sensors

MDPI

Article

HFR-Video-Based Stereo Correspondence Using High Synchronous Short-Term Velocities

Qing Li, **Shaopeng Hu**, **Kohei Shimasaki** and **Idaku Ishii** *

Smart Robotics Laboratory, Graduate School of Advanced Science and Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan
* Correspondence: iishii@robotics.hiroshima-u.ac.jp; Tel.: +81-82-424-7692; Fax: +81-82-422-7158

Abstract: This study focuses on solving the correspondence problem of multiple moving objects with similar appearances in stereoscopic videos. Specifically, we address the multi-camera correspondence problem by taking into account the pixel-level and feature-level stereo correspondences, and object-level cross-camera multiple object correspondence. Most correspondence algorithms rely on texture and color information of the stereo images, making it challenging to distinguish between similar-looking objects, such as ballet dancers and corporate employees wearing similar dresses, or farm animals such as chickens, ducks, and cows. However, by leveraging the low latency and high synchronization of high-speed cameras, we can perceive the phase and frequency differences between the movements of similar-looking objects. In this study, we propose using short-term velocities (STVs) of objects as motion features to determine the correspondence of multiple objects by calculating the similarity of STVs. To validate our approach, we conducted stereo correspondence experiments using markers attached to a metronome and natural hand movements to simulate simple and complex motion scenes. The experimental results demonstrate that our method achieved good performance in stereo correspondence.

Keywords: high-speed vision; stereo correspondence; motion information; high synchronous velocity

check for updates

Citation: Li, Q.; Hu, S.; Shimasaki, K.; Ishii, I. HFR-Video-Based Stereo Correspondence Using High Synchronous Short-Term Velocities. *Sensors* **2023**, *23*, 4285. https://doi.org/10.3390/s23094285

Academic Editor: Cosimo Distante

Received: 20 March 2023
Revised: 19 April 2023
Accepted: 23 April 2023
Published: 26 April 2023

1. Introduction

Stereo vision offers a straightforward way for computers to comprehend the world and can reconstruct the three-dimensional geometric information of scenes [1]. It is widely used in various fields such as autonomous navigation systems for mobile robots [2], aerial and remote sensing measurements [3], medical imaging [4], SLAM [5], and more. Stereo correspondence is a crucial element of stereo vision that plays a vital role in finding corresponding point pairs between two images to calculate the depth information of the stereo image [6].

The goal of this study is to achieve stereo correspondence for multiple moving objects with similar appearances. Over the past few decades, extensive research has been dedicated to stereo correspondence. Traditional stereo correspondence algorithms can be categorized into local, global, and semi-global methods. These methods use manually extracted features, such as sum of absolute difference (SAD) [7], normalized cross-correlation (NCC) [8], SIFT (Scale-Invariant Feature Transform) [9], and ORB (Oriented FAST and Rotated Brief) [10], to provide similarity measures between left and right image patches. However, the performance of traditional stereo correspondence methods is severely limited by the handcrafted features used in the cost function. In Ref. [11], convolutional neural networks (CNNs) were first introduced for stereo correspondence, demonstrating advantages in both speed and accuracy over traditional methods. Currently, deep learning-based image similarity measurement methods mainly rely on feature extraction from deep networks [12] and similarity comparison through metric learning [13].

However, appearance-based correspondence methods face significant challenges due to variations in camera viewpoints, lighting conditions, and pose changes [14]. Motion information, on the other hand, is independent of object appearance and exhibits excellent performance in scenes with similar appearances and drastic changes in appearance. Currently, a significant amount of research has been devoted to cross-camera multi-object correspondence based on motion information [15,16]. Existing motion similarity measurement methods can be divided into two categories: spatial similarity and spatio-temporal similarity [17]. Spatial similarity only considers the same geometric shape and ignores the temporal dimension, which is not suitable for real-time stereo correspondence systems. The update of motion information is delayed due to the limited speed of traditional visual image input (30 or 60 fps) [18], making trajectory synchronization of high-speed moving objects difficult. However, high-speed vision sensors operate at hundreds or even higher frequencies, enabling them to observe moving objects and capture phase differences with extremely low latency [19]. Additionally, viewing angles significantly affect trajectory matching performance. First-order motion velocity and second-order acceleration directions are relatively insensitive to viewing angles.

In this study, we propose a high-speed stereo correspondence system for multiple moving objects with similar appearances, based on their high synchronous velocities. We designed stereo correspondence experiments for moving objects with different frequencies and amplitudes, as well as for high-speed moving hands with occlusions. The subsequent parts of this study are organized as follows: related works and research are presented in Section 2; a detailed algorithm analysis and concept illustration are presented in Section 3; Section 4 presents a full description of the experimental test platform, followed by a discussion of the test results; and finally, the conclusions are presented in Section 5.

2. Related Works

This study aims to correspond multiple moving objects with similar appearances in a stereoscopic video, which is closely related to research on image similarity measurement and trajectory similarity measurement. In the following sections, we will provide a brief review of related works.

2.1. Image Similarity Measurement

The computation of image matching serves as the initial step in stereo correspondence, relying primarily on the similarity of target pixel blocks surrounding the stereo images. Over time, the measurement of similarity between image blocks has evolved from region-based approaches to feature-based approaches, and finally to deep learning techniques.

Region-based matching methods can be classified into two categories. The first approach minimizes differences in pixel information by using methods such as cross-correlation [20], mean square error (MSE) [21], and mutual information [22]. The second approach transforms images from the time domain to the frequency domain and performs similarity analysis in the transformed domain using techniques such as Fourier transform [23], Walsh transform [24], and wavelet transform [25]. However, region-based image matching methods require high-quality images because noise, lighting, and changes in shape can greatly affect the quality of the match. Feature-based methods can significantly reduce the impact of image quality on similarity and have been extensively researched to date [26]. These features are often manually designed, such as SURF [27], ORB [28], and LBP [29]. Feature-based methods require additional computational power to find matching points with similar features between image blocks. The Structural Similarity Index (SSIM) combines brightness, contrast, and structure to achieve matching results similar to human visual perception and has been widely used for comparing image similarity [30].

Recently, convolutional neural networks (CNNs) have replicated the huge success in image recognition and have become a research hotspot in image region matching. Based on CNNs, image matching can be mainly divided into two research directions: (1) using deep networks such as ResNet [31] and VGG [32] to extract image features and then using similarity metrics such as Euclidean distance and cosine distance to measure the similarity of high-dimensional features; (2) using metric learning to directly output the similarity of two image blocks. In Ref. [33], the ResNet model was used to extract periocular features from different spectral bands, and cosine similarity was used for image verification, achieving high accuracy. In Ref. [34], the VGGNet was used to extract multi-scale features from segmented patches and achieved detection of forged images. Compared to manually extracted features, features extracted by CNNs are more effective in handling noise and morphological changes. In Ref. [35], MatchNet was proposed, which uses CNN for region feature extraction and then computes similarity through a three-layer fully connected network. The DeepCompare method was proposed in Ref. [36], which improved the performance of the Siamese network using the Center-Surround Two-Stream Network and Spatial Pyramid Pooling (SPP) [37]. DeepCD based on the Triplet network was proposed in Ref. [38]. This method describes image patches as complementary descriptors and improves the performance in various applications. Currently, methods based on deep learning are difficult to output calculation results in extreme time and are not suitable for high-speed vision systems. However, the matching performance they provide is unmatched by traditional algorithms.

2.2. Matching Based on Motion

When objects are well tracked under good conditions of a single camera, their motion information is less affected by lighting, shape changes, and noise. Motion-based matching has been widely used in cross-camera multi-object matching, such as in smart traffic [39], user behavior analysis [40], and motion pose estimation [41].

There are various ways to represent motion information, such as trajectories, angles, and velocities. Trajectories, as an easily obtainable form of motion information, have been widely used in multi-object tracking. Trajectories can be classified into two types: sequence-only trajectories and spatiotemporal trajectories, depending on whether the temporal property is considered [42].

Different methods have been developed for measuring the similarity between different target trajectories, which are mainly divided into three directions: distance-based, feature-based, and deep learning-based trajectory similarity calculation methods. Distance-based trajectory similarity calculation methods mainly measure the similarity between trajectories by calculating the distance between trajectory points. Some classic methods include Dynamic Time Warping (DTW) [43], Edit Distance on Real sequence (EDR) [44], and Longest Common Subsequence (LCSS) [45]. For instance, LCSS is used to calculate the similarity of the 3D GPS trajectories of the trucks in Ref. [46] to identify the movement patterns of the trucks. In Ref. [47], a trajectory evaluation method based on Dynamic Time Warping was proposed to evaluate the discrepancy between robot trajectories and human motion. However, these methods have limitations in dealing with data noise and missing values.

Feature-based trajectory similarity calculation methods extract features from trajectories and then calculate the similarity between features to measure the similarity between trajectories. Some classic methods include Shape Context [48], Histogram of Oriented Gradients (HOG) [49], and Global Alignment Kernel (GAK) [50]. For example, a skeleton-based action recognition method is proposed in Ref. [51], which combines trajectory images and visual features to simulate human actions. Based on the Fréchet distance, a shape-based local spatial association metric is proposed in Ref. [52] for detecting anomalous activities of moving ships. However, these methods are more complex in feature extraction and computation, and require a larger amount of computation.

