ITC 2/53
Information Technology and Control
Vol. 53 / No. 2 / 2024
pp. 542-553
DOI 10.5755/j01.itc.53.2.36044

Optimization of Human Posture Recognition Based on Multi-view Skeleton Data Fusion

Received 2024/01/12 | Accepted after revision 2024/04/29

HOW TO CITE: Xu, Y., Wei, S., Yin, J. (2024). Optimization of Human Posture Recognition Based on Multi-view Skeleton Data Fusion. *Information Technology and Control*, 53(2), 542-553. https://doi.org/10.5755/j01.itc.53.2.36044

Optimization of Human Posture Recognition Based on Multi-view Skeleton Data Fusion

Yahong Xu, Shoulin Wei, Jibin Yin

Faculty of Information Engineering and Automation; Kunming University of Science and Technology; 665000, Yunnan, China

Corresponding author: Jibin Yin, 41868028@qq.com

This research introduces a novel method for fusing multi-view skeleton data to address the limitations encountered by a single vision sensor in capturing motion data, such as skeletal jitter, self-pose occlusion, and the reduced accuracy of three-dimensional coordinate data for human skeletal joints due to environmental object occlusion. Our approach employs two Kinect vision sensors concurrently to capture motion data from distinct viewpoints extract skeletal data and subsequently harmonize the two sets of skeleton data into a unified world coordinate system through coordinate conversion. To optimize the fusion process, we assess the contribution of each joint based on human posture orientation and data smoothness, enabling us to fine-tune the weight ratio during data fusion and ultimately produce a dependable representation of human posture. We validate our methodology using the FMS public dataset for data fusion and model training. Experimental findings demonstrate a substantial enhancement in the smoothness of the skeleton data, leading to enhanced data accuracy and an effective improvement in human posture recognition following the application of this data fusion method.

KEYWORDS: human posture recognition, vision sensor, multi-view, data fusion, coordinate transformation.

1. Introduction

Human posture recognition involves the extraction of human skeleton data from video sequences. Computer algorithms utilize this skeleton data to identify specific human posture classifications. This research direction is pivotal within computer vision and finds applications across diverse fields such as human-computer interaction, medical sports rehabilitation, pedestrian recognition, and virtual reality. In

these scenarios, vision sensors capture video image data containing human movements, which are subsequently processed and integrated to derive human skeleton data for analysis.

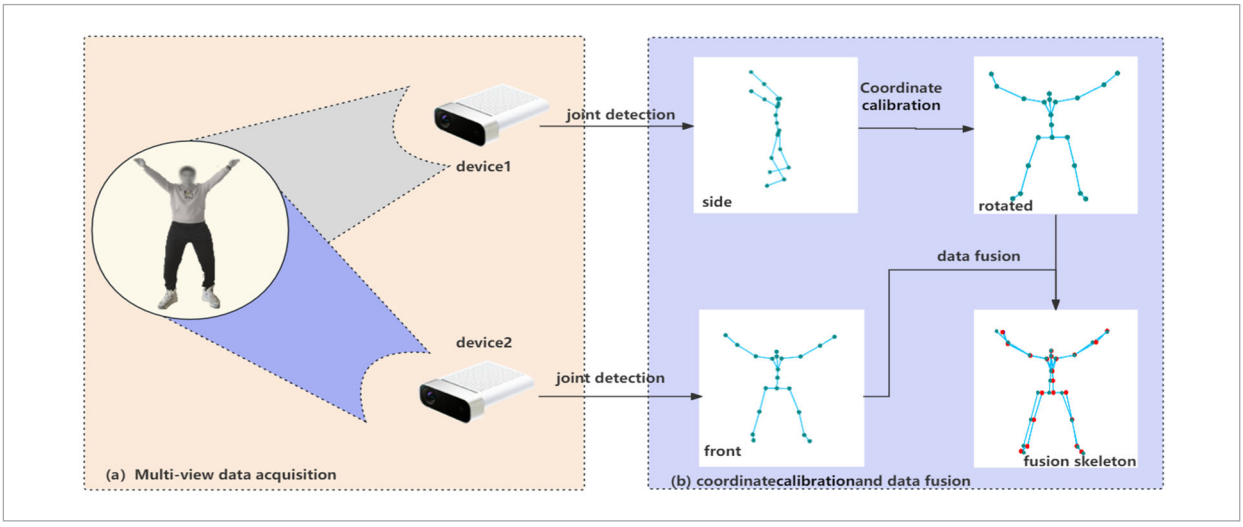
Many researchers have introduced effective methods for human posture recognition. These approaches predominantly rely on video images captured from a single-view camera to discern human movements and trajectories, constituting single-view human posture recognition. However, the accuracy of human pose recognition from a solitary camera perspective is limited due to inherent constraints in capturing comprehensive human joint data, resulting in diminished accuracy. Common challenges encompass self- occlusion of joint movements, jitter in joints due to random noise, data loss, and data oscillation, among others. These issues compromise data quality and accuracy, consequently diminishing the precision of human posture recognition. To address these challenges, this paper proposes a data fusion method employing multi-view Kinects to enhance the reliability of generated human poses.

This research presents an innovative approach employing two visual sensors to simultaneously capture human body motion data from distinct angles (Figure 1. a). This optimization of the existing single-view human posture recognition method leverages multi-view data fusion techniques to harness the organic complementarity inherent in data collected from various perspectives. After aligning the data from different devices into a common coordinate system, fusion weights are determined based on the angle of the human body relative to the cameras and the smoothness of the data, resulting in the generation of a more trustworthy representation of human posture (Figure 1. b). This enhances the reliability of the fused data and addresses prevalent issues in current algorithms, notably the lack of interactivity in the data fusion process and the overly rigid selection of high-credibility data. The improved data fusion algorithm presented in this article yields more dependable three-dimensional coordinate data for human skeletal joint points, substantially enhancing the accuracy of human posture recognition.

