “move” timings.

A sequence of these detected timings of the $j$th joint in a video captured by $C_i$ is defined as the motion rhythm $R_{i,j} = \{t^r_{m_1}, t^r_{s_1}, \cdots; t^r_{m_N}, t^r_{s_N}\}_{i,j}$, where $N_{r_m}$ and $N_{r_s}$ are the total number of detected timings. Note that $R_i$ denotes a set of $R_{i,j}$; that is, $R_i = \{R_{i,j}\}$ in the following.

(3) Harmonizing Motion Rhythms

Suppose motion rhythms $R_1$ and $R_i$ are extracted from two videos of $C_1$ and $C_i$, respectively. When these two cameras are synchronized, the motion rhythms should strictly harmonize with each other. That is, some parts of $R_{1,j} = \{t^r_{m_1}, t^r_{s_1}\}_{1,j}$ and $R_{i,j} =$
\[ \{ t_{m,i}^{r}, t_{s,j}^{r} \}_{i,j} \] match the appropriate temporal offset \( \delta_i \). Figure 7 illustrates an example in which motion rhythms detected from \( C_1 \) and \( C_i \) harmonize with best temporal offset \( \delta_i \). In addition, the motion rhythms of active moving joints tend to precisely harmonize with each other. For example, joints of both feet tend to be active in a soccer juggling data set\(^{24} \).

These observations lead us to define the best temporal offset \( \delta_{i}^{\text{opt}} \) as a temporal offset that maximizes a matching score of motion rhythm as

\[
\delta_{i}^{\text{opt}} = \arg \max_{\delta_{i}} H(R_1, R_i, \delta_{i}), \quad (3)
\]

where the function \( H \) outputs the matching score for each time offset \( \delta_{i} \). The matching score is derived from the largest times of co-occurrences of motion rhythms among those of each joint as follows,

\[
H(R_1, R_i, \delta_{i}) = \max_{j} \left( \sum_{r_{m}} h(t_{m,i,j}^{r}, t_{m,i,j}^{*} + \delta_{i}) \right) + \sum_{r_{s}} h(t_{s,i,j}^{r}, t_{s,i,j}^{*} + \delta_{i})). \quad (4)
\]

The function \( h \) increments the score if the difference in any pairs of all detected timings (that is, \( t_{1} \) from \( C_1 \) and \( t_{i} \) from \( C_i \)) with offset \( \delta_{i} \) is less than a threshold \( \text{th}_{\text{near}} \) as follows:

\[
h(t_{1} + \delta_{i}) = \begin{cases} 1 & (|t_{1} - (t_{i} + \delta_{i})| < \text{th}_{\text{near}}) \\ 0 & (\text{otherwise}) \end{cases}, \quad (5)
\]

where \( \text{th}_{\text{near}} \) denotes number of frames correspondent to 0.033 seconds. For example, in a 30-fps input videos, \( \text{th}_{\text{near}} \) is 1, and in a 120-fps input videos, \( \text{th}_{\text{near}} \) is 4.

### 3.3 Hybrid approach

As Table 1 shows, while our motion-rhythm-based approach is robust against detection errors and is available for a large initial temporal offset, it is not suitable for estimating the time shifts with sub-frame precision.

On the other hand, while epipolar-based approaches sometimes suffer from the detection errors of feature points such as SIFT and a large initial temporal offset. However, they can estimate the temporal offset with sub-frame precision.

Based on these observations, we additionally propose a hybrid approach that take advantage of both approaches, as illustrated in Figure 8.

This approach firstly reduce the temporal offset by using motion-rhythm-based approach. Then, we expect to improve accuracy by applying Albl et al.’s method\(^{17} \) using human poses as corresponding points, where the original Albl et al.’s method\(^{17} \) uses SIFT based correspondence points.

### 4. Preliminary Experiment

This section demonstrates how to decide parameters for our proposed methods. Time window \( N_{\text{move}} \), threshold for detecting the motion rhythms \( \text{th}_{\text{move}} \) and \( \lambda \) described in Section 3.2 are decided based on preliminary experiments as follows.

To determine the values for these, we performed a preliminary experiment using four videos of different scenes to find which parameters could precisely detect ground truth motion rhythms. All four scenes show a single person performing repetitive movements: 1) walking back and forth in a room, 2) bouncing a ball on the ground, 3) throwing and catching a ball, and 4) stepping to the side. We take the first scene (Fig. 9) as an example to show how we decided \( \text{th}_{\text{move}} \). The input video has 1888 frames corresponding to 31.5 seconds.