Deep learning-based trajectory similarity calculation methods use machine learning to model and learn trajectory data, and then calculate the similarity between trajectories. Some classic methods include neural network-based methods, decision tree-based methods [53], and support vector machine-based methods [54]. For instance, an RNN-based Seq2Seq autoencoder model is proposed in Ref. [55], which improves the calculation of similarity. In Ref. [56], an attention-based robust autoencoder model is proposed, which learns low-dimensional representations of noisy ship trajectories. An unsupervised learning method is proposed in Ref. [57], which can automatically extract low-dimensional data features through convolutional autoencoders. The similarity between trajectories can be obtained from the similarity between low-dimensional data, which ensures high-quality trajectory clustering performance. However, these methods require a large amount of training data and computation resources, but they offer higher accuracy and robustness in trajectory similarity calculation.

3. HFR Stereo Correspondence Based on High Synchronous Short-Term Velocities

3.1. Concept

As mentioned in previous sections, matching multiple moving objects with similar appearances in stereoscopic video is a significant challenge. To address this issue, we propose a High Frame Rate (HFR) stereo vision system, as depicted in Figure 1. The entire process of stereo correspondence for multiple objects is divided into two steps: independent multiple-object tracking and stereo correspondence based on high synchronous short-term velocities. In the independent multiple-object tracking step, we define the pixel-scale movement of an object between HFR frames as its velocity, which comprises horizontal and vertical components. As shown in Figure 1a, we utilize n velocities over a period of time before the current time as the motion feature of the objects, referred to as short-term velocities. In the object stereo correspondence step, we analyze the similarity between the high synchronous short-term velocities of multiple objects frame-by-frame to establish correspondences among different objects, as illustrated in Figure 1b.

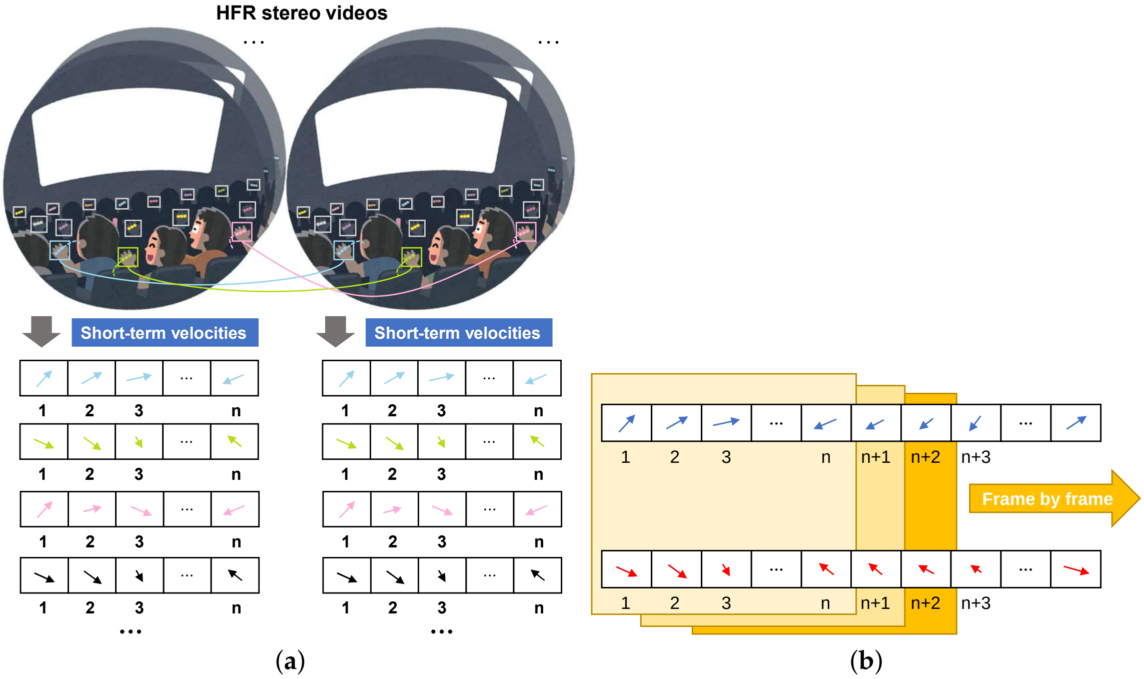


Figure 1. Concept of stereo correspondence based on high synchronous short-term velocities. (**a**) Motion features composed of short-term velocities. (**b**) Correspondence based on short-term velocities.

3.2. Independent Multiple-Object Tracking in HFR Stereoscopic Video

In this study, we conducted offline experiments using the HFR stereoscopic videos to validate the effectiveness of our algorithm. The first step involves the fast tracking of multiple objects using the HFR stereo camera, which enables the real-time update of the motion positions and velocities of the objects. However, HFR stereoscopic videos not only provide more image information but also impose a higher computational burden on multiple object tracking. HFR stereoscopic videos usually run at 200 frames per second or higher, leaving us with only 5 milliseconds or less for computation. However, detectors that yield good detection performance usually require longer running times. For instance, in this study, the hand detection using MediaPipe takes approximately 30 milliseconds, while the marker detector takes about 10 milliseconds. Therefore, we proposed a hybrid tracking approach that combines object detection with template matching to enable the tracking of multiple objects with very low processing time. This approach exhibits good tracking performance for objects with drastic appearance changes due to the constantly updated object templates. Due to the low latency of high frame rate (HFR) videos, the motion speed of objects between frames is relatively low. To quickly locate objects near their image blocks, template matching can be utilized. Figure 2 illustrates the hybrid detection method based on template matching and object detection. The time interval between input HFR images is denoted as τ milliseconds. The detector continuously performs object detection with a time interval of δ milliseconds, where $\delta(\delta > \tau)$ represents the processing time of the detector. The detection results $D(I_t)$ obtained from the detector in the input image I_t at time $t = k \times \delta$ $(k = 0, 1, 2, \dots)$ can be expressed as follows:

$$D(I_t) = \{d_t^1, d_t^2, \dots, d_t^l, \dots, d_t^L\}(l = 1, 2, \dots, L). \tag{1}$$

Each detection result d_t^l comprises six parameters:

$$d_t^l = \{x_l, y_l, w_l, h_l, p_l, c_l\}. \tag{2}$$

x_l, y_l, w_l, and h_l represent the starting image coordinates, width, and height of the l-th object image block, respectively. p_l and c_l represent the confidence score and category of the detection result, respectively. As indicated in Figure 2, we obtained object templates T_t updated at time intervals of δ.

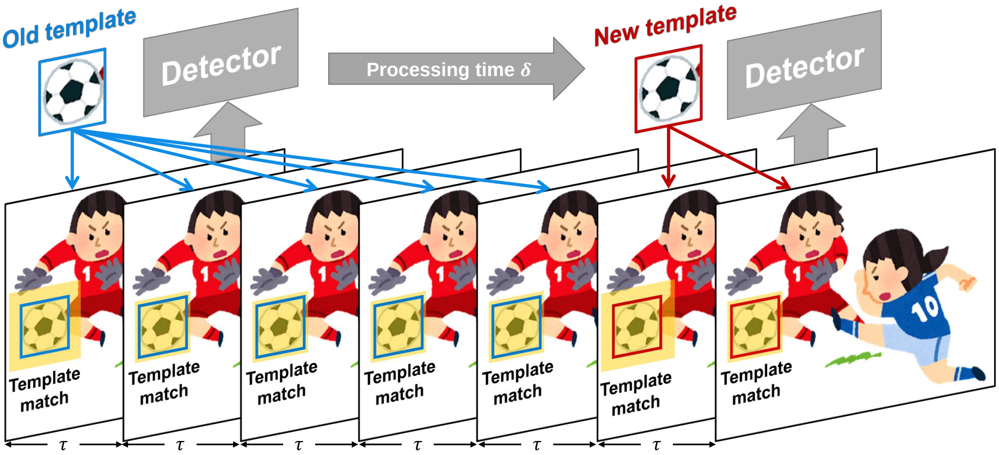


Figure 2. Hybrid detection method based on template matching and object detector.

Simultaneously, we perform template matching using the most recently updated templates to detect objects at time intervals of τ. In high-speed visual systems where the system's operational speed is a priority, a trade-off between speed and accuracy is often necessary. Therefore, we employ the sum of absolute differences (SAD) as the similarity metric for image-template matching. The detection process for objects between adjacent HFR frames is as follows:

$$P_l(t) = P_l(t - \tau) + \underset{|x| \le R, |y| \le R}{\arg\min} \; E(x, y), \tag{3}$$

$$E(x, y) = \sum_{x', y'} (T_l(x', y') - I_t(x'_t + x + x', y'_t + y + y')). \tag{4}$$

$P_l(t - \tau)$ and $P_l(t)$ represent the coordinates of the center of the l-th object in the previous and current frames, respectively. T_l is the template image of the l-th object that is most recently updated. I_t represents the region of interest (ROI) being searched in the current image, as highlighted in yellow in Figure 2. $(x't, y't)$ represents the top-left point coordinate of the ROI region in the current image. R is the search range of the template matching. To mitigate the impact of object appearance changes on tracking, we perform template updates by searching in a larger region each time, as depicted in the yellow area in the figure.