To validate the proposed method, we applied it to the publicly available FMS dataset for fusing the skeleton data collected by two vision sensors and integrated it into the training and testing phases of the human posture recognition algorithm. The principal contributions of this study are twofold: (1) We introduce a data fusion method that combines skeletal data from multiple devices to produce reliable human posture data. This method enhances the accuracy of human posture recognition and bolsters the robustness of the entire data collection system, effectively addressing potential issues such as data loss and data jitter; (2) We apply this fusion method to the FMS dataset, utilizing the merged data with multiple advanced human pose recognition algorithms to further substantiate its effectiveness.

Figure 1

(a) Multi-view data collection (b) Coordinate calibration and data fusion



2. Related Work

2.1. Skeleton-Based Human Posture Recognition

In light of the continuous advancements in deep learning techniques for data analysis and processing, the adoption of deep learning for skeleton-based human posture recognition has emerged as a dominant trend. The Convolutional Neural Network (CNN) [11] approach, while effectively representing the skeleton as a pseudo image and capturing local correlations, is not ideally suited for sequential tasks. Li et al. [13] addressed this limitation by dividing the human skeleton into five segments, transforming them into two-dimensional action images, and subsequently applying image classification techniques. Simonyan and Zisserman [22] first proposed a dual stream framework to capture spatio-temporal information in video frame sequences, this framework consists of two separately running CNNs, one extracting spatial information from a single RGB image and the other extracting motion information from video optical flow sequences, the two sets of features are fused in the final classification layer. Liu et al. [15] added a spatio-temporal interactive learning block in the middle of the network to complete feature fusion in the early stages. Wang et al. [24] changed the spatial feature extraction and recognition of RGB images from single frames to multiple frames, improving the network's ability to describe spatial features. In contrast, the Recurrent Neural Network (RNN) [29] method constructs the skeleton as a sequence of joint coordinate vectors. Wu and Shao [26] introduced a dynamic framework, pioneering the extraction of high-level bone joint features. Meanwhile, Liu et al. [14] maintained a single-stream approach, treating the human body as a tree structure and feeding the human body joint nodes into the Long Short-Term Memory network (LSTM) [8] in a depth-first order traversal manner. This allowed them to capture temporal relationships by stacking LSTM modules. Han and Shan [7] increase feature extraction channels by optimizing feature extraction methods or improving feature extraction efficiency to improve the recognition performance of the model.

However, representing the human skeleton as a sequence of vectors or a two-dimensional mesh falls short of capturing the intricate dependencies among interconnected joints. The structure of human joint points and the skeleton naturally forms a graph-like structure, where the Graph Convolutional Network (GCN) [10] emerges as a more adept tool for capturing the intricate relationships among different joint points during human motion. ST-GCN [28] pioneers the application of graph convolution networks to the realm of human action recognition. This groundbreaking approach leverages both spatial edges naturally connecting joint points and the temporal edges stemming from the same joint points across continuous time. It constructs a spatio-temporal graph convolution, significantly enhancing the model's ability to grasp the temporal and spatial relationships. 2S-AGCN [21] introduces a dual-level graph representation, featuring a global graph capturing common patterns across all data and individual graphs tailored to unique data instances. This innovative approach overcomes the constraints associated with predefined and unmodifiable skeleton graphs in ST-GCN, introducing a new paradigm known as Adaptive Graph Convolutional Networks. In the same vein, MS-G3D [16] puts forth a multi-scale adjacency matrix and a unified spatio-temporal model. The multi-scale convolution effectively mitigates issues related to biased weighting, while the unified spatio-temporal model introduces cross-spatio-temporal jump connections. Furthermore, it incorporates a time window mechanism in the temporal dimension to enhance the flow of spatio-temporal information. In a parallel development, Shift-GCN [4] introduces the concept of shift convolution [25] to the realm of human motion recognition. By combining the convolution operator with spatial shift operations, this approach simultaneously integrates information from both spatial and channel domains, strategically offsetting channels in the temporal and spatial dimensions. This novel approach more accurately represents human joint constraints and temporal information. Lee et al. [12] sorted all nodes and generated a new adjacency matrix, improving the algorithm's robustness against certain key point exchanges, displacements, or losses.

2.2. Multi-view Data Fusion

Currently, research on the fusion of multi-view data is relatively limited. Tong et al. [23] proposed a method utilizing multiple Kinects to scan distinct areas of the human body and fuse them into a comprehensive

three-dimensional human body model. Geerse et al. [5] positioned four Kinect devices along one side of a corridor to extract gait parameters and analyze human gait. Müller et al. [17] strategically deployed six Kinects on both sides of a corridor, fusing skeletal data from both sides to enhance three-dimensional reconstruction for gait assessment, thereby improving the accuracy of skeletal data. Guo et al. [6] introduced a dual Kinect data fusion technology based on posture angles. They devised a corresponding data selection mechanism that leveraged the angle relationship between different body posture directions and sensor positions, offering a straightforward and rapid data fusion solution. This approach effectively addressed data loss arising from the self-occlusion of the human skeleton. Jiang et al. [9] proposed a data fusion method based on joint angles. They performed data fusion by calculating the weighted average of data from two Kinect devices placed orthogonally. However, this approach did not account for compensating for data loss in cases of missing data. Peng et al. [18] introduced a fusion algorithm that used redundant data to compensate for occlusion. Nonetheless, there was a substantial disparity between redundant data and original data, leading to poor correlation with previous and subsequent frame data, resulting in reduced action recognition accuracy. Chen et al. [2] adopted a data screening approach based on the physiological constraints of human joints. They exclusively used data from the primary device when it successfully tracked the data and only resorted to data fusion with another device when the primary device failed to track the data. However, the effectiveness of this data fusion approach requires further improvement.

