Article

Not peer-reviewed version

Viewpoint-agnostic Taekwondo Action Recognition using a Synthesized Two-dimensional Skeletons

Chenglong Luo , Sung-Woo Kim , Hun-Young Park , Kiwon Lim , Hoeryong Jung *

Posted Date: 21 August 2023

doi: 10.20944/preprints202308.1432.v1

Keywords: Taekwondo poomsae; action recognition; skeletal data; camera viewpoint; martial arts

Article

Viewpoint-agnostic Taekwondo Action Recognition using a Synthesized Two-dimensional Skeletons

Chenglong Luo [1], Sung-Woo Kim [2], Hun-Young Park [2,3], Kiwon Lim [2,3,4] and Hoeryong Jung [1,3,*]

[1] Division of Mechanical and Aerospace Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea; luo0611@konkuk.ac.kr (C.L.); junghl80@konkuk.ac.kr (H.J.)

[2] Physical Activity and Performance Institute, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea; kswrha@konkuk.ac.kr (S.-W.K.); parkhy1980@konkuk.ac.kr (H.-Y.K.) exercise@konkuk.ac.kr (K.L.)

[3] Department of Sports Medicine and Science, Graduate School, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea

[4] Department of Physical Education, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea

* Correspondence: junghl80@konkuk.ac.kr; Tel.: +82-2-450-3903

Abstract: Issues of fairness and consistency in Taekwondo poomsae evaluation have emerged owing to the lack of an objective evaluation method. This study proposes a three-dimensional (3D) convolutional neural network (CNN)-based action recognition model for the objective evaluation of Taekwondo poomsae. The model exhibits robust recognition performance regardless of variation in perspective by reducing the discrepancies between training and test images. The model uses 3D skeletons of the poomsae unit action collected using a full-body motion-capture suit to generate synthesized two-dimensional (2D) skeletons from the desired perspective. This approach aids in obtaining 2D skeletons from diverse perspectives as part of the training dataset and ensures consistent recognition performance regardless of the viewpoint. The model was trained using 2D skeletons projected from diverse viewpoints, and its performance was evaluated using various test datasets, including projected 2D skeletons and RGB images captured from various viewpoints. Comparison of the performance of the proposed model with that of previously reported action recognition models demonstrated the superiority of the model, underscoring its effectiveness in recognizing and classifying Taekwondo poomsae actions.

Keywords: Taekwondo poomsae; action recognition; skeletal data; camera viewpoint; martial arts

1. Introduction

Taekwondo is a traditional Korean martial art that has become one of the most popular sports worldwide. Two types of Taekwondo competitions are conducted: gyeorugi and poomsae, which involve various movements and complex techniques. Gyeorugi requires two competing players, and objective judgments are made using a quantitative and accurate electronic scoring system. In Taekwondo poomsae, a single player demonstrates basic attack and defense techniques in a specific order. In this case, evaluation is subjective and qualitative, based on the judges' opinions, except for penalties (e.g., stopping or exceeding boundaries). Owing to situational constraints, judges must evaluate multiple participants simultaneously, which may cause issues related to the fairness and consistency of evaluations not only in competitions but also in promotional tests. To address these issues, quantitative evaluation methods using vision-based action recognition techniques have been proposed [1,2].

Vision-based human action recognition (HAR) has emerged as a prominent area of interest in computer vision and artificial intelligence. Its primary objective is to detect and analyze human actions from unknown video sequences, thereby enabling a deeper understanding and interpretation of such actions. HAR has been applied to various domains, including security [3–5], healthcare [6–8], and sports [9–11]. Vision-based HAR systems have been employed to support quantitative evaluation

doi:10.20944/preprints202308.1432.v1

and judgment in various sports [12–18]. However, few studies have reported the application of action recognition technology in martial arts disciplines, such as Taekwondo [1,2]. In previous studies, action recognition approaches using RGB and RGB-D images have been proposed. These included methods that emphasized the dominant poses associated with each action in RGB-D videos as an input to convolutional neural networks (CNNs) [19] as well as techniques that enhanced the structural information of body parts, joints, and temporal scales by representing sequences of depth maps as structured dynamic images [20]. However, the rapid-action characteristics of martial arts pose challenges to motion capture because of insufficient sharpness and intermittent frame loss. Furthermore, the dynamic nature and wide range of actions in martial arts render the RGB-D methods inadequate. These approaches are also susceptible to domain shifts caused by environmental changes and cannot accurately predict dynamic actions.