Figure 10 (a) shows the 2D coordinates of left ankle joint and (b) shows \( L2-norm \) of left ankle joint normalized by human size \( H^1 \). We firstly visually checked the ground truth timings (ground truth timings of changing from the “stop” state to the “move” state marked by red dashed lines.) and found that when \( N_{\text{move}} = 0.3 \) seconds, \( \lambda = 0.7, \text{th}_{\text{move}} = 0.008 \) (navy horizontal line), most of ground truth timings can be automatically detected (red circles).
Fig. 9: A preliminary experiment example to decide parameters for detecting the motion rhythms manually: an input video with a single person walking back and forth in the room.

Fig. 10: Preliminary experiment of Motion rhythm detection for input video shown in Figure 9. (a): the 2D coordinates of left ankle joint and (b): L2−norm of left ankle joint normalized by human size $H^t$. Ground truth timing of changing from the “stop” state to the “move” state is marked by red dashed lines. When $N_{move} = 0.3$ seconds, $\lambda = 0.7$, $th_{move} = 0.008$ (navy horizontal line), detected timings are marked by red circles.

Human size $H^t$ in frame $t$ is calculated using Equation 6. When cameras are set with wide baselines, it is possible for the length of human limbs at each viewpoint to be quite different. We therefore calculate the length of four body parts, and then consider the longest parts to be human size $H^t$. At $t$ ($t = 1, \cdots, N_t$), let

$$H^t = \max(\eta_1^t, \eta_2^t, \eta_3^t, \eta_4^t) \quad (6)$$

5. Evaluations

This section demonstrates the performance of our method with indoor data\textsuperscript{24} and outdoor data in a practical scenario.

5.1 Indoor Scene

(1) Experimental Environment

Figure 11 illustrates the capture setup used in public indoor data\textsuperscript{24}. This setup consists of a set of four calibrated cameras ($C_1$, $C_2$, $C_3$, and $C_4$). These cameras capture 1032×778 resolution videos with 30 fps. In these evaluations, we used the 530 frames of video corresponding to 17 seconds.

In our experiment, we manually shifted 10% and 20% of the total number of frames to simulate the ground truth of the temporal offset, which correspond to 1.8 and 3.5 seconds in time scale, respectively.

(2) Results

Table 2 reports the average estimation errors of camera pairs $C_1C_2$, $C_1C_3$, $C_2C_3$ and $C_3C_4$. For the epipolar-based and hybrid approaches, we use the average result of 10 trials for each pair of cameras to get estimation errors because RANSAC is used in Albl et al.’s method\textsuperscript{17}. According to Table 2, since no correspondent points have been detected, Albl et al.’s method\textsuperscript{17} is not available for this wide baseline scene. When we use 2D human poses as correspondent points into Albl et al.’s method\textsuperscript{17}, the estimation errors are extremely high in cases of large initial temporal offset. However, our motion-rhythm-based approach could estimate temporal offset with errors of the 0.25 to 0.5 frame (that is, 0.008 to 0.017 seconds). Furthermore, our hybrid
Fig. 12: Example of matching scores of indoor data camera pair $C_1C_2$ corresponding to estimation temporal offset: (a) no temporal shift. Ground truth temporal shift is 0 frame and estimation temporal offset is 0 frame. The matching score of 0 frame is 14. (b) Manually created 10% temporal shift. Ground truth temporal shift is 53 frame and estimation temporal offset is also 53 frame. The matching score of 53 frame is 13. (c) Manually created 20% temporal shift. Ground truth temporal shift is 106 frame and estimation temporal offset is also 106 frame. The matching score of 106 frame is 10.

<table>
<thead>
<tr>
<th>Shift percentage (%)</th>
<th>Albl et al.\cite{17}</th>
<th>2D human poses\cite{23} + Albl et al.\cite{17}</th>
<th>Motion rhythm (ours)</th>
<th>Hybrid (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>–</td>
<td>0.19 / 0.006</td>
<td>0.25 / 0.008</td>
<td>0.11 / 0.004</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>36.23 / 1.208</td>
<td>0.50 / 0.017</td>
<td>0.19 / 0.006</td>
</tr>
<tr>
<td>20</td>
<td>–</td>
<td>99.44 / 3.315</td>
<td>0.50 / 0.017</td>
<td>0.10 / 0.003</td>
</tr>
</tbody>
</table>

Fig. 13: Example of matching scores of outdoor data camera pair $C_1C_2$ corresponding to estimation temporal offset. (a) no temporal shift. Ground truth temporal shift is 0 frame and estimation temporal offset is 6 frame. The matching score of 0 frame is 7. (b) Manually created 10% temporal shift. Ground truth temporal shift is 300 frame and estimation temporal offset is 306 frame. The matching score of 306 frame is 7. (c) Manually created 20% temporal shift. Ground truth temporal shift is 600 frame and estimation temporal offset is 606 frame. The matching score of 606 frame is 6.