3. Data Fusion

The visual sensor is proficient in capturing human motion data, and then bone extraction techniques can be used to extract skeletal data, but when utilized as a single sensor for motion capture, it encounters self-occlusion issues that compromise the accuracy of the acquired three-dimensional coordinate data for human skeletal joint points. This research introduces a novel data fusion model predicated on human body posture orientation and data smoothness. It fuses and processes two sets of data acquired from different perspectives. Initially, a coordinate transformation is employed to align the joint data captured at varying angles, into a shared world coordinate system, establishing a uniform data platform. Subsequently, the choice of data fusion method depends on the successful data capture by device. If only one camera effectively captures data, that data is utilized directly. However, when both cameras successfully record data, data fusion is implemented. Lastly, the model integrates joint posture orientation and data smoothness to assess the data contributions of the two devices, optimize the data fusion weight coefficients, and generate a reliable representation of human posture. This model significantly enhances the accuracy of motion-captured human posture data and resolves the issue of data loss due to self-occlusion in single-view data capture.

3.1. Coordinate Calibration

In practice, when employing a single vision sensor device, it captures human skeleton joint data based on its own camera coordinate system. Therefore, when using two or more vision sensors to collect data, ensuring accurate data fusion necessitates the harmonization of the data acquired by these devices into a common world coordinate system-referred to as coordinate calibration. The primary objective of coordinate calibration is to establish a shared reference coordinate system, guaranteeing uniformity in the coordinate standards across different datasets. This standardization enables consistent processing and analysis of skeletal joint data gathered from diverse devices, serving as a fundamental prerequisite for effective data fusion. It is only when the data exists within the same coordinate system that meaningful data comparisons and fusion can occur, ultimately enabling precise motion capture.

During the process of converting coordinate systems, it is crucial to account for variations in scale, rotation, and translation. In our study, we specifically focus on the transformation between the reference coordinate systems of two Kinects. Given that the parameters and characteristics of the two Kinect sensors are identical, including their scale, our coordinate transformation considerations primarily revolve around rotation and translation transformations. During the coordinate transformation procedure, we apply rotations around different coordinate axes at specific angles to derive the coordinate transformation matrix.

The three-dimensional coordinate conversion merely necessitates knowledge of the corresponding joint point coordinates in the two coordinate systems, allowing us to accurately calculate the rotation matrix R and translation matrix T that facilitate the transformation between these coordinate systems.

We establish the original coordinate system, denoted as S2, of device2 as the reference world coordinate system. Subsequently, we perform the conversion of data obtained from device 1 into the S2 world coordinate system. The corresponding three-dimensional coordinate transformation is articulated in Equation (1). In this equation, the rotation matrix R characterizes the angular adjustment between the new coordinate system and the original coordinate system, while the translation matrix T defines the spatial displacement of the origin of the new coordinate system relative to the old one. Notably, (x, y, z) denotes the coordinates within the original coordinate system S1 of device 1, whereas (x′, y′, z′) represents the new coordinates post-conversion into S2. Consequently, the converted coordinates (x′, y′, z′) in conjunction with the coordinates acquired by device2 collectively pinpoint the three-dimensional coordinate location of the same bone joint point.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + T \tag{1}$$

This study derives the rotation matrix and translation matrix utilizing the geometric principles of Singular Value Decomposition (SVD) [1]. SVD is a widely employed matrix decomposition technique within the realm of mathematics and computational science. From a geometric perspective, applying singular value decomposition to a matrix is akin to executing a sequence of transformations encompassing rotation, scaling, and vector space adjustments. Our method relies on knowledge of corresponding points in the two coordinate systems, subsequently employing SVD to calculate the rotation and translation matrices. This process facilitates the comprehensive transformation of points from one coordinate system to another, thereby enabling seamless data fusion between disparate coordinate systems.

To streamline the calculations, we begin by relocating both sets of data, centering their centroid coordinates at the origin. This simplification eliminates the need to factor in translation operations, allowing us to solely focus on the rotation operation, aligning the two coordinate systems for straightforward computation of the rotation matrix, denoted as R. The calculation formula for determining the center of mass is illustrated in Formula (2).

$$centroid_A = \frac{1}{N}\sum_{i=1}^{N} A^i \tag{2}$$

Among them, A represents the collection of data points obtained from device 1, where A^i signifies the coordinates of the i-th joint. The value of N is set to 32, representing the total number of joint points on the human body. The calculation for determining the center of mass for device 2 aligns with the principles described in Equation (3).