In this work, we employ a distance matrix Φ between I objects in the previous frame and J objects in the current frame as a replacement for the Intersection over Union (IOU) method for object tracking.

$$\Phi = \begin{bmatrix} \psi(1,1) & \psi(1,2) & \cdots & \psi(1,J) \\ \psi(2,1) & \psi(1,2) & \cdots & \psi(2,J) \\ \cdots & \cdots & \cdots & \cdots \\ \psi(I,1) & S(I,2) & \cdots & \psi(I,J) \end{bmatrix}. \tag{5}$$

$\psi(i, j)$ represents the Euclidean distance between the i-th object in the previous frame and the j-th object in the current frame, measured in pixels. We employ the Hungarian matching algorithm to obtain tracking results quickly and efficiently.

In high-speed imaging, where object motion is relatively slow and motion between adjacent frames is approximately uniform, we use a Kalman filter for optimal estimation of motion. The Kalman filter can also be used for short-term motion prediction when object detection is temporarily lost.

3.3. Correspondence Based on High Synchronous Velocities

3.3.1. Velocity-Based Correspondence

Once the optimal tracking state of the object is obtained, we can obtain highly synchronized spatiotemporal velocities (STVs). As shown in Figure 3, we sampled the velocities of the object at the pixel scale within N high-speed frames to extract the motion feature of the object. The STVs V of the object were then obtained as follows:

$$V = \{v_{N-1}, \ldots, v_n, \ldots, v_1, v_0\}, (n = 0, 1, \ldots, N-1), \tag{6}$$

where $v_n = [dx_n, dy_n]$ is the velocity vector at the pixel scale in the n-th frame before the current frame.

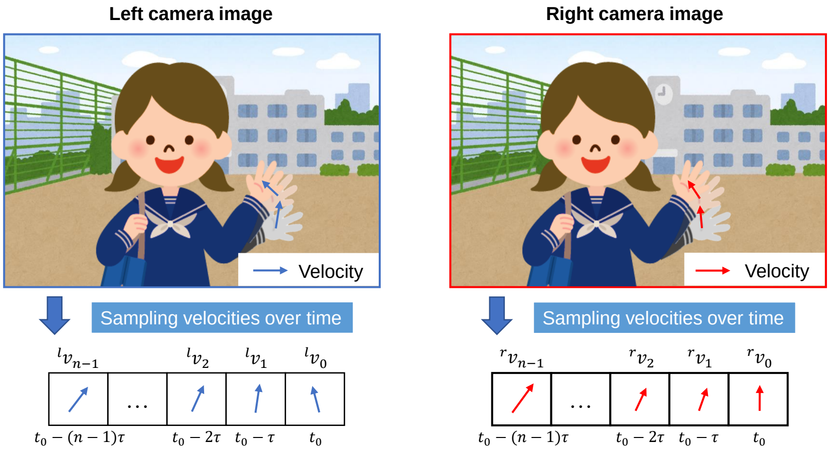


Figure 3. Sampling velocities over time in HFR stereoscopic video.

In this study, we propose the concept of the scale cosine distance. While the calculation of the cosine distance yields the cosine of the angle between both vectors, which is close to 1 when the angle is small, the cosine distance does not consider the length of the vector. This means that two parallel vectors with different lengths would have a cosine distance of 1, even though their similarity is very low. To overcome this limitation, we introduce the scale cosine distance s between vectors A and B, which takes into account the length of the vector, as expressed below:

$$s = \frac{A \cdot B}{\max(|A|, |B|)^2}, \tag{7}$$

where $|A|$ and $|B|$ are the modulo lengths of vectors A and B, respectively. When the lengths of both vectors are similar and the included angle is small, the scale cosine distance is larger, with a higher similarity close to 1.

Hence, for the N-dimensional high-synchronization STVs ${}^l V_i$ and ${}^r V_j$ extracted from the left and right HFR stereo cameras, we calculated the scale cosine similarity $S_v(i,j)$ between them as follows:

$$S_v(i,j) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{{}^l v_k \, .^r \, v_k}{\max\left(|{}^l v_k|, |{}^r v_k \,| \right)^2}. \tag{8}$$

3.3.2. Direction-Based Correspondence

The correlation of velocity decreases in the presence of a large viewing angle in the HFR stereo camera. The correlation between the direction of velocity change and the change in camera viewing angle is relatively small. We extract the cosine values of the angle changes between velocities to form a short-term angle for measuring the similarity A of direction changes.

$$A = \{a_{N-2}, \ldots, a_n, ..., a_1, a_0\}, (n = 0, 1, \ldots, N-2). \tag{9}$$

a_n is the cosine value between adjacent velocity angles,

$$a_n = \frac{v_{n+1} \cdot v_n}{|v_{n+1}| \cdot |v_n|}, (n = 0, 1, \ldots, N-2). \tag{10}$$

Hence, for the $(N-1)$-dimensional high-synchronization STVs ${}^l A_i$ and ${}^r A_j$ extracted from the left and right HFR stereo cameras, we calculated the direction similarity $S_a(i,j)$ between them as follows:

$$S_a(i,j) = 1 - \frac{1}{2(N-1)} \sum_{i=0}^{N-2} |{}^l a_k - {}^r a_k|. \tag{11}$$

3.3.3. Mixed Correspondence

The similarity measure of object motion is contributed by both the similarity of velocities and the similarity of velocity change directions. We define the mixed similarity $S(i,j)$ between the short-term velocities of the i-th target in the left camera and the j-th target in the right camera as follows:

$$S(i,j) = \omega_v S_v(i,j) + \omega_a S_a(i,j), \tag{12}$$

$$\text{s.t. } \omega_v + \omega_a = 1. \tag{13}$$

where ω_v and ω_a are scale factors that reflect the contribution of velocity and direction to the similarity metric in different camera perspectives. Generally, when the HFR stereo camera has a large field of view, the direction similarity $S_a(i,j)$ should contribute a larger proportion. Finally, based on the mixed similarity of short-term velocities, a bipartite graph S can be reconstructed for I targets in the left camera and J targets in the right camera,

$$S = \begin{bmatrix} S(1,1) & S(1,2) & \cdots & S(1,J) \\ S(2,1) & S(1,2) & \cdots & S(2,J) \\ \cdots & \cdots & \cdots & \cdots \\ S(I,1) & S(I,2) & \cdots & S(I,J) \end{bmatrix}. \tag{14}$$

Using the Hungarian matching algorithm, we can easily obtain the correspondence relationship based on motion information.

4. Experiment

The proposed stereo correspondence algorithm was implemented offline using an HFR stereo camera system that operated at a speed of 200 fps. The system was composed of two high-speed USB 3.0 camera heads from Imaging Source Corp. (DFK 37BUX273, Germany) and a personal computer. The cameras were compact, measuring 36 × 36 × 25 mm in size, weighing 70 g, and had no mounted lens. They were capable of capturing and transferring 10-bit color images of 1440 × 1080 pixels to RAM at a rate of 238 fps via a USB 3.0 interface. We used a PC with the following hardware specifications to record the HFR stereoscopic video: Intel Core i9-9900K @ 3.2 GHz CPU, 64 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU.

To evaluate the performance of our stereo correspondence algorithm, we analyzed HFR stereo offline videos that were captured at a rate of 200 fps ($\tau = 5$ ms) with a 2-ms exposure time. In this study, we chose the hand as the detection target because it had a high similarity in texture and color across different people, and moved at a high speed relative to other body parts, making it difficult to use appearance-based methods for correspondence. We conducted three experiments to evaluate our algorithm: stereo correspondence evaluation, correspondence of fast-moving hands, and correspondence in a meeting room scene. For the hand detection task, we used Google's MediaPipe toolkit, which provided accurate and rapid hand detection.

4.1. Stereo Correspondence Evaluation

We conducted an evaluation of the correspondence performance of our HFR stereo correspondence algorithm when implemented offline in our system. Figure 4 illustrates the experimental setup for the stereo correspondence evaluation, where two metronomes were fixed 800 mm away from the HFR stereo camera. The small metronomes operated at frequencies of 3.0 and 2.6 Hz, respectively. OpenCV-generated markers were attached to different positions on the pointers of both metronomes. As a result, markers on a similar pointer exhibited similar movements when shaking, but with different magnitudes of movement. During the operation of the metronomes, we captured a 200-fps HFR stereoscopic video using 12-mm lens fixed cameras. The positions of the individual markers were easily detected using OpenCV.

In Figure 5, we show the input stereo images of size 1440 × 1080 pixels, with the correspondence results at intervals of 0.06 s for t = 7.00∼7.25 s. After applying the stereo correspondence algorithm, the same marker in the HFR stereoscopic video was marked with numerical symbols of a similar color. The xy coordinate values of the image centroids of the markers in the left HFR stereoscopic video are presented in Figure 6. From the image, markers 0 and 1 exhibited similar movement with different magnitudes than markers 2 and 3. The mixed similarities of the moving markers' STVs over time are shown in Figure 7. Figure 7a–d depict the mixed similarities between markers 0, 1, 2, and 3 in the left HFR stereo image and those in the right HFR stereo images, respectively. The graph indicates that similar markers in the HFR stereo images have a high degree of similarity, which is almost greater than 0.8. Markers 0 and 1 on a similar pointer have a similar angular velocity, but different linear velocities. However, our scale cosine distance includes a scale factor that can easily distinguish between markers 0 and 1. The same applies to markers 2 and 3. We also considered the effect of the duration of STVs on multi-object stereo correspondence. Figure 8 presents the results of stereo correspondence using a 30 fps stereo camera in the same scene. It is evident that marker 0 and marker 3 do not match in the correspondence. Figure 9 shows the short-term velocity features of marker 0 within a 0.3-second interval in the stereo camera at t = 7.710 s. It is evident that traditional low-speed cameras have synchronization issues when tracking fast-moving objects. Velocity information is delayed by approximately 30 milliseconds, which significantly affects the correspondence results, especially when the object changes direction frequently. Figure 10 shows the short-term velocity features of marker 0 within a 0.3-second interval in the HFR stereo camera. In contrast, the HFR camera not only provides more motion information in a short time but also has much higher synchronization.