<table>
<thead>
<tr>
<th>Shift percentage (%)</th>
<th>Albl et al.\cite{17}</th>
<th>2D human poses\cite{23} + Albl et al.\cite{17}</th>
<th>Motion rhythm (ours)</th>
<th>Hybrid (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>–</td>
<td>0.05 / 0.000</td>
<td>5.00 / 0.042</td>
<td>1.93 / 0.016</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>288.35 / 2.403</td>
<td>2.67 / 0.022</td>
<td>2.02 / 0.017</td>
</tr>
<tr>
<td>20</td>
<td>–</td>
<td>605.84 / 5.049</td>
<td>3.00 / 0.025</td>
<td>1.08 / 0.009</td>
</tr>
</tbody>
</table>

approach improves the results with error of 0.10 to 0.19 frame (that is, 0.003 to 0.006 seconds). Figure 12 illustrates temporal changes in matching scores. Each graph shows the case of different temporal offsets, (a)
is for 0, (b) is for 53, and (c) is for 106. Because the higher matching score means better harmonization, the highest score marked by blue circle is selected as estimation result. For example, in Figure 12 (a), estimated time 0 shows the highest matching score 14, which is denoted by (0, 14). Thus, the estimation temporal offset here becomes 0.

These results show that our methods attain the highest matching score at the appropriate temporal offsets even with the large initial temporal offset. From the above, we can conclude that 2D poses as corresponding points are effective for multi-view video synchronization with wide baselines. Moreover, our motion-rhythm-based approach works with large initial temporal offsets and is robust against 2D pose detection errors.

5.2 Outdoor Scene

(1) Experimental Environment

Figure 14 shows the configuration of the evaluation that used real outdoor data. Three cameras with 1920×1080 resolution and 120 fps: $C_1$, $C_2$ and $C_3$ (Casio EX100) were set with wide baselines. These cameras captured one player throwing a ball. Each input video has 3,000 frames corresponding to 25 seconds.

(2) Results

Table 3 reports the average errors of camera pairs $C_1C_2$, $C_1C_3$, and $C_2C_3$. The Albil et al.’s method17 fails with a large initial temporal offset and indoor data but can estimate the appropriate offset with a small initial offset. On the other hand, the motion-rhythm-based approach estimates the temporal offset with around 3 to 5 frames errors (that is, 0.042 to 0.025 seconds), and the hybrid approach outperforms them with around 1 to 2 frames errors (that is, 0.009 to 0.016 seconds). Figure 13 illustrates temporal changes in matching scores as well as Figure 12. These results show that our method attains the highest matching score with temporal offsets within 6 frames of their ground truth in all settings. As Figs. 12 and 13 show, no temporal offsets with a high matching score are comparable to the highest one, so the motion-rhythm-based approach does not fall into the local minima and can robustly estimate the appropriate matching score even with large initial temporal offsets and detection errors. These results prove our methods are robust even in practical outdoor scenes.

6. Discussion

(a) Robustness to 2D joint noise

In order to investigate how the accuracy of the correspondence affects to the final accuracy of synchronization, we performed an additional evaluation in which noise was added to the detected 2D poses. The experimental conditions are as follows. The input data ($C_0C_1$
Table 4: Synchronization accuracy when adding zero-mean Gaussian noise with standard deviation(SD) 5, 10, 15 to 2D joints. The ground truth value of ten trials is 53 frame.