We proceed by accumulating two sets of decentralized data, denoted as (A – $centroid_A$) and (B – $centroid_B$), into a symmetric matrix H.

$$H = \left(A - centroid_A\right)\left(B - centroid_B\right)^T \tag{3}$$

Employ singular value decomposition technology to decompose H (Formula (4)). Where U is an orthogonal matrix, representing a rotation operation, while S is a diagonal matrix with elements along the diagonal representing singular values, indicating a scaling operation. Additionally, V^T corresponds to another orthogonal matrix, representing an additional rotation operation.

$$H = USV^T \tag{4}$$

The rotation matrix can be calculated according to Equation (5).

$$R = VU^T \tag{5}$$

Once the rotation matrix is determined, it can be employed to compute the translation matrix. This computation relies on identifying corresponding points between the two datasets, and the specific translation matrix can be computed following Equation (6). Here, R represents the rotation matrix obtained earlier, while $centroid_A$ and $centroid_B$ denote the central points of the data collected by device 1 and device 2 devices, respectively. Employing centroids for the

translation matrix calculation yields results that are both more stable and accurate, as it mitigates the influence of sampling variance or outliers.

$$T = centroid_B - \left(R * centroid_A\right) \quad (6)$$

3.2. Data Fusion

Following the coordinate calibration procedure, each joint is characterized by two sets of coordinates, denoted as g_{1i} and g_{2i}. Here, g_{1i} represents the coordinates of the i-th joint after conversion by device 1, while g_{2i} corresponds to the original coordinate data for the same joint collected by device 2. To merge these two sets of data, we employ the fusion algorithm outlined in Equation 7, where w_{1i} represents the fusion weight assigned to the i-th joint from device 1, and w_{2i} signifies the fusion weight for the identical joint from device 2.

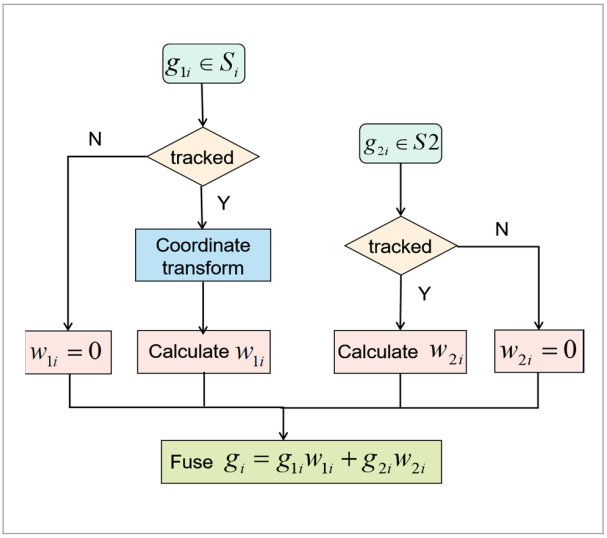
$$g_i = g_{1i}w_{1i} + g_{2i}w_{2i} \quad (7)$$

In instances where the device fails to successfully capture data for a specific joint point within a data frame, it becomes imperative to compensate for this missing data. A commonly employed method involves utilizing the corresponding joint point data from the preceding and subsequent frames and computing their average value to serve as the predicted value for the absent data. This approach offers the advantage of requiring minimal computational effort, ensuring speedy predictions, and achieving an acceptable level of prediction accuracy.

The Kinect SDK provides information regarding the tracking status of skeletal joint points, classifying them into tracked, inferred, and lost statuses. Among these, the data in the tracked status is the most reliable, followed by the inferred status, while data in the lost status cannot be deemed trustworthy. In cases where both Kinect devices are in the tracked or inferred states, their respective fusion weights for skeletal joint point data depend on the posture orientation and data stability of the human body in relation to the two Kinect devices. When only one Kinect is in either the tracked or inferred state, its data should be prioritized, and its fusion weight set to 1. In this case, the data from the other Kinect device, which is experiencing data loss, should be disregarded, with its fusion weight set to 0.

In summary, the fusion process is shown in Figure 2.

Figure 2
Fusion Process



It is worth noting that the accuracy of vision sensor data detection is influenced by the orientation of the human body posture. Different orientations yield varying levels of data accuracy. When the human body is directly facing the device, the quality of the data it captures is notably higher, as the sensor can obtain more direct visual information. To quantify the human posture orientation, we generate a posture vector for the human body using the position data of the left and right shoulder joints. The difference between these two positions constitutes the vector N. We then calculate the angles α and β between the human posture vector N and the XOY plane of the two devices using the following formulas:

$$\alpha = \arccos\frac{|Nx_1 + Ny_1|}{|N|} \quad (8)$$

$$\beta = \arccos\frac{|Nx_2 + Ny_2|}{|N|} \quad (9)$$

Among them, N_{x1} and N_{y1} are the vector N components along the X and Y axes, respectively, in the S1 coordinate system, while N_{x2} and N_{y1} represent the vector N components along the X and Y axes, respectively, in the S2 coordinate system. The angle α characterizes the orientation between the human body's shoulder joint and the front of device1, serving as a measure of the extent to which the human body faces

device1. A smaller α signifies a more direct alignment of the human body with Kinect1, thereby enhancing the credibility of the data collected by device1. Similarly, β signifies the angle between the human shoulder joint vector and the front of device2. A smaller β implies a more direct alignment of the human body with device2, resulting in increased reliability of the data collected by device2.