Recent research on action recognition incorporated skeletal data into complex human action recognition [21–30]. For instance, Du et al. proposed an architecture that divided the human skeleton into five parts and fed them into separate subnetworks instead of using recurrent neural networks to process the entire skeleton as an input [31]. Yan et al. introduced an action recognition method based on graph convolutional networks (GCNs) considering the spatiotemporal features of skeletons [19]. Subsequently, GCN-related studies [23,25–27,29], including those by Duan et al., generated heatmaps using skeletons to address the limitations of GCN methods related to the accuracy of skeleton coordinates and integration with other modality data [22]. In skeleton-based action recognition, skeleton representation provides core information that is highly relevant to human behavior. Unlike RGB-D models, it remains robust against variations in illumination, changes in clothing, and environmental factors.

Previous approaches used for action recognition primarily relied on images obtained from a single perspective. However, in the context of the poomsae evaluation, the same movement may appear differently when captured from different viewpoints, posing challenges for accurate recognition. Furthermore, single-view action recognition requires training models specific to each viewpoint, thereby necessitating retraining efforts when dealing with images captured from other viewpoints. This results in potential time and resource constraints. Moreover, single-view action recognition predominantly focuses on discerning individual movements, thereby presenting difficulties in recognizing complex movements involving multiple actions.

In this study, we propose a novel action recognition model for the evaluation of poomsae that exhibits robust recognition performance regardless of viewpoint variations. The model uses 3D skeletons collected using a full-body motion-capture suit to create 2D skeletons from a desired perspective. Thus, the proposed approach obtains 2D skeletal data from diverse viewpoints as part of the training data and effectively addresses the effect of observation perspectives, ensuring consistent and reliable performance in action recognition regardless of the viewpoint. The main contributions of this study are as follows:

1. A 3D skeletal dataset comprising 16 unit actions in Taekwondo poomsae was constructed using motion data collected using full-body motion-capture suits.
2. This study proposes methods for generating 2D skeletons by projecting 3D skeletons from diverse viewpoints and synthetic joint and bone heatmaps to incorporate viewpoint-dependent action characteristics into the training dataset. This ensured consistent and reliable performance, regardless of the viewpoint.
3. The optimal camera viewpoint for action recognition of Taekwondo poomsae was determined via the analysis and evaluation of recognition performance.

2. Materials and Methods

2.1. Data Collection

Primary motion data were collected from Taekwondo experts using a full-body motion-capture suit (Xsens MVN, Xsens Corp., Netherlands). The suit was equipped with 17 IMU sensors, and each sensor measured the 3-axes acceleration, angular velocity, and orientation of the body segment at the attached point. Subsequently, the raw data obtained from the motion-capture suit were processed to extract the positions of 23 joints in the human skeleton [33]. To enhance the generalizability of the action recognition model, the skeleton with 23 joints was converted to one with 16 joints, as illustrated in Figure 1. Forty Taekwondo experts participated in data collection. The experimental protocol was approved by the Konkuk University Institutional Review Board (7001355-202004-HR-372). Each subject was instructed to sequentially perform the 16-unit actions of Taekwondo poomsae as a predefined data-collection protocol while wearing the motion-capture suit. Each subject executed each action four times and repeated the process three times, resulting in 12 sets of executions for each unit action. Consequently, a Taekwondo unit-action dataset comprising 7,680 unit action data (16-unit actions per subject, repeated 12 times, with 40 participants) was prepared, and the motion database, which was named the Taekwondo Unit Action Dataset of 3D skeletons (TUAD-3D), was constructed, as depicted in Figure 2.