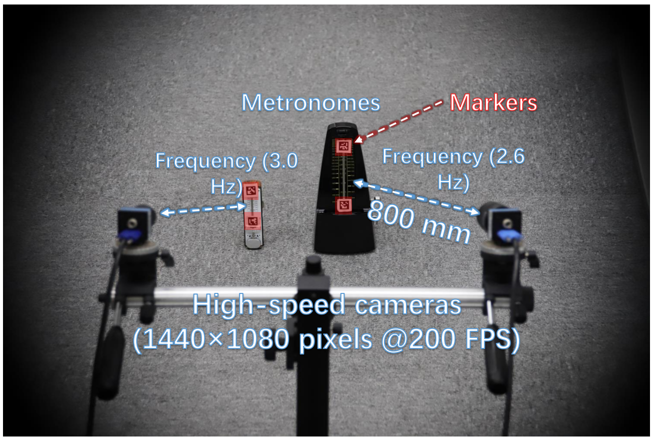


Figure 4. Experiment setup for similar motion correspondence.

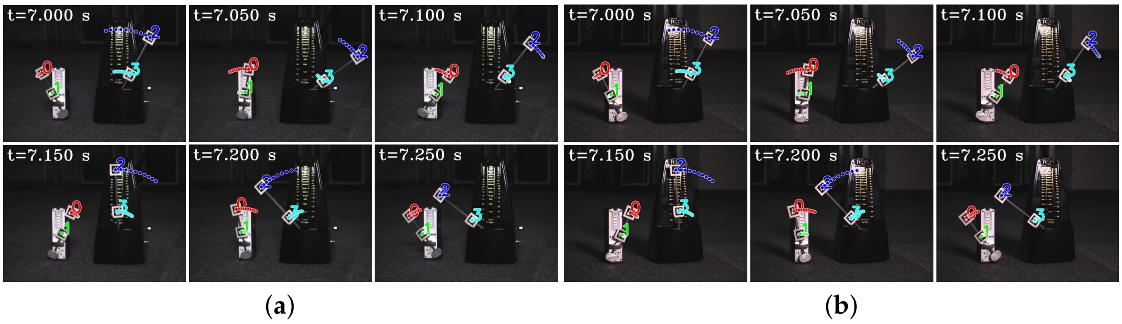


Figure 5. Input images and correspondence result in evaluation. (**a**) Left HFR stereoscopic video. (**b**) Right HFR stereoscopic video.

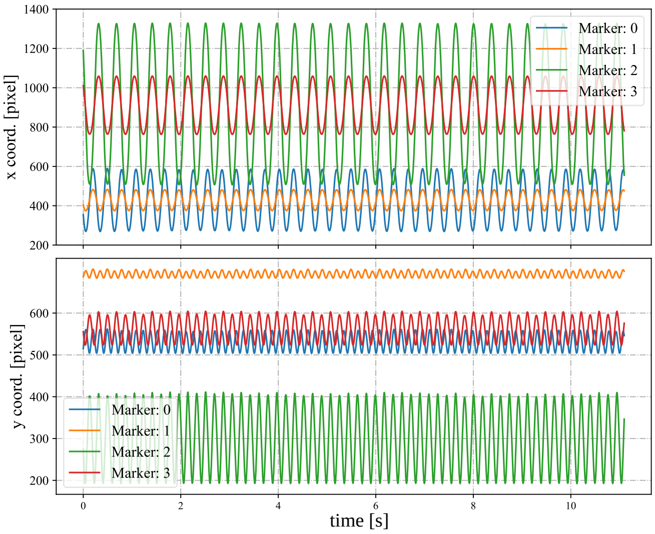
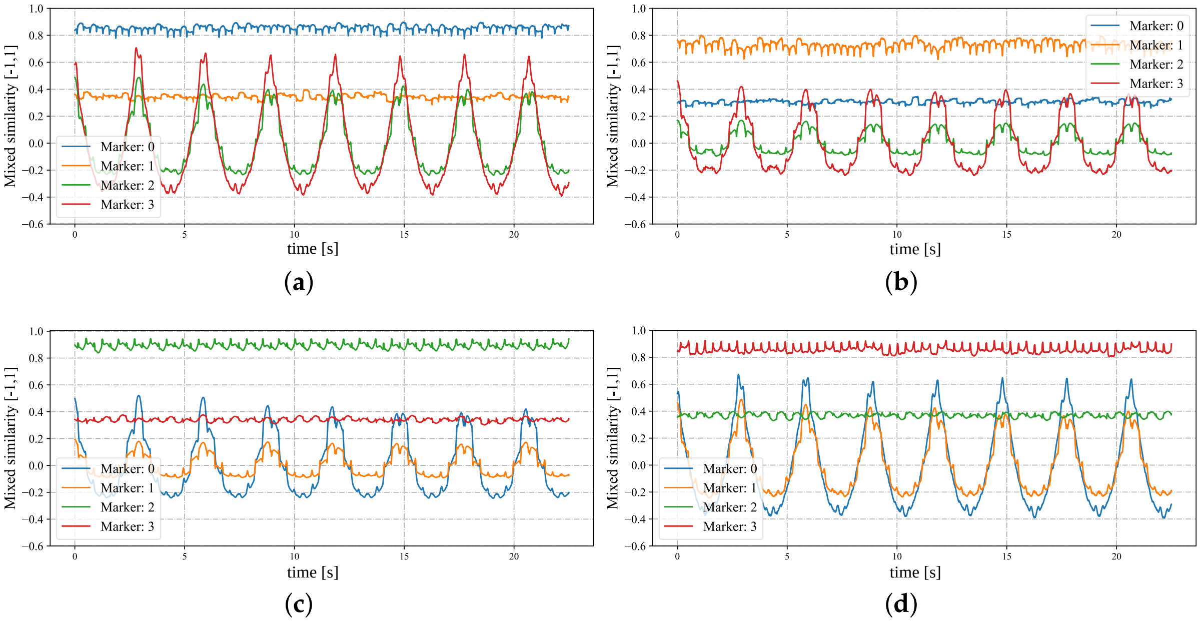


Figure 6. Image centroids of the markers in the stereo correspondence evaluation.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. Mixed similarities of different markers in the HFR stereoscopic video when the STVs length is 64. (**a**) Marker 0 in the left video. (**b**) Marker 1 in the left video. (**c**) Marker 2 in the left video. (**d**) Marker 3 in the left video.

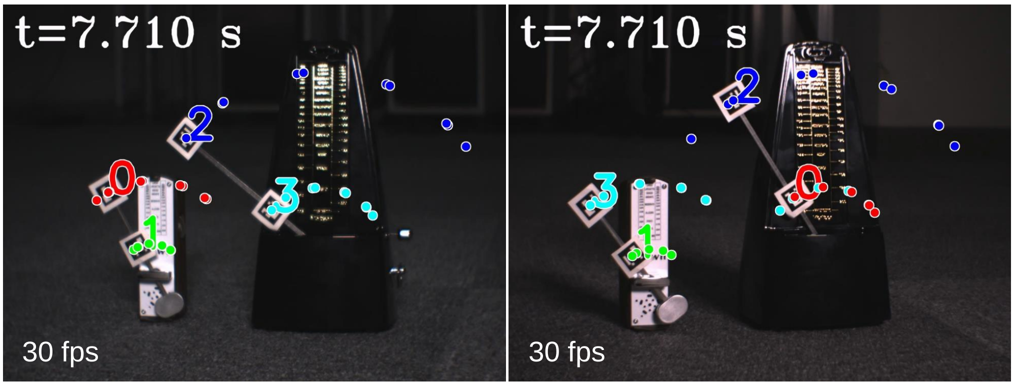


Figure 8. Correspondence results using a stereo camera at 30 fps (t = 7.710 s).

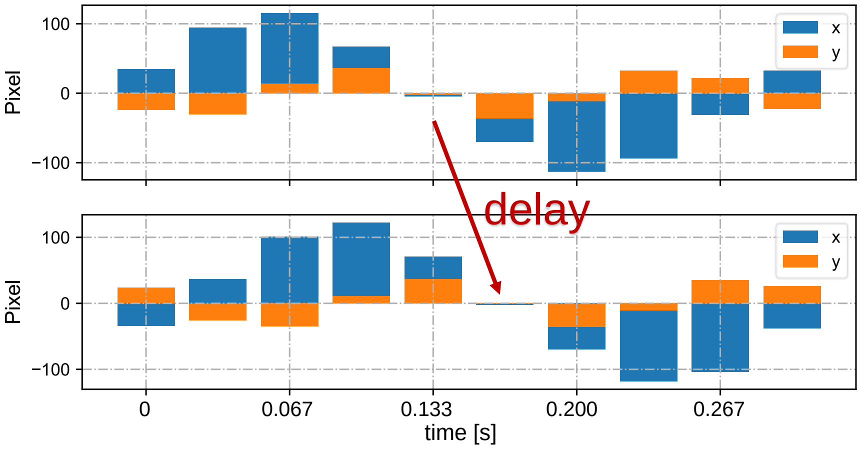


Figure 9. Short-term velocities of marker 0 in the stereo video in 0.3 s at 30 fps (t = 7.710 s).