Inherent in the data obtained by the vision sensor are noise and outliers that necessitate identification and processing. Given the inherent coherence of human motion sequences, discrepancies between frames corresponding to different actions should generally be modest. This implies that the data should exhibit a certain degree of smoothness. We can assign weights to the fusion operation based on the smoothness of the data, thereby diminishing the influence of outliers. Consequently, the fused skeletal data becomes smoother and more natural, elevating its credibility. By computing the positional error of the same joint point between two consecutive frames, as elucidated in Equation (10), we determine the data weight. A larger error indicates a higher likelihood of the data being an outlier with low credibility, warranting a lower weight. Conversely, a smaller error suggests less data irregularity and higher credibility, justifying a higher weight.

$$err1 = \sqrt{\sum_{i=2}^{N}\left(A^{i} - A^{i-1}\right)^{2}} \tag{10}$$

Considering the factors outlined above, it becomes evident that a smaller angle between the human posture and the vision sensor device corresponds to a reduced smoothness error between frames, resulting in a higher fusion weight. This signifies that when the human posture is more closely aligned with the sensor device or when the data captured from the device exhibits lower inter-frame smoothness error, the device will carry greater weight in the fusion process of skeletal joints. Adhering to this principle, we can establish a weight for device1, as expressed in Equation (11), while a comparable weight for device2 is defined in Equation (12).

$$W_{1i} = \frac{\beta}{2(\alpha+\beta)} + \frac{err2}{2(err1+err2)} \tag{11}$$

$$W_{2i} = \frac{\alpha}{2(\alpha+\beta)} + \frac{err1}{2(err1+err2)} \tag{12}$$

4. Experimental Verification and Analysis

4.1. Dataset

To evaluate the reasonableness and effectiveness of the data fusion method proposed in this paper and to validate the effectiveness of the fusion algorithm for pose recognition, this paper uses the FMS (functional movement screen) dataset [27] for data fusion and model training.

The FMS dataset contains 3624 motion sequence samples, covering 7 major categories and 15 subcategories. This dataset was captured from two perspectives using two Azure Kinect cameras simultaneously, and can be used to validate the data fusion method proposed in this study. Each perspective contributed 1812 samples, totaling 3624 samples. These exercise samples were completed by 45 volunteers. In addition, two auxiliary Azure Kinect cameras were used to collect color images to supplement the data. Therefore, the dataset contains 3624 sets of color images and 3624 sets of depth images, each corresponding to an action sequence. In order to train and validate the data, this paper divides the dataset into training and validation sets. Specifically, actions performed by 30 participants were used as the training set, while the remaining 15 participants performed actions as the validation set.

4.2. Coordinate Transformation Verification

Verify the feasibility of coordinate conversion by performing coordinate calibration processing on bone data collected from two devices with different perspectives. To this end, we execute coordinate transformations on the data sourced from the FMS public dataset, following the procedure outlined in Section 2.1. This process serves to consolidate the coordinates into a unified world coordinate system. To evaluate the effectiveness of the coordinate transformation algorithm, the study randomly selected eight sets of sample data. Using the SpineBase joints, we compute two key metrics, namely RMSE (Root Mean Square Error) and MAE (Mean Absolute Error), to gauge the performance of the coordinate transformation algorithm. RMSE and MAE serve as measures of the coordinate transformation error, with smaller values signifying higher accuracy in the coordinate transfor-

mation process. Experimental results are shown in Tables 1-2, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - h\left(x_i\right)\right)^2} \quad (13)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left| y_i - h\left(x_i\right)\right| \quad (14)$$

Among them, y_i is the actual value, while $h(x_i)$ is the value following coordinate conversion.

Table 1
Root mean square error (RMSE) of X-axis, Y-axis, and Z-axis (unit: mm)

X-axis	Y-axis	Z-axis	Triaxial average
30.5	26.6	22.4	26.5
24.6	28.9	21.5	25.0
26.2	29.7	18.4	24.7
28.3	30.0	24.6	27.6
14.0	12.1	17.5	14.5
16.9	18.7	30.5	22.0
20.1	26.2	15.4	20.6
27.0	30.8	24.9	27.6

Table 2
Mean absolute error (MAE) of X-axis, Y-axis, Z-axis (unit: mm)

X-axis	Y-axis	Z-axis	Triaxial average
24.2	21.3	15.2	20.3
22.4	22.5	18.6	21.2
23.4	24.3	15.9	21.2
21.1	26.4	18.9	22.2
12.0	9.6	14.4	12.0
13.7	14.9	25.1	17.9
15.8	19.9	12.4	16.0
23.8	25.4	20.3	23.2

According to the results shown in the tables, the maximum root mean square errors for the X, Y, and Z axes are 30.5, 30.8, and 30.5, respectively, while the average maximum root mean square error for the three axes is 27.6. In addition, the maximum average absolute errors of the X, Y, and Z axes are 24.2, 26.4, and 23.2, respectively, while the average maximum absolute error of the three axes is 23.2.

These indicators collectively demonstrate that the data transformation method yields favorable fitting results on this dataset. Specifically, its root mean square error and mean absolute error are both low, indicating that this method has high accuracy. Therefore, it can be concluded that the coordinate transformation method used has shown significant superiority and effectiveness on this dataset.

4.3. Data Fusion Experimental Verification

To verify the effectiveness of data fusion methods in compensating for data loss and handling noise issues, we used the standard deviation of all joints for overall validation. Meanwhile, the square root of SpineBase joints was also used to verify the effectiveness of the fusion method in handling individual joint outliers.

By calculating the standard deviation of all joints, the overall effectiveness of data fusion methods in handling outliers can be evaluated. Standard deviation is an indicator of the degree of dispersion of points in a dataset, and a larger standard deviation implies greater data dispersion and variability. However, due to the inherent dispersion of human joints, using a total of 32 nodes to calculate the overall standard deviation may result in larger values of standard deviation. Therefore, when evaluating the effectiveness of data fusion algorithms, the continuity and smoothness of standard deviation should be mainly observed to verify whether the algorithm can effectively solve problems such as data missing and jumping, so as to make the generated motion data coherent and consistent. In addition, we chose to use the square root of SpineBase joints as an indicator to verify the effectiveness of handling individual joint outliers. The SpineBase joint is located at the base of the human spine, and its movement characteristics have a significant impact on body posture and movement patterns. By observing the outlier handling effect of the joint, we can evaluate the robustness and accuracy of the data fusion method at the individual joint level.