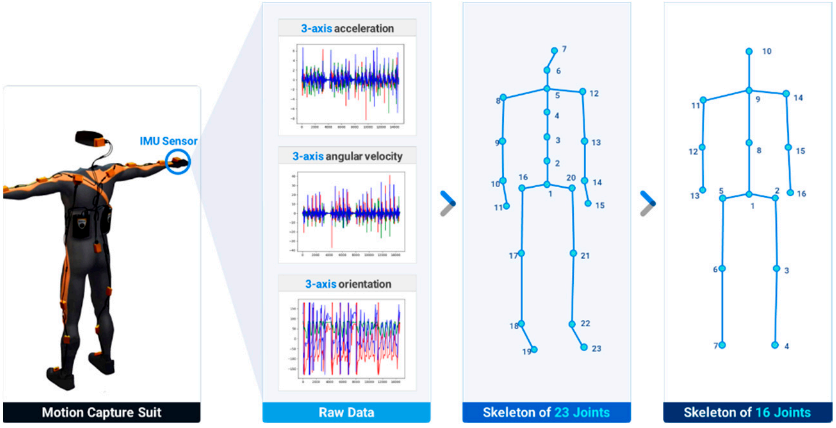


Figure 1. Three-dimensional (3D) skeleton data-collection procedure

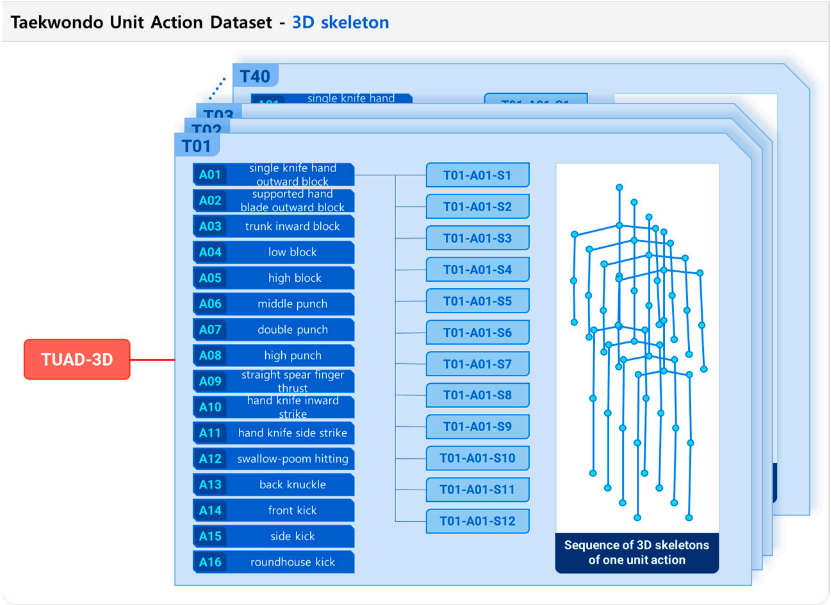


Figure 2. Structure of Taekwondo Unit Action Dataset of 3D skeletons (TUAD-3D)

2.2. Three-dimensional (3D) Convolutional Neural Network (CNN)-based Viewpoint-agnostic Action Recognition

The viewpoint-agnostic action recognition proposed in this study adopted a previously reported posec3d framework, which utilized a sequence of 2D skeleton heatmaps as input to 3D CNNs as the primary action recognition architecture [22]. To address the performance degradation caused by viewpoint mismatch in training and test images, this study proposed a method for using diverse-viewpoint 2D skeletons, which were generated through the projection of 3D skeletons, as the training dataset. Finally, the 2D skeletons were converted into synthetic heatmap images and used to train the action recognition network. Figure 3 illustrates the action recognition architecture proposed in this study.

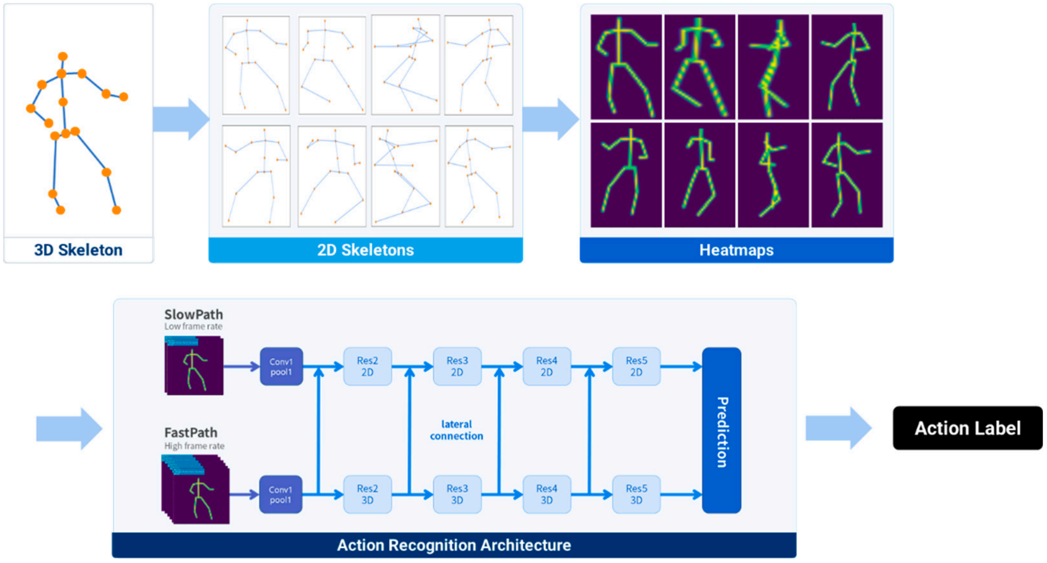


Figure 3. Overall architecture of three-dimensional (3D) convolutional neural network (CNN)-based viewpoint-agnostic action recognition.