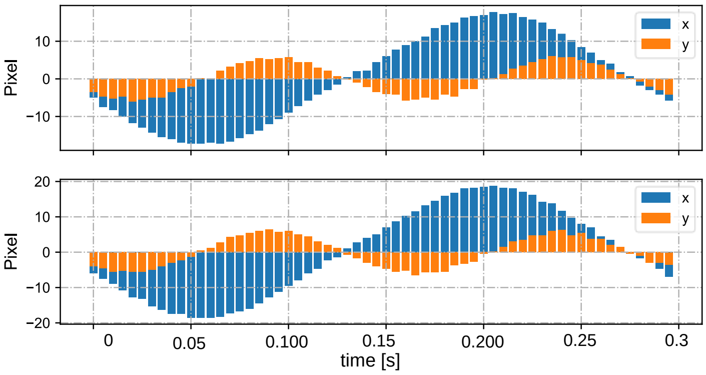


Figure 10. Short-term velocities of marker 0 in the stereo video in 0.3 s at 200 fps.

4.2. *Stereo Correspondence of Hands with Complex Movements*

We present the stereo correspondence results of hand movements during complex actions such as overlap and reappearance. The experimental setup is illustrated in Figure 11. Two individuals waved their hands approximately 8 m away from the HFR stereo cameras. Similar to the previous experiment, we captured a 200-fps HFR stereoscopic video using 12-mm fixed lens cameras. The hand movements in the video included mutual occlusion, static states, disappearance, and reappearance. There were four hands in the HFR stereoscopic video, represented by hand 2, hand 0, hand 1, and hand 3 from left to right. In the offline detection process, we utilized MediaPipe to detect the hands.

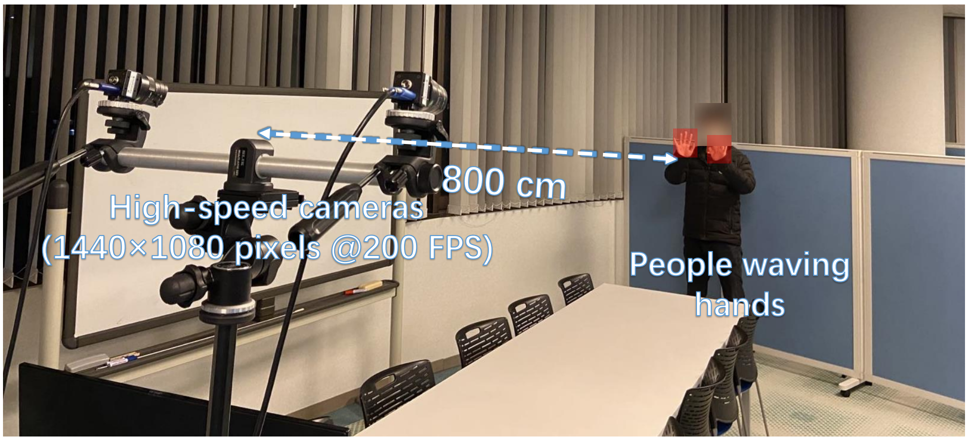


Figure 11. Experiment setup for hand stereo correspondence.

In Figure 12, we depict the input HFR stereo images with a resolution of 1440 × 1080 pixels and the correspondence results at intervals of 0.05 s for t = 12.84∼13.09 s. In the HFR stereoscopic video, similar hands are marked with similar colors from left to right. As shown in the graph, there is an overlap between hands 2 and 0, which belong to the person on the left. Hands 0 and 1, belonging to different people, also overlap. Our method correctly predicts the position of the hands and completes the hand correspondence even in the case of missing objects. The xy coordinate values of the image centroids of the hands in the left HFR stereoscopic video are shown in Figure 13. By analyzing the trajectories of the four hands, we can decompose the entire motion process into multiple actions. From 1.8 to 9.0 s, the hands belonging to the same person crossed each other and moved. From 23.0 to 28.0 s, hands 2 and 3 disappeared and reappeared. For the rest of the time, the four hands were stationary. The mixed similarities of different hands' STVs over time are shown in Figure 14. Figure 14a–d show the mixed similarities of STVs between hands 0, 1, 2, and 3 in the left HFR stereo image and those in the right HFR stereo images, respectively. Similar to the metronome correspondence, the motion features of a similar hand in the HFR stereoscopic video have a higher similarity. Since our features are motion-based, it can be seen from the figure that missing motion features introduced more uncertainty when the hand was stationary. Furthermore, we added appearance-based correspondence methods and calculated the accuracy of each method for hand correspondence every 0.25 s, as shown in Figure 15. The deep learning methods, ResNet and DeepCompare, achieved significantly better results throughout the process and were clearly superior to traditional methods. Our method maintained an accuracy of almost 100% during hand movement. The accuracy rate was lower than that of the appearance-based methods only when the hand was stationary. In calibrated stereo cameras, when similar objects are found in the stereo camera, their spatial positions can be calculated. In Figure 16, we plotted the 3D trajectories of hands 0, 1, 2, and 3 over 7 to 8 s. From the image, we can see that the four hands moved up and down at a distance of approximately 8 m from the camera. The acquisition of spatial information helped us to better analyze the movement of objects.

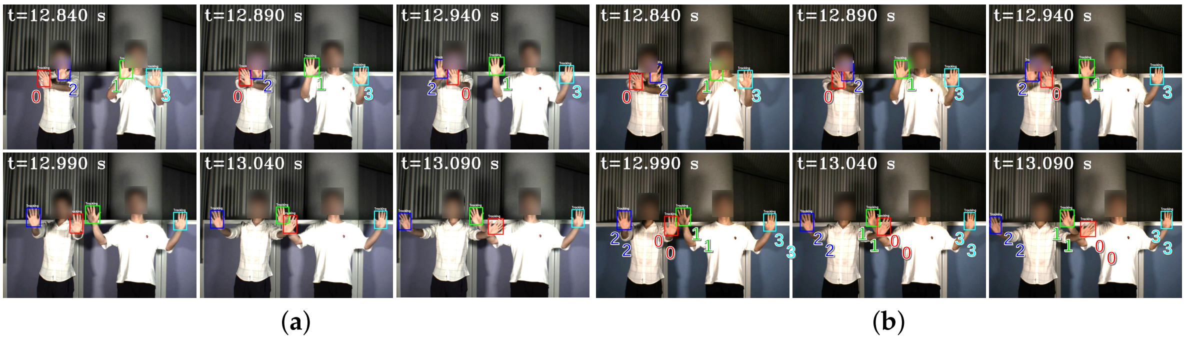


Figure 12. Input images and hand correspondence result. (**a**) Left HFR stereo images. (**b**) Right HFR stereo images.

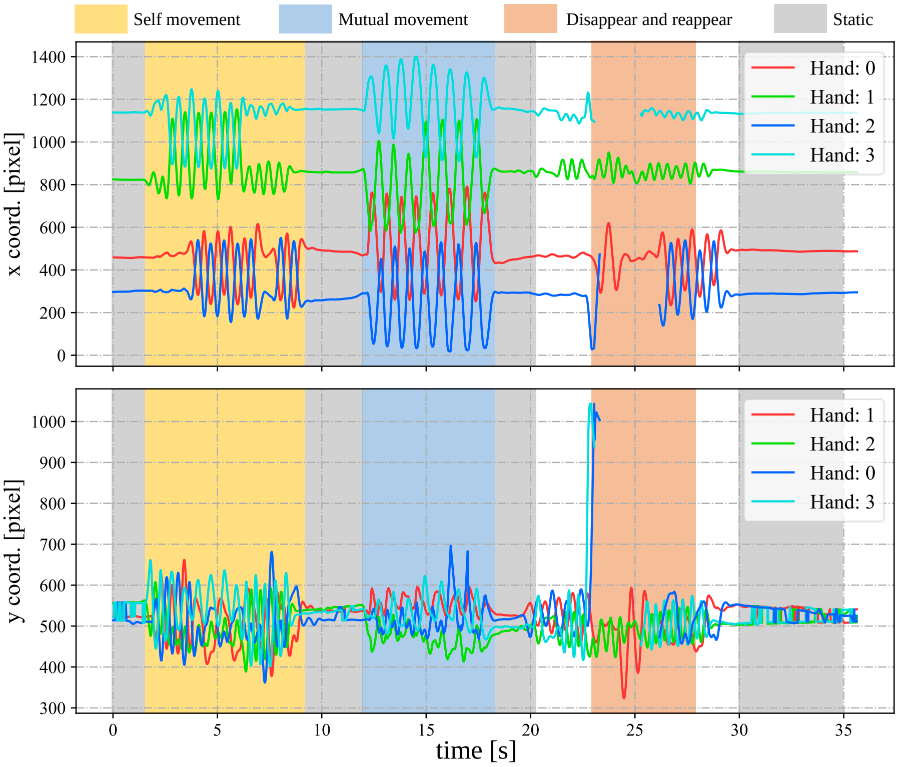
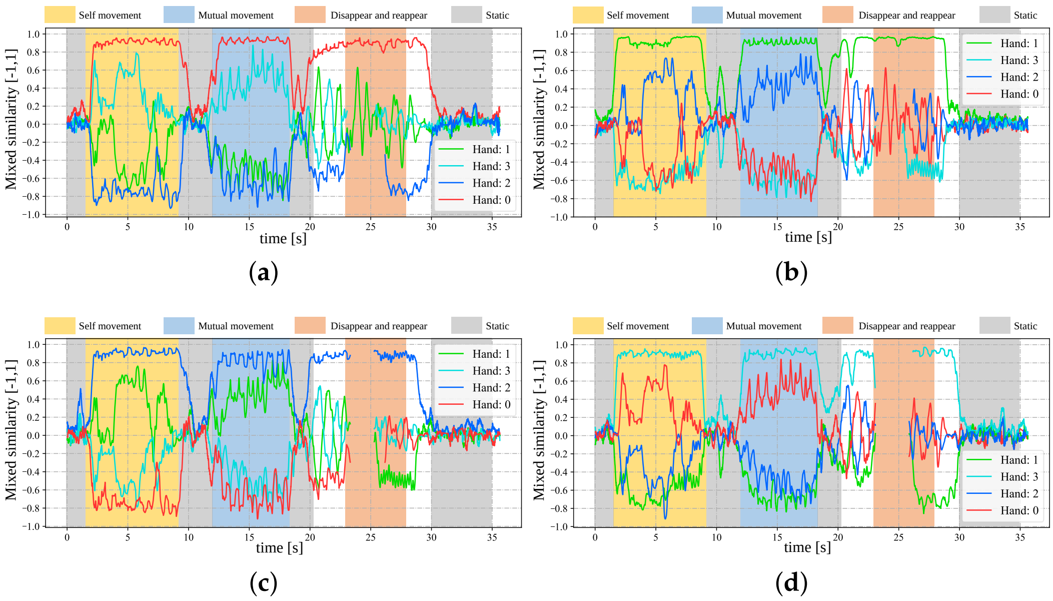