In summary, by comprehensively considering the standard deviation of all joints and the outlier processing effect for SpineBase joints, we can comprehensively evaluate and verify the outlier processing effect of the data fusion method. The standard deviation of all joints is shown in Figure 3, and the square root result of the SpineBase joint is shown in Figure 4.

Figure 3

(a) Standard deviation of X-axis in consecutive frames.
(b) Standard deviation of Y-axis in consecutive frames.
(c) Standard deviation of Z-axis in consecutive frames

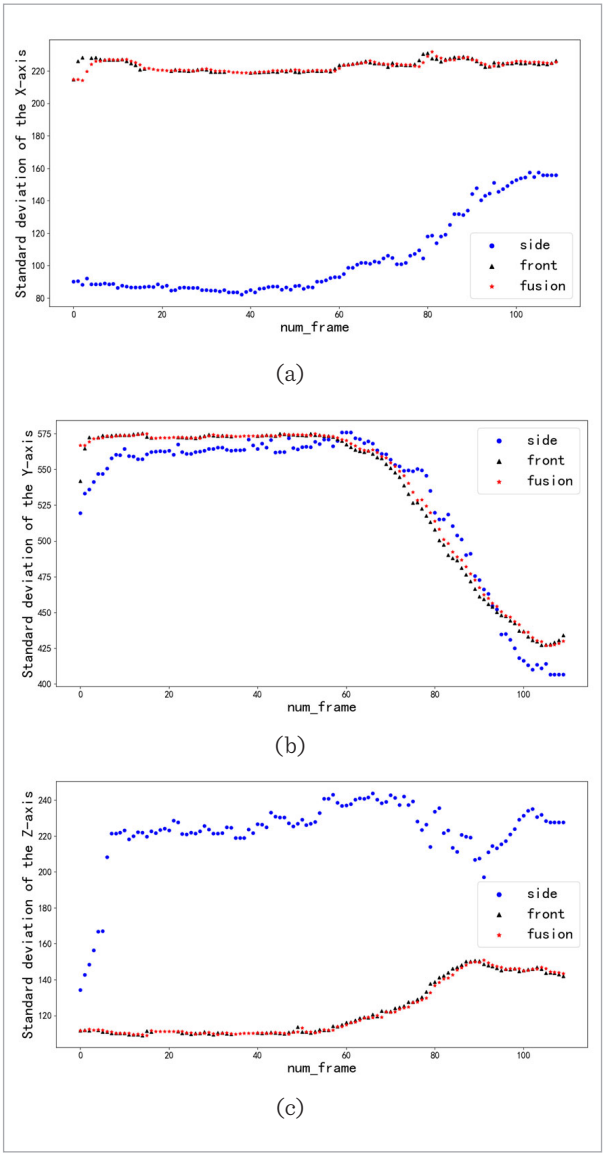
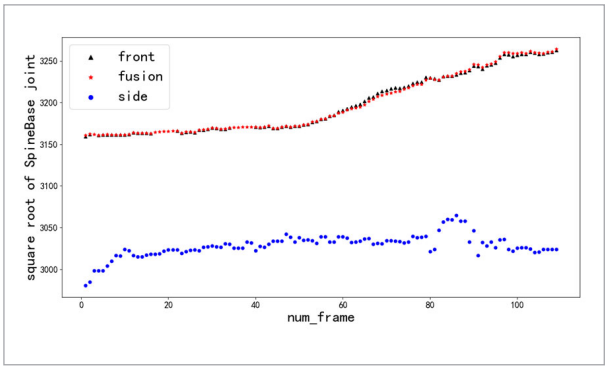


Figure 4

Square root of SpineBase joint in consecutive frames



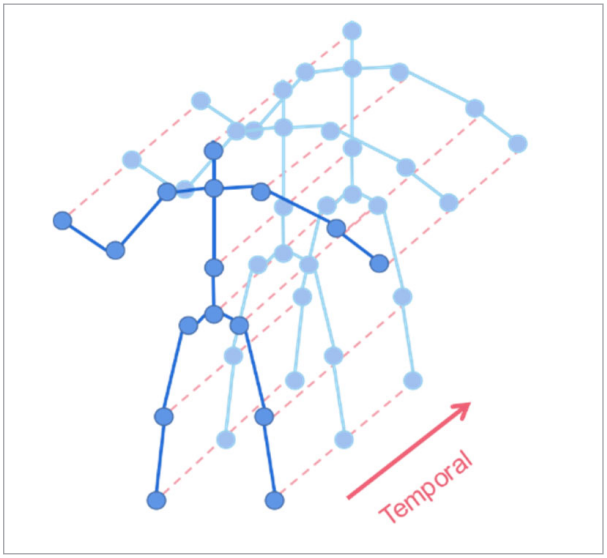
As shown in the figures, the data resulting from the fusion process exhibits a relatively comprehensive and continuous profile, effectively addressing issues such as data gaps and discontinuities. These observations affirm the effectiveness of the fusion algorithm proposed in this article. Additionally, data fusion enhances the smoothness of the acquired data, consequently enhancing the reliability of data and accuracy of posture recognition.