2.2.1. Generation of Diverse Viewpoint Two-dimensional (2D) Skeletons from 3D Skeleton

Figure 4 illustrates the projection of the 3D skeleton onto the image planes of various camera viewpoints to generate 2D skeletons with diverse perspectives. In this procedure, we assumed that the camera could be rotated in a fixed orbit around the center of the 3D skeleton, as depicted in Figure 4(a). The position of the camera was calculated by multiplying the rotation matrix $\boldsymbol{R}_z(\theta)$ with its initial position $\boldsymbol{p}_0$, as follows:

$$\boldsymbol{p}_\theta = \boldsymbol{R}_{z,\theta}\,\boldsymbol{p}_0, \tag{1}$$

where $\boldsymbol{p}_0$ and $\boldsymbol{p}_\theta$ denote the initial and rotated camera positions, respectively. To incorporate various perspectives, the joint positions of the 3D skeleton were rotated in intervals of 10°, 45°, and 90°. This rotation operation facilitated the projection of the 3D skeleton keypoints onto a 2D image plane, thereby transforming the skeleton information in 3D space into the corresponding 2D skeleton information. Although the process of rotation resulted in a partial loss of position and orientation information of the rotated skeleton, it effectively enabled the representation of 2D skeleton information from diverse perspectives. The 2D skeleton at the rotated camera position $\boldsymbol{p}_\theta$ was projected by multiplying the projection matrix $\boldsymbol{P}_\theta$ with the 3D skeleton coordinates, as follows:

$$\boldsymbol{s}_{i,\theta}^{2D} = \boldsymbol{P}_\theta\,\boldsymbol{s}_i^{3D}, \tag{2}$$

where $\boldsymbol{s}_i^{3D}$ denotes the ith joint position of the 3D skeleton, and $\boldsymbol{s}_{i,\theta}^{2D}$ is the joint position of the 2D skeleton corresponding to the camera position $\boldsymbol{p}_\theta$. The projection matrix $\boldsymbol{P}_\theta$ can be acquired using intrinsic and extrinsic camera parameters. An intrinsic parameter characterizes the optical properties of the camera, whereas an extrinsic parameter matrix describes its position and orientation.

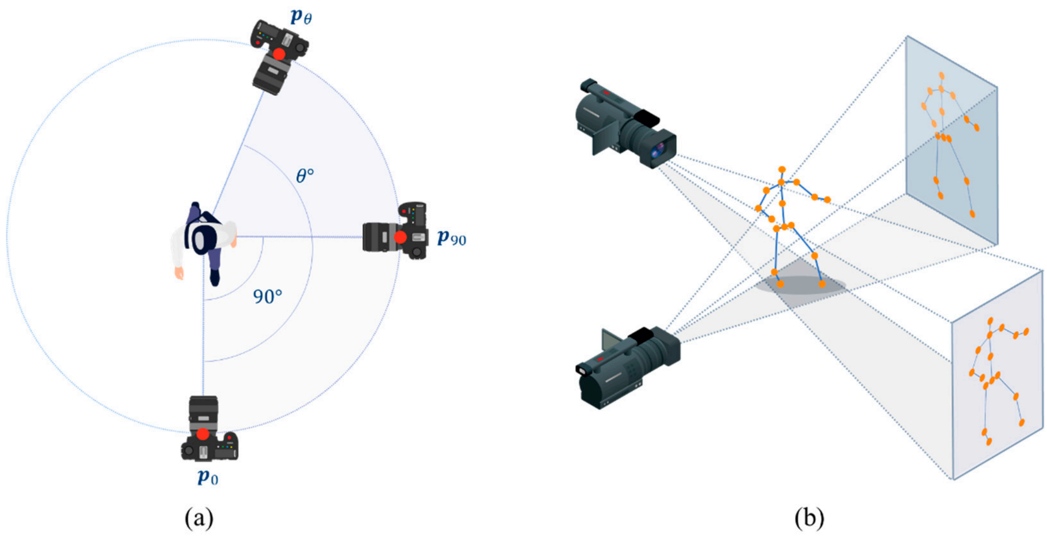


Figure 4. Generation of 2D skeletons by projecting the 3D skeleton onto various image planes. (a) The camera viewpoints can be determined by rotating the initial viewpoint. (b) Projection of the 3D skeleton onto the desired image plane.

2.2.2. Generation of Synthetic Heatmap Image from 2D Skeleton

The joint positions of the 2D skeletons were employed to generate synthetic 2D heatmap images. The value assigned to each pixel coordinate within the heatmap image was determined by applying a Gaussian kernel to that coordinate. The heatmaps generated were categorized into bone and joint heatmaps. To reduce the volume of the 3D heatmaps, we implemented two techniques. The first technique was subject-centered cropping, which involved cropping all the frames based on a minimum bounding box that enclosed the subject in a 2D pose. The frames were cropped as the participant moved within a confined area. Subsequently, the cropped frames were resized to the desired target size. The second technique was uniform sampling, which entailed selecting a subset of frames to capture the temporal dimension. This sampling approach ensured that the frames were uniformly distributed throughout the sequence, thereby effectively reducing the computational load.