<table>
<thead>
<tr>
<th>SD</th>
<th>Approach</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
<th>Trial 4</th>
<th>Trial 5</th>
<th>Trial 6</th>
<th>Trial 7</th>
<th>Trial 8</th>
<th>Trial 9</th>
<th>Trial 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Motion rhythm</td>
<td>53</td>
<td>52</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>52</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>52.91</td>
<td>53.10</td>
<td>53.12</td>
<td>53.37</td>
<td>52.88</td>
<td>52.83</td>
<td>53.12</td>
<td>52.94</td>
<td>52.31</td>
<td>53.18</td>
</tr>
<tr>
<td>10</td>
<td>Motion rhythm</td>
<td>53</td>
<td>54</td>
<td>52</td>
<td>53</td>
<td>52</td>
<td>-37</td>
<td>53</td>
<td>51</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>53.35</td>
<td>53.11</td>
<td>52.51</td>
<td>52.69</td>
<td>53.65</td>
<td>-69.90</td>
<td>52.71</td>
<td>52.49</td>
<td>53.55</td>
<td>52.42</td>
</tr>
<tr>
<td>15</td>
<td>Motion rhythm</td>
<td>12</td>
<td>50</td>
<td>48</td>
<td>52</td>
<td>52</td>
<td>5</td>
<td>22</td>
<td>6</td>
<td>52</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>8.65</td>
<td>54.33</td>
<td>53.13</td>
<td>53.39</td>
<td>53.05</td>
<td>1.55</td>
<td>-26.65</td>
<td>53.11</td>
<td>52.52</td>
<td>52.64</td>
</tr>
</tbody>
</table>

Fig. 16: Left foot of y-coordinate from camera $C_1$, $C_2$, and $C_3$.

camera pair of indoor data) is perturbed by the addition of zero-mean Gaussian noise whose standard deviation $\sigma$ are 5, 10, 15. The input data has 10 % temporal offset (53 frames, 1.77 seconds). Since the experiment results vary depending on random Gaussian noise, the results of 10 trials are shown in Table 4. The results present us with two key findings:

(1) When the noise for 2D joint detection increases, the synchronization accuracy of the Motion rhythm-based method decreases significantly. This is because Motion rhythms can not be detected correctly due to the noise. Figure 15 shows an example of detection omission, where ground truth timings of changing from the “stop” state to the “move” state are indicated by red dashed lines. The timing detected correctly when $\sigma = 0$, 5 but miss detected when $\sigma = 10$, 15 is indicated by a blue triangle.

(2) Even if the noise increases, the Hybrid approach improves synchronization accuracy compared to the Motion-rhythm-based approach under the same conditions. In other words, the accuracy improvement by the Albl et al.’s method$^{[17]}$ is effective even if the amount of noise becomes large. From the above, we expect that the hybrid approach can improve accuracy even in sequences where a large amount of noise is added to 2D joint positions.

(b) Comparison to 1D signal approaches

We could consider human joint coordinates as 1D audio time series signals. In some papers$^{[9][25]}$, cross-correlation synchronizes audio signals; other methods are based on dynamic time warping (DTW)$^{[26][27]}$. These methods rely on a similar audio signal or features extracted from the signal. However, human joint coordinates detected from wide baselines sometimes change quite differently. Figure 16 illustrates the y-coordinate of the left foot in $C_1$, $C_2$, and $C_3$. While the y coordinate serials of interval (A) from $C_1$ and interval (B) from $C_3$ are similar, those of interval (A) and interval (C) from $C_2$ are extremely different and even look like inverse signals. Therefore, 1D signal processing methods for synchronization are unsuitable for camera configurations with wide baselines.

(c) Limitations

(1) Our approaches only work for human action scenes with individuals. Scenes with a person performing no actions or actions with periodic motions such as juggling are unsuitable. (2) The camera setup needs to be fixed, which is not available for moving cameras.

7. Conclusion

We proposed a novel method of estimating the temporal offset between multi-view unsynchronized videos by using motion rhythms of 2D human poses as a cue for synchronization with a wide baseline camera setting. Motion rhythm is a sequence of timing when each 2D human joint starts to move and stop, which is consistent even from wide baseline views and is robust against detection errors. Our experiments proved our motion-rhythm-based approach is robust against detection er-
rors of 2D poses and can obtain frame-level accuracy even with a large initial time offset. Moreover, our hybrid synchronization approach that combines epipolar- and motion-rhythm-based approaches is more precise. Future work involves applying our methods to groups of people.

References

15) F. Zhou and F. De la Torre : “Generalized time warping for multi-modal alignment of human motion”, IEEE Conference on Computer Vision and Pattern Recognition, pp.1282-1289 (June 2012)
24) N. Luca and M. C. Guido : “Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouette”, International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission (June 2018)
Yoshinori Kusachi is a senior research engineer at the Universe Data Handling Laboratory of NTT Media Intelligence Laboratories. He received a Master’s from the Nara Institute of Science and Technology and joined NTT in 1997. He has been engaged in research and practical application development in the fields of image processing, computer vision, and pattern recognition. He received a Ph.D in Information Engineering from the Nara Institute of Science and Technology in 2007.