4.4. Human Posture Recognition Verification

4.4.1. Comparisons with Other Fusion Methods

ST-GCN (Spatial-Temporal Graph Convolutional Network [28]) stands out as the pioneer network to introduce graph convolutional networks into the field of human posture recognition. To capture the intricate relationships among human skeletal components, ST-GCN incorporates a graph structure, which serves as a representation of the connections between joints in the human body. It introduces temporal convolution and spatial convolution techniques tailored to this graph structure, enabling the handling of temporal sequencing relationships. Figure 5 visually delineates the human body graph structure. Within ST-GCN, each joint point is treated as a node within the graph, with the interconnections between them represented by the graph edges. These edges delineate the adjacency relationships among human body joint points, encompassing aspects such as bone connections and limb movement directions. Additionally, in the temporal dimension, the same node corresponding to the human body is linked by supplementary edges, accounting for temporal relationships.

Figure 5
Spatial temporal graph of a skeleton sequence proposed by ST-GCN



Through the construction of a graph structure representing the human skeleton, ST-GCN harnesses the power of graph convolution operations in the realm of posture recognition. These graph convolution operations excel at aggregating information from proximate regions and disseminating it across the global scale, enabling a more comprehensive understanding of spatial and temporal interdependencies within human poses.

In order to compare the effect of the data fusion method proposed in this article with the previous fusion method in terms of human posture recognition, we input the fused joint three-dimensional coordinate data into ST-GCN for training. Considering that the direction and length of human bones are crucial for posture recognition, we also input bone data and joint-bone data into ST-GCN for training to compare the results. The experimental results are shown in Table 3.

The experimental results demonstrate that the data fusion method proposed in this study outperforms previous methods in human pose recognition tasks. Specifically, this fusion method has achieved superior performance in both joint and joint bone scenarios. These findings provide substantial evidence to support the advantages of the data fusion method.

4.4.2. Human Posture Recognition Experiment

The objective of this experiment is to investigate the influence of data fusion on human posture recognition and validate the efficacy of the data fusion algorithm in enhancing posture recognition accuracy. Our approach involved conducting posture recognition on the original data and on fused skeleton data utilizing various state-of-the-art algorithms. Subsequently, we conducted a comparative analysis of the recognition accuracy, and the results are presented in Table 4.

The aforementioned experimental results clearly demonstrate that the data fusion method proposed in this study leads to a substantial enhancement in the accuracy of human posture recognition. Among the six human posture recognition models examined, the fused data consistently yielded the most favorable

Table 4
Comparison of accuracy (%) of data before and after fusion on state-of-the-art methods

Method	Fusion (ours)	Front	Side
Shift-GCN [4]	95.9	95.5	95.9
2s-AGCN [21]	98.2	96.9	95.7
DualHead-Net [3]	98.0	97.1	97.8
DGNN [20]	97.3	96.9	96.7
MS_G3D [16]	97.6	97.5	95.7
GCN-NAS [19]	98.0	98.0	95.2

Table 3
Comparison of accuracy (%) of different fusion methods on ST-GCN model

	Front	Side	Fusion1 [2]	Fusion2 [9]	Fusion3 **(ours)**
Joint	86.9	83.5	88.1	86.9	**89.1**
Bone	82.5	**86.2**	81.6	85.2	85.7
Joint-Bone	85.2	85.8	80.9	86.1	**87.5**

outcomes. This validation underscores the efficacy of the data fusion approach presented in this study, particularly in addressing issues related to self-occlusion of skeletal joints and underscores the practical applicability of the study.

5. Conclusion

This article introduces a data fusion method designed to amalgamate skeletal data collected from two vision sensor devices. The approach leverages human body posture orientation and data smoothness to determine the fusion weights for skeletal data, resulting in more dependable human posture data and an enhancement in the accuracy of skeletal data acquired through the vision sensor. The study draws upon the FMS public datasets for data fusion and the training of advanced human posture recognition algorithms. A comparative analysis of recognition accuracy between the original and fused data validates the efficacy of the proposed data fusion method, contributing significantly to the advancement of the field of human posture recognition.

In future endeavors, we can explore the integration of additional sensor devices for skeletal data fusion, expanding the scope of experiments and investigating optimal device placement strategies to attain the highest data accuracy and reliability. Furthermore, the inclusion of data from other sensors, such as inertial measurement units (IMU) or cameras, in skeletal data fusion is worth considering. This multi-modal data integration has the potential to further elevate the accuracy and resilience of posture recognition, offering broader possibilities for developments in related domain