The bone heatmap served as a visual representation of skeletal connectivity. It depicted the interconnections between different segments of the skeleton, facilitating the comprehension of its structural arrangement and the tracking of joint movements. In contrast, the joint heatmap focused on representing the central point of each skeletal segment. This enabled the precise localization of joint positions and provided a more detailed understanding of the skeletal shape, which was utilized for motion analysis. The training and validation procedures were conducted separately for the two types of heatmaps to ensure their individual accuracies. The pixel value of the joint heatmap $J_{i,j}$ was calculated as follows:

$$J_{i,j} = \sum_{k=1}^{NbJoint} \exp\left(-\frac{D(i,j,\boldsymbol{u}_k)}{2\sigma^2}\right), \tag{3}$$

where σ is the variance of the Gaussian map, and $D(i,j,\boldsymbol{u}_k)$ represents the Euclidean distance between pixels $(i,\ j)$ and the k^{th} joint position of the 2D skeleton $(\boldsymbol{u}_k)$. The pixel value of the bone heatmap $B_{i,j}$ was calculated as follows:

$$B_{i,j} = \sum_{k=1}^{NbBone} \exp\left(-\frac{D(i,j,\boldsymbol{b}_k)}{2\sigma^2}\right), \tag{4}$$

where $D(i,j,\boldsymbol{b}_k)$ denotes the shortest distance between pixel $(i,\ j)$ and the k^{th} bone segment $\boldsymbol{b}_k$, which is defined by the two joint positions of the 2D skeleton. After the abovementioned process, 2D joint and bone heatmaps were generated for each 2D skeleton. This process resulted in a 3D heatmap with the dimensions of $T \times H \times W$ for each action sequence, where T is the number of frames in each action, and H and W represent the height and width of the heatmap image, respectively.

doi:10.20944/preprints202308.1432.v1

2.2.3.3. D CNN Architecture

The SlowFast architecture was employed to construct a 3D CNN action-classification model [34]. It consisted of two distinct pathways, namely slow and fast pathways, as illustrated in Figure 5. The slow pathway was designed to effectively retain spatial information, whereas the fast pathway preserved temporal information. By combining these two pathways, the SlowFast architecture demonstrated an enhanced capability in capturing both spatial and temporal features, which resulted in improved accuracy for action-classification tasks.

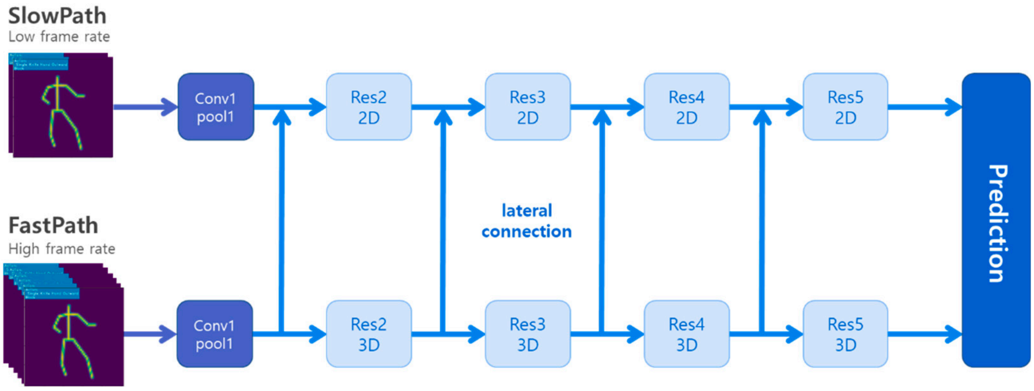


Figure 5. 3D CNN architectures

2.2.4. Training Procedure

The 3D skeletal databases of the 16-unit actions of poomsae, collected from 40 Taekwondo experts, were used to train the proposed action-classification model. During the training phase, we assessed the generalization performance via 5-fold cross-validation. The model was trained using the stochastic gradient descent optimizer with a maximum of 240 epochs, and cross-entropy loss was employed as the chosen loss function.