Figure 13. Image centroids of the hands in the left HFR stereoscopic video.



(**a**) (**b**)

(**c**) (**d**)

Figure 14. Mixed similarities between different hands in the HFR stereoscopic video when the STVs length is 64. (**a**) Hand 0 in the left video. (**b**) Hand 1 in the left video. (**c**) Hand 2 in the left video. (**d**) Hand 3 in the left video.

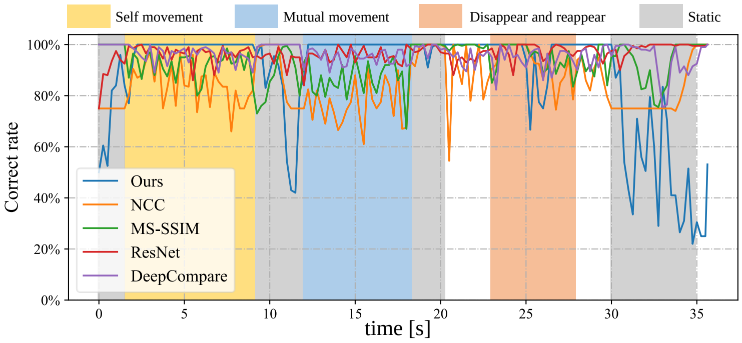


Figure 15. Correct rate of different stereo correspondence methods updated every 0.25 s.

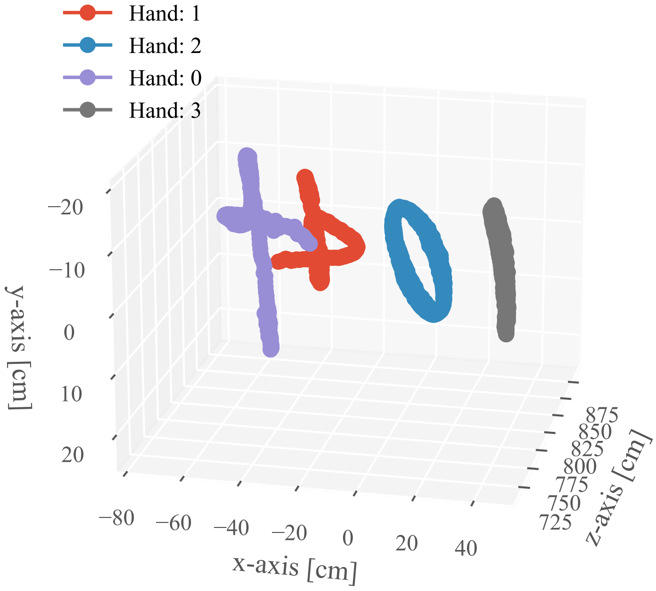


Figure 16. 3D trajectory of each hand with 7∼8 s.

4.3. Stereo Correspondence in the Meeting Room

Finally, we present the experimental results for stereo correspondences when the stereo cameras operate at 200 fps in a meeting room. To obtain a larger field of view, the stereo cameras are equipped with 6-mm lenses. The experimental setup is illustrated in Figure 17. In the meeting room, several students were more than 2 m away from the stereo cameras. Due to factors such as privacy and occlusion, it was difficult to detect and identify different students by their faces. Obtaining the spatial position using stereo correspondence of the hands is a feasible solution to identify different students.

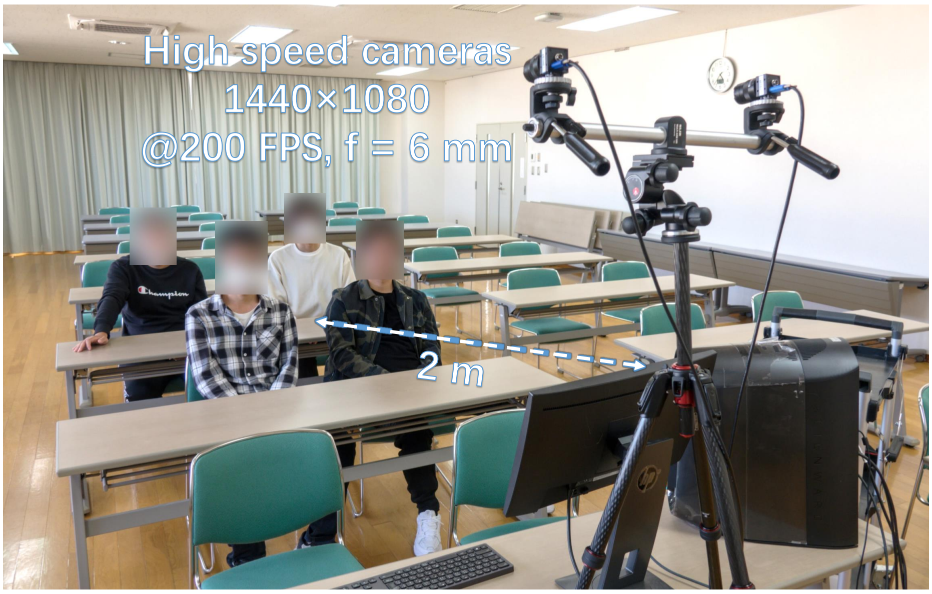
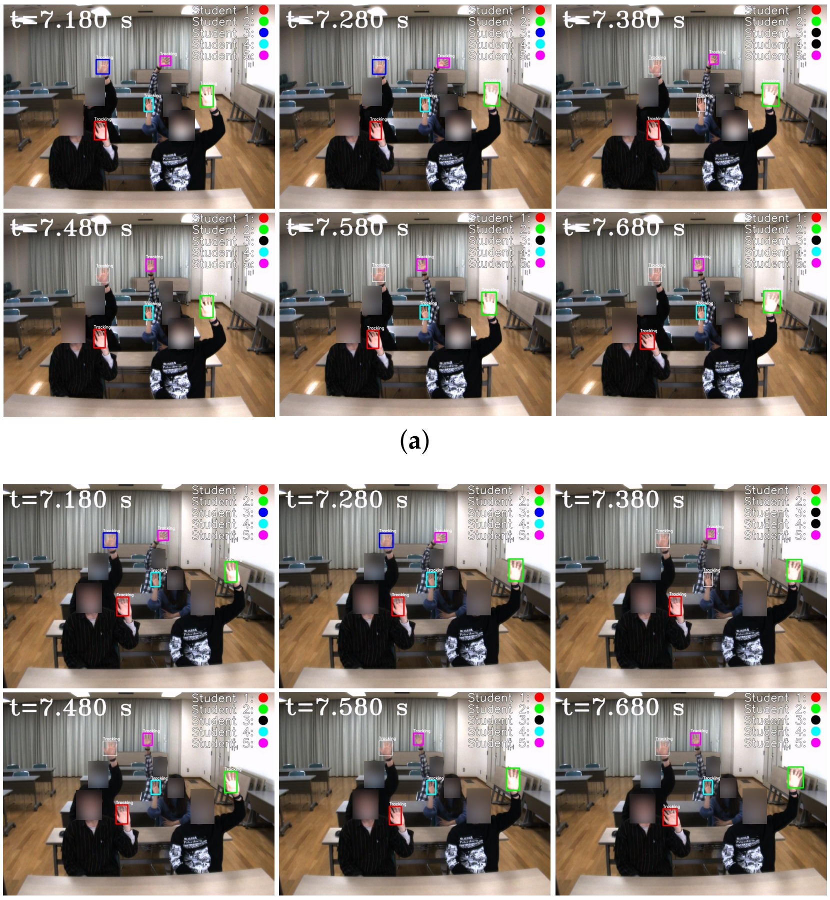


Figure 17. Experimental environment for stereo correspondence in the meeting room.