References

1. Austin, D. Feature Column from AMS. http://www.ams.org/publicoutreach/feature-column/fcarc-svd. Access 21 June 2023.
2. Chen, N., Chang, Y., Liu H., Huang, L., Zhang, H. Human Pose Recognition Based on Skeleton Fusion from Multiple Kinects. 2018 37th Chinese Control Conference (CCC), 2018, 5228-5232. https://doi.org/10.23919/ChiCC.2018.8483016
3. Chen, T., Zhou, D., Wang, J., Wang, S., Guan, Y., He, X., Ding, E. Learning Multi-Granular Spatio-Temporal Graph Network for Skeleton-Based Action Recognition. Proceedings of the 29th ACM International Conference on Multimedia, 2021, 4334-4342. https://doi.org/10.1145/3474085.3475574
4. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-tion, 2020, 183-192. https://doi.org/10.1109/CVPR42600.2020.00026
5. Geerse, D. J., Coolen, B. H., Roerdink, M. Kinematic Validation of a Multi-Kinect V2 Instrumented 10-Meter Walkway for Quantitative Gait Assessments. PloS One, 2015, 10(10), e0139913. https://doi.org/10.1371/journal.pone.0139913
6. Guo, T., Chen, Y., Lin, L. Gesture Recognition Based on the Gesture Angle of Dual Kinect. Science Technology and Engi-Neering, 2019, 19(29), 172-178
7. Han, Y., Shan T. TwinLSTM: Two-Channel LSTM Net-Work for Online Action Detection. 2022 26th Interna-tional Conference on Pattern Recognition (ICPR), 2022, 3310-3317. https://doi.org/10.1109/ICPR56361.2022.9956717
8. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. Neural Computation, 1997, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
9. Jiang Y, Song K, Wang J. Action Recognition Based on Fusion Skeleton of Two Kinect Sensors. 2020 International Conference on Culture-Oriented Science & Technology (ICCST), 2020, 240-244. https://doi.org/10.1109/ICCST50977.2020.00052
10. Kipf, T. N., Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv Preprint ArXiv:1609.02907, 2016. https://doi.org/10.48550/arXiv.1609.02907
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient-Based Learning Applied to Document Recognition. Pro-ceedings of the IEEE, 1998, 86(11), 2278-2324. https://doi.org/10.1109/5.726791
12. Lee, H., Park, U., Kim, I. J., Cho, J. Rank-GCN for Robust Action Recognition. IEEE Access, 2022, 10, 91739-91749. https://doi.org/10.1109/ACCESS.2022.3202164
13. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M. Skele-Ton Based Action Recognition Using Translation-Scale Invariant Image Mapping and Multi-Scale Deep CNN.2017 IEEE Inter-National Conference on

Multime-dia & Expo Workshops (ICMEW), 2017, 601-604. https://doi.org/10.1109/ICMEW.2017.8026282

14. Liu, J., Shahroudy, A., Xu, D., Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. Com-Puter Vision-ECCV 2016: 14th Euro-pean Conference, Am-Sterdam, the Netherlands, October 11-14, 2016, Proceed-Ings, Part III 14. Springer Interna-tional Publishing, 2016, 816-833. https://doi.org/10.1007/978-3-319-46487-9_50
15. Liu, T., Ma, Y., Yang, W., Ji, W., Wang, R., Jiang, P. Spatial-temporal interaction learning based two-stream network for action recognition. Information Sciences, 2022, 606: 864-876. https://doi.org/10.1016/j.ins.2022.05.092.
16. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W. Dis-entangling and Unifying Graph Convolutions for Skele-ton-Based Action Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 143-152. https://doi.org/10.1109/CVPR42600.2020.00022
17. Müller, B., Ilg, W., Giese, M. A., Ludolph, N. Improved Kinect Sensor Based Motion Capturing System for Gait Assessment. BioRxiv, 2017, 098863. https://doi.org/10.1371/journal.pone.0175813
18. Peng, Q., Chen, W., Wu, X., Wang, J. A Novel Vision-Based Human Motion Capture System Using Dual-Kinect. 2015 IEEE 10th Conference on Industrial Elec-tronics and Applications (ICIEA), 2015, 51-56. https://doi.org/10.1109/ICIEA.2015.7334083
19. Peng, W., Hong, X., Chen, H., Zhao, G. Learning Graph Convolutional Network for Skeleton-Based Human Ac-tion Recognition by Neural Searching. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(03), 2669-2676. https://doi.org/10.1609/aaai.v34i03.5652
20. Shi, L., Zhang, Y., Cheng, J., Lu, H. Skeleton-Based Ac-tion Recognition with Directed Graph Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 7912-7921. https://doi.org/10.1109/CVPR.2019.00810
21. Shi, L., Zhang, Y., Cheng, J., Lu, H. Two-Stream Adap-tive Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition, 2019, 12026-12035. https://doi.org/10.1109/CVPR.2019.01230
22. Simonyan, K., Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems, 2014, 27.
23. Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H. Scanning 3D Full Human Bodies Using Kinects. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(4), 643-650. https://doi.org/10.1109/TVCG.2012.56
24. Wang, Z., Lu, H., Jin, J., Hu, K. Human Action Recogni-tion Based on Improved Two-Stream Convolution Net-work. Applied Sciences, 2022, 12(12), 5784. https://doi.org/10.3390/app12125784
25. Wu, B., Wan, A., Yue, X., Jin, P., Zhao, S., Golmant, N., Keutzer, K. Shift: A Zero Flop, Zero Parameter Alterna-tive to Spatial Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion, 2018, 9127-9135. https://doi.org/10.1109/CVPR.2018.00951
26. Wu, D., Shao, L. Leveraging Hierarchical Parametric Networks for Skeletal Joints Based Action Segmentation and Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, 724-731. https://doi.org/10.1109/CVPR.2014.98
27. Xing, Q., Shen, Y., Cao, R., Zong, S., Zhao, S., Shen, Y. Functional Movement Screen Dataset Collected with Two Azure Kinect Depth Sensors. Scientific Data, 2022, 9(1), 104. https://doi.org/10.1038/s41597-022-01188-7
28. Yan, S., Xiong, Y., Lin, D. Spatial Temporal Graph Con-volutional Networks for Skeleton-Based Action Recogni-tion. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1). hhttps://doi.org/10.1609/aaai.v32i1.12328
29. Zaremba, W., Sutskever, I., Vinyals, O. Recurrent Neural Network Regularization. arXiv Preprint ArXiv:1409.2329, 2014. https://doi.org/10.48550/arXiv.1409.232's