2.3. Evaluation Metrics

The evaluation metrics used in the experiment were F1-score, precision, recall, and accuracy.

$$precision = \frac{TP}{TP + FP}, \tag{5}$$

$$recall = \frac{TP}{TP + FN}, \tag{6}$$

$$accuracy = \frac{TP + TN}{TP+FP+TN + FN}, \tag{7}$$

where TP (true positive) represents samples that are predicted as positive, and the ground truth also labels them as positive, whereas FP (false positive) represents samples that are predicted as positive, but the ground truth labels them as negative. TN (true negative) represents samples that are predicted as negative, and the ground truth labels them as negative, whereas FN (false negative) represents samples that are predicted as negative, but the ground truth labels them as positive. The F1-score is a metric that balances precision and recall and measures the performance of the model. A higher F1-score indicates better performance.

$$F1 - Score = \frac{2*precision*recall}{precision+recall} \tag{8}$$

3. Results

The action recognition model was trained on four 2D skeletal databases. Table 1 lists the configurations of the training dataset. Each training dataset comprised 2D skeletons generated by projecting a 3D skeleton at several viewing angles. The models trained using these datasets were denoted as Models A–D, as listed in Table 1. The performance of these models was compared to deduce the optimal configuration of the projection viewpoints of the 2D skeletons for the highest recognition performance.

Table 1. Viewpoint configuration of 2D skeletons of four training datasets.

ID	Viewpoint configuration	Number of viewpoints	Number of training data
Model A	0°, 90°	2	12,360
Model B	0°, 90°, 180°, 270°	4	24,720
Model C	0°, 45°, 90°, 135°, ⋯ , 315°	8	49,440
Model D	0°, 10°, 20°, 30°, ⋯ , 350°	36	222,480

3.1. Performance Evaluation using Synthetic 2D Skeleton Datasets

The performance of the model was assessed using a synthesized 2D skeletal dataset. Test samples for the 2D skeletal data were generated by projecting the 3D skeletons of ten individuals selected from TUHA-3D at 10° intervals across viewpoints. Tables 2 and 3 present the evaluation results of the joint and bone heatmaps, respectively. While training and testing with the joint heatmap, the highest performance was observed for Model D, with an accuracy of 0.9802. Similarly, while training and testing the bone heatmap, the highest performance was observed for Model D, with an accuracy of 0.9783. The performance comparison results show that the recognition accuracy increased as more 2D skeletons projected at distinct viewing angles were included in the training dataset.

Table 2. Performance evaluation results of the joint heatmap model tested using a random projection 2D skeletal dataset.

Model	Precision	Recall	F1-score	Accuracy
Model A	0.8611	0.8398	0.8373	0.7997
Model B	0.9680	0.9669	0.9670	0.9669
Model C	0.9769	0.9764	0.9765	0.9764
Model D	0.9803	0.9802	0.9802	0.9802

Table 3. Performance evaluation result of the bone heatmap model tested using a random projection 2D skeletal dataset.

Model	Precision	Recall	F1-score	Accuracy
Model A	0.8611	0.8398	0.8373	0.7997
Model B	0.9686	0.9682	0.9682	0.9682
Model C	0.9769	0.9764	0.9765	0.9764
Model D	0.9786	0.9783	0.9784	0.9783

3.2. Performance Evaluation using 2D Skeletons Extracted from Front- and Side-view RGB Images

The performance of the proposed model was evaluated using 2D skeleton data extracted from RGB images. Test samples of the 2D skeleton data were extracted from the poomsae unit-action images captured using an RGBD camera (Realsense 435d; Intel Corporation, USA). In the data collection procedure, two additional RGBD cameras were installed in the front and on the left-hand side of the participants to collect test sample images. Figure 6 depicts the RGB images captured by the frontal and lateral cameras. Overall, 5,527 test samples of 2D skeletons were generated from the RGB images using the HRnet pose-estimation algorithm. Tables 4 and 5 present the evaluation results using the joint and bone heatmaps, respectively. While training and testing with the joint heatmap, the highest performance was observed for Model D, with an accuracy of 0.8705. Similarly, while training and testing the bone heatmap, the highest performance was observed for Model C, with an accuracy of 0.8761. The performance comparison results demonstrate the effectiveness of the action recognition model trained with synthetic 2D skeletons for the RGB test samples.

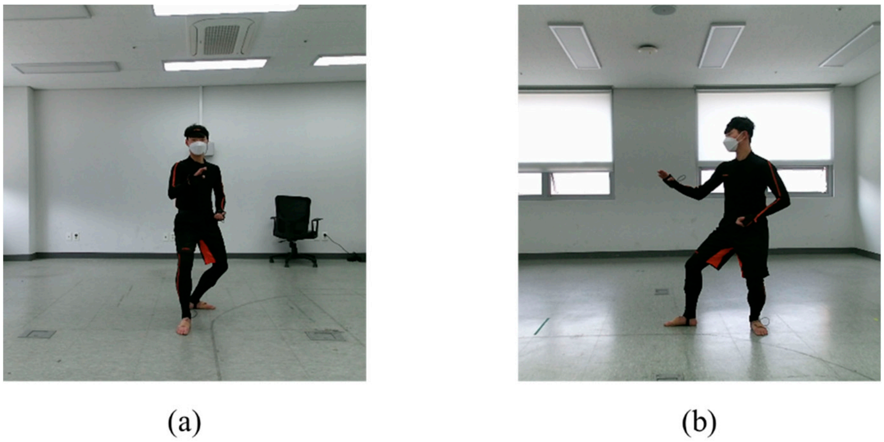


Figure 6. RGB image samples used for performance evaluation. (a) Frontal image and (b) lateral image

Table 4. Performance evaluation results of the joint heatmap model using 2D skeletal dataset extracted from RGB images.