In Figure 18, we depict the input HFR stereo images of 1440 × 1080 pixels with correspondence results at intervals of 0.1 s for t = 7.180∼7.680 s. We numbered the students from 1 to 5, from the nearest to the farthest. Similar hands in the stereo HFR video were marked with similar colors, as in the previous experiment. When the students raised their hands, we performed stereo correspondence using hand movements. Furthermore, we calculated the 3D positions of the different hands. In this experiment, we knew the seating distribution of each student in advance, and we could identify who raised their hand through the position of the hand. In Figure 18, we marked the hand-raising action of classmates in the upper right corner. When the hands were raised, circles belonging to different students were filled with different colors; otherwise, they were filled with black. The xy coordinate values of the hand images in the left HFR stereoscopic video are shown in Figure 19 at t = 0–16 s. Simultaneously, Figure 20 shows the time variation of the mixed similarity between similar hands at t = 0–16 s. From the graph, the hands of students 1 to 5 appeared individually in the HFR stereoscopic video. The students' hands moved at 0–7 and 10.5–16 s. During motion, the same hand in the HFR stereoscopic video had a high mixed similarity of approximately 0.8. We stopped the hand from moving at 6.8–10.2 s. As seen in Figure 20, the mixed similarities of the same hand dropped rapidly, greatly reducing the accuracy of the correspondence. Figure 21 shows the Gantt chart of the detected students' hands raised over time. When the hand stopped moving, we could not accurately complete the correspondence. Our algorithm is currently limited regarding stereo correspondence in the static state. These results show that our method can accurately match objects in a stereoscopic video in moving scenes and use spatial information to complete certain applications.



(**a**)

(**b**)

Figure 18. Input images and stereo correspondence result. (**a**) Left HFR stereo images. (**b**) Right HFR stereo images.

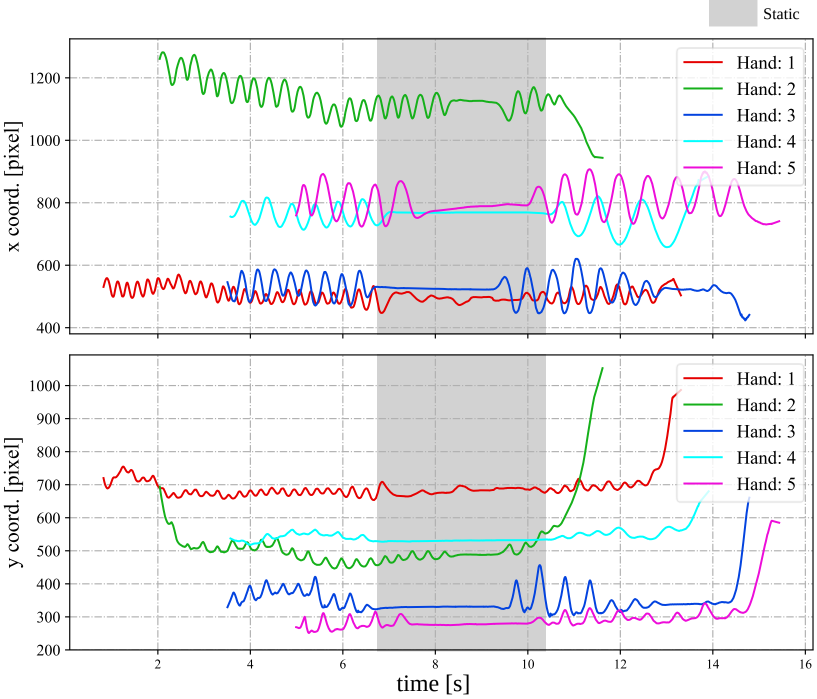


Figure 19. Image centroids of hands in the left HFR stereoscopic video.

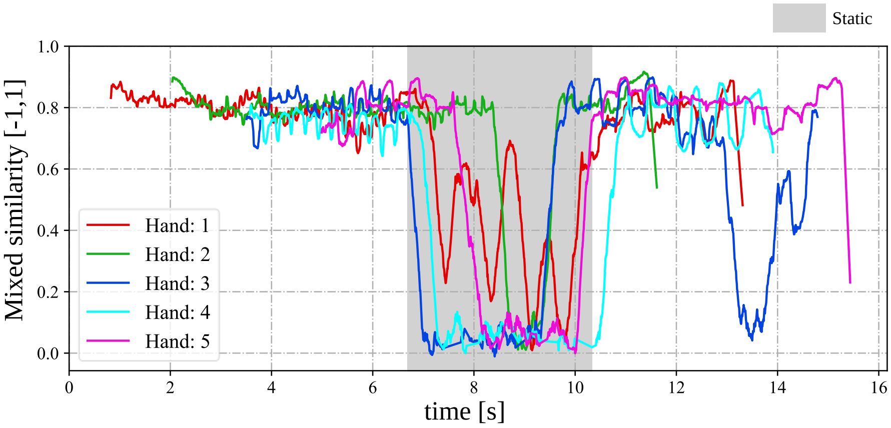


Figure 20. Mixed similarities between a similar hand in the HFR stereoscopic video when the STVs length is 64.

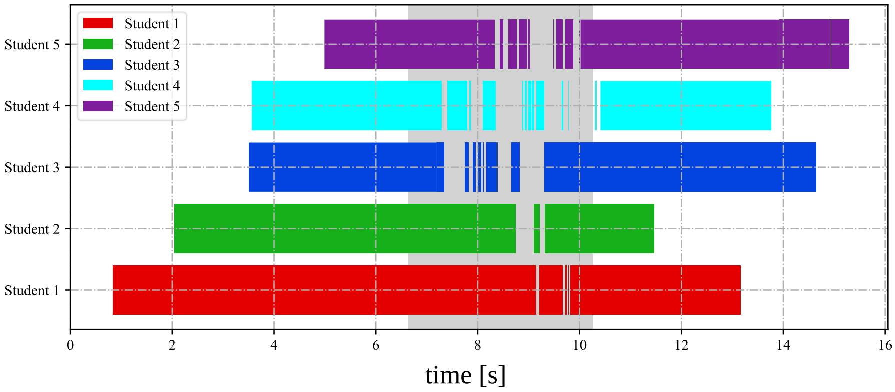


Figure 21. Statistical analysis of raised hands.

5. Conclusions

In this study, we addressed the problem of stereo correspondence of objects with similar appearances. Traditional appearance-based algorithms do not provide effective performance, so we proposed a method that uses highly synchronized motion information to overcome this limitation. Our approach involves using high-synchronous short-term velocities acquired by high-speed vision systems as features for stereo correspondence of moving objects. We demonstrated the effectiveness of our method through experiments on (1) the correspondence of markers for regular motion on a metronome and (2) motion tracking and correspondence of multiple hands in indoor scenes. These experiments confirmed the potential of high-speed vision technology to improve the stereo correspondence of objects with similar appearances. However, our current method cannot provide accurate results when objects are static.

In the future, we aim to address the following issues to further improve our method: (1) We will add appearance-related factors to our algorithm to achieve higher accuracies in stereo correspondences when objects are static. (2) Currently, our algorithm is run offline, but we plan to implement it using real-time HFR stereo camera systems in the future.

Author Contributions: All authors contributed to designing this study and writing the manuscript. I.I. contributed to the concept of stereo correspondence based on motion information. S.H. and K.S. designed and assembled a high-speed stereo camera system. Q.L. developed a high-speed hybrid tracking of multiple objects and a stereo matching method based on motion information. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [China Scholarship Council] grant number [202006540007].

Institutional Review Board Statement: I think this study did not require ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. And I think this study did not involve humans.