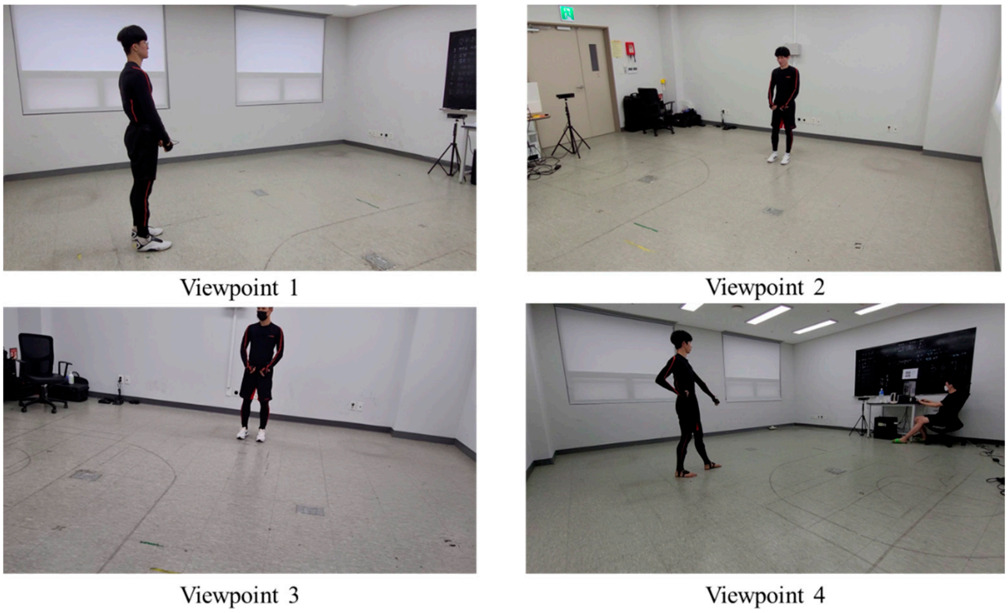
Model	Precision	Recall	F1-score	Accuracy
Model A	0.7638	0.7638	0.5854	0.5795
Model B	0.8761	0.8533	0.8500	0.8516
Model C	0.8705	0.7647	0.7763	0.7626
Model D	0.8998	0.8717	0.8706	0.8705

Table 5. Performance evaluation result of bone heatmap model using 2D skeletal dataset extracted from RGB images.

Model	Precision	Recall	F1-score	Accuracy
Model A	0.7453	0.6559	0.6549	0.6549
Model B	0.8750	0.8303	0.8300	0.8294
Model C	0.8903	0.8766	0.8752	0.8761
Model D	0.8891	0.8743	0.8717	0.8732

3.3. Performance Evaluation using 2D Skeletons Extracted from Random View RGB Images

The performance of the model was evaluated using a 2D skeletal dataset extracted from RGB images captured from random viewpoints. The RGB image dataset, obtained using four smartphone cameras from four distinct viewpoints, was used for the assessment. The dataset encompassed 639 samples of poomsae unit actions. Figure 7 illustrates the RGB image samples used for performance evaluation. Tables 6 and 7 present the evaluation results using the joint and bone heatmaps, respectively. While training and testing with the joint heatmap, the highest performance was observed for Model C, with an accuracy of 0.9381. Similarly, when training and testing the bone heatmap, the highest performance was observed for Model D, with an accuracy of 0.8670. A performance comparison shows that the action recognition model trained with synthetic 2D skeletons works for test samples obtained from random-view RGB images.



Viewpoint 1

Viewpoint 2

Viewpoint 3

Viewpoint 4

Figure 6. RGB image samples captured from random viewpoints used for performance evaluation.

Table 6. Performance evaluation results of the joint heatmap model using 2D skeletal dataset extracted from RGB images captured in random viewpoints.

Model	Precision	Recall	F1-score	Accuracy
Model A	0.8298	0.8185	0.7944	0.8398
Model B	0.9217	0.9009	0.9037	0.9010
Model C	0.9432	0.9381	0.9384	0.9381
Model D	0.8977	0.8702	0.8682	0.8623

Table 7. Performance evaluation results of the bone heatmap model using 2D skeletal dataset extracted from RGB images captured in random viewpoints.