Data Availability Statement: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hamid, M.S.; Abd Manap, N.; Hamzah, R.A.; Kadmin, A.F. Stereo matching algorithm based on deep learning: A survey. *J. King Saud-Univ.-Comput. Inf. Sci.* **2022**, *34*, 1663–1673. [CrossRef]
2. Oroko, J.A.; Nyakoe, G. Obstacle avoidance and path planning schemes for autonomous navigation of a mobile robot: A review. In Proceedings of the Sustainable Research and Innovation Conference, Nairobi, Kenya, 5–7 October 2022; pp. 314–318.
3. Liu, C.; Xing, C.; Hu, Q.; Wang, S.; Zhao, S.; Gao, M. Stereoscopic hyperspectral remote sensing of the atmospheric environment: Innovation and prospects. *Earth-Sci. Rev.* **2022**, *226*, 103958. [CrossRef]
4. Schlinkmann, N.; Khakhar, R.; Picht, T.; Piper, S.K.; Fekonja, L.S.; Vajkoczy, P.; Acker, G. Does stereoscopic imaging improve the memorization of medical imaging by neurosurgeons? Experience of a single institution. *Neurosurg. Rev.* **2022**, *45*, 1371–1381. [CrossRef] [PubMed]
5. Macario Barros, A.; Michel, M.; Moline, Y.; Corre, G.; Carrel, F. A comprehensive survey of visual slam algorithms. *Robotics* **2022**, *11*, 24. [CrossRef]
6. Shabanian, H.; Balasubramanian, M. A novel factor graph-based optimization technique for stereo correspondence estimation. *Sci. Rep.* **2022**, *12*, 15613. [CrossRef]
7. Hamzah, R.A.; Azali, M.N.Z.; Noh, Z.M.; Zahari, M.; Herman, A.I. Development of depth map from stereo images using sum of absolute differences and edge filters. *Indones. J. Electr. Eng. Comput. Sci.* **2022**, *25*, 875–883. [CrossRef]
8. Chang, Q.; Zha, A.; Wang, W.; Liu, X.; Onishi, M.; Lei, L.; Er, M.J.; Maruyama, T. Efficient stereo matching on embedded GPUs with zero-means cross correlation. *J. Syst. Archit.* **2022**, *123*, 102366. [CrossRef]
9. Wang, F.; Ding, L. Object recognition and localization based on binocular stereo vision. In Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics, Nanjing, China, 24–26 June 2022; pp. 196–201.
10. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.; Tardós, J.D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
11. Zbontar, J.; LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1592–1599.
12. Laga, H.; Jospin, L.V.; Boussaid, F.; Bennamoun, M. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1738–1764. [CrossRef]
13. Kaya, M.; Bilge, H.Ş. Deep metric learning: A survey. *Symmetry* **2019**, *11*, 1066. [CrossRef]
14. Köhl, P.; Specker, A.; Schumann, A.; Beyerer, J. The MTA Dataset for Multi Target Multi Camera Pedestrian Tracking by Weighted Distance Aggregation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 4489–4498. [CrossRef]
15. Li, P.; Zhang, J.; Zhu, Z.; Li, Y.; Jiang, L.; Huang, G. State-aware re-identification feature for multi-target multi-camera tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
16. He, Y.; Wei, X.; Hong, X.; Shi, W.; Gong, Y. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Trans. Image Process.* **2020**, *29*, 5191–5205. [CrossRef]
17. Magdy, N.; Sakr, M.A.; Mostafa, T.; El-Bahnasy, K. *Review on Trajectory Similarity Measures*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2016; pp. 613–619. [CrossRef]
18. Li, Q.; Chen, M.; Gu, Q.; Ishii, I. A Flexible Calibration Algorithm for High-speed Bionic Vision System based on Galvanometer. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 4222–4227.
19. Gu, Q.Y.; Ishii, I. Review of some advances and applications in real-time high-speed vision: Our views and experiences. *Int. J. Autom. Comput.* **2016**, *13*, 305–318. [CrossRef]
20. Costa, L.d.F. Comparing cross correlation-based similarities. *arXiv* **2021**, arXiv:2111.08513.
21. Sara, U.; Akter, M.; Uddin, M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *J. Comput. Commun.* **2019**, *7*, 8–18. [CrossRef]
22. Zhao, S.; Wang, Y.; Yang, Z.; Cai, D. Region mutual information loss for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
23. Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; Zhu, Q. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote. Sens.* **2019**, *57*, 9059–9070. [CrossRef]
24. Zermi, N.; Khaldi, A.; Kafi, R.; Kahlessenane, F.; Euschi, S. A DWT-SVD based robust digital watermarking for medical image security. *Forensic Sci. Int.* **2021**, *320*, 110691. [CrossRef]

25. Yang, L.; Su, H.; Zhong, C.; Meng, Z.; Luo, H.; Li, X.; Tang, Y.Y.; Lu, Y. Hyperspectral image classification using wavelet transform-based smooth ordering. *Int. J. Wavelets, Multiresolut. Inf. Process.* **2019**, *17*, 1950050. [CrossRef]
26. Pautrat, R.; Larsson, V.; Oswald, M.R.; Pollefeys, M. Online invariance selection for local feature descriptors. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Lecture Notes in Computer Science. Springer: Cham, Switzerland; Volume 12347, pp. 707–724.
27. Gupta, S.; Thakur, K.; Kumar, M. 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. *Vis. Comput.* **2021**, *37*, 447–456. [CrossRef]
28. Pang, Y.; Li, A. An improved ORB feature point image matching method based on PSO. In Proceedings of the Tenth International Conference on Graphics and Image Processing (ICGIP 2018), Chengdu, China, 12–14 December 2018; SPIE: Bellingham, WA, USA, 2019; Volume 11069, pp. 224–232.
29. Chengtao, C.; Mengqun, L. Tire pattern similarity detection based on template matching and LBP. In Proceedings of the 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, 18–20 October 2019; pp. 419–423.
30. Venkataramanan, A.K.; Wu, C.; Bovik, A.C.; Katsavounidis, I.; Shahid, Z. A hitchhiker's guide to structural similarity. *IEEE Access* **2021**, *9*, 28872–28896. [CrossRef]
31. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
32. Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; Roy, K. Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **2019**, *13*, 95. [CrossRef] [PubMed]
33. Hernandez-Diaz, K.; Alonso-Fernandez, F.; Bigun, J. Cross Spectral Periocular Matching using ResNet Features. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019; pp. 1–7. [CrossRef]
34. Agarwal, R.; Verma, O.P. An efficient copy move forgery detection using deep learning feature extraction and matching algorithm. *Multimed. Tools Appl.* **2020**, *79*, 7355–7376. [CrossRef]
35. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. Matchnet: Unifying feature and metric learning for patch-based matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3279–3286.
36. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4353–4361. [CrossRef]
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
38. Yang, T.Y.; Hsu, J.H.; Lin, Y.Y.; Chuang, Y.Y. DeepCD: Learning Deep Complementary Descriptors for Patch Representations. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3334–3342. [CrossRef]
39. Hsu, H.M.; Cai, J.; Wang, Y.; Hwang, J.N.; Kim, K.J. Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model. *IEEE Trans. Image Process.* **2021**, *30*, 5198–5210. [CrossRef]
40. Gou, M.; Karanam, S.; Liu, W.; Camps, O.; Radke, R.J. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 10–19.
41. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient online pose tracking. *arXiv* **2018**, arXiv:1802.00977.
42. Su, H.; Liu, S.; Zheng, B.; Zhou, X.; Zheng, K. A survey of trajectory distance measures and performance evaluation. *VLDB J.* **2020**, *29*, 3–32. [CrossRef]
43. Zhao, L.; Shi, G. A novel similarity measure for clustering vessel trajectories based on dynamic time warping. *J. Navig.* **2019**, *72*, 290–306. [CrossRef]
44. Maergner, P.; Pondenkandath, V.; Alberti, M.; Liwicki, M.; Riesen, K.; Ingold, R.; Fischer, A. Combining graph edit distance and triplet networks for offline signature verification. *Pattern Recognit. Lett.* **2019**, *125*, 527–533. [CrossRef]
45. Rubinstein, A.; Song, Z. Reducing approximate longest common subsequence to approximate edit distance. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, PA, United States, 5–8 January 2020; pp. 1591–1600.
46. Ying, L.; Li, Z.; Xiang-mo, Z.; Ke, C. Effectiveness of trajectory similarity measures based on truck GPS data. *China J. Highw. Transp.* **2020**, *33*, 146.
47. Gong, L.; Chen, B.; Xu, W.; Liu, C.; Li, X.; Zhao, Z.; Zhao, L. Motion similarity evaluation between human and a tri-co robot during real-time imitation with a trajectory dynamic time warping model. *Sensors* **2022**, *22*, 1968. [CrossRef] [PubMed]
48. Zhu, C.; Yang, J.; Shao, Z.; Liu, C. Vision based hand gesture recognition using 3D shape context. *IEEE/CAA J. Autom. Sin.* **2019**, *8*, 1600–1613. [CrossRef]
49. Patel, C.I.; Labana, D.; Pandya, S.; Modi, K.; Ghayvat, H.; Awais, M. Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences. *Sensors* **2020**, *20*, 7299. [CrossRef] [PubMed]
50. Zhao, X.; Rao, Y.; Cai, J.; Ma, W. Abnormal trajectory detection based on a sparse subgraph. *IEEE Access* **2020**, *8*, 29987–30000. [CrossRef]

51. Liang, X.; Zhang, H.B.; Zhang, Y.X.; Huang, J.L. JTCR: Joint Trajectory Character Recognition for human action recognition. In Proceedings of the 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 3–6 October 2019; pp. 350–353. [CrossRef]
52. Cao, J.; Liang, M.; Li, Y.; Chen, J.; Li, H.; Liu, R.W.; Liu, J. PCA-based hierarchical clustering of AIS trajectories with automatic extraction of clusters. In Proceedings of the 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), Shanghai, China, 9–12 March 2018; pp. 448–452.
53. Xiao, Z.; Wang, Y.; Fu, K.; Wu, F. Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS Int. J. -Geo-Inf.* **2017**, *6*, 57. [CrossRef]
54. Bagheri, M.A.; Gao, Q.; Escalera, S. Support vector machines with time series distance kernels for action classification. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–7.
55. Yao, D.; Zhang, C.; Zhu, Z.; Huang, J.; Bi, J. Trajectory clustering via deep representation learning. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 3880–3887.
56. Zhang, R.; Xie, P.; Jiang, H.; Xiao, Z.; Wang, C.; Liu, L. Clustering noisy trajectories via robust deep attention auto-encoders. In Proceedings of the 2019 20th IEEE International Conference on Mobile Data Management (MDM), Hong Kong, China, 10–13 June 2019; pp. 63–71.
57. Liang, M.; Liu, R.W.; Li, S.; Xiao, Z.; Liu, X.; Lu, F. An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation. *Ocean. Eng.* **2021**, *225*, 108803. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.