Model	Precision	Recall	F1-score	Accuracy
Model A	0.7142	0.6222	0.6095	0.6041
Model B	0.8665	0.7862	0.7930	0.7715
Model C	0.8848	0.8471	0.8482	0.8419
Model D	0.9040	0.8736	0.8764	0.8670

3.4. Performance Comparison with Previous Models

The performance of the proposed model was compared with those of previously reported action recognition models, including posec3d [22], stgcn [29], stgcn++ [35], ctrgcn [26], and aagcn [36]. For the evaluation, the proposed model was trained using the 2D skeletal databases of Models C and D. The previously reported models were trained using 2D skeletal databases extracted from the RGB images of the poomsae unit action captured from frontal and lateral viewpoints presented in Section 3.2. Random-viewpoint RGB image datasets were used as test datasets for the proposed and previously reported models. The detailed outcomes of this evaluation for the random-view RGB image test dataset are listed in Table 8. Notably, the proposed model trained with the 2D skeletal databases of Model D exhibited superior performance, achieving an accuracy of 0.8670.

Table 8. Performance comparison between proposed and previously reported models.

Model	Precision	Recall	F1-score	Accuracy
stgcn [29]	0.7257	0.6915	0.6715	0.6667
stgcn++ [35]	0.8144	0.7210	0.7417	0.7058
ctrgcn [26]	0.7944	06509	0.6587	0.6275
aagcn [36]	0.8040	0.7442	0.7541	0.7261
posec3d [22]	0.8825	0.7520	0.7865	0.7340
Proposed (Model C)	0.8848	0.8471	0.8482	0.8419
Proposed (Model D)	0.9040	0.8736	0.8764	0.8670

4. Discussion

This study examined the efficacy of action recognition of Taekwondo poomsae. This was achieved by examining various training and testing datasets. The performance of four models, namely A, B, C, and D, trained using 2D skeletal representations obtained by projecting 3D skeletons from diverse camera viewpoints was evaluated and contrasted across distinct testing datasets. The evaluation outcomes of the 2D skeletal data obtained by projecting the 3D skeleton at 10° intervals across the viewpoints revealed that Model A achieved an accuracy of only 0.7997. In contrast, Models B, C, and D achieved an accuracy of 0.96. This observation underscores the insufficiency of relying solely on frontal and lateral viewpoint data to recognize actions from other perspectives.

The performance of the proposed model was evaluated using RGB images captured from the frontal and lateral viewpoints. Among the four models assessed, Model A exhibited the lowest accuracy with values of 0.5795 and 0.6549 for the joint and bone heatmap models, respectively. This discrepancy can be attributed to subtle variations in the relative positioning of key points between the 2D skeletal representations obtained via the projection of 3D skeletons and those extracted from the images, leading to a degradation of recognition performance. Conversely, Model D demonstrated the highest accuracy of 0.8705 in the joint heatmap model, whereas Model C achieved the highest accuracy of 0.8761 in the bone heatmap model. This highlights the potential enhancement of the recognition performance achieved via the incorporation of projection data from different viewpoints.

The assessment encompassed image data captured from random viewpoints, and 2D skeletal representations were extracted from images obtained using HRNet from those viewpoints. Model C achieved the highest joint heatmap precision of 0.9381, whereas Model D attained a peak skeletal heatmap precision of 0.8670. The observed decline in accuracy during image data evaluation can be attributed to the disparities between the 2D skeletal representations obtained via the projection of 3D skeletons and those estimated using the estimation algorithm. Despite efforts to align them by discarding significantly discrepant key points, inherent misalignment effects on accuracy persisted.

To compare existing models with the proposed approach, they were trained on sample images acquired from two RGB-D cameras positioned frontally and laterally. The images were employed to extract the 2D skeletal representations that were used for training. A comparative assessment was carried out by juxtaposing the models on a test dataset of random-view RGB images, where Model C exhibited better performance than the existing models by more than 10 %. This observation underscores the high recognition performance of the proposed method for Taekwondo poomsae from arbitrary viewpoints. The motion-recognition methodology demonstrates high precision across distinct viewpoints and datasets. The outcomes demonstrate the elevated recognition performance of Models C and D. Furthermore, compared with previously reported models the proposed model exhibits superiority in action recognition, which highlights its efficacy in real-world scenarios.

This study employed various training and testing datasets, yet the variability in the data might not fully encapsulate the complexity and diversity of real-world scenarios. The effectiveness of the proposed method could be constrained by the degree of diversity in the data. When tested on images from different perspectives, the model's performance exhibited a decline, indicating sensitivity to varying viewpoints. However, the assessment was limited to the Z-axis, and the extent of viewpoint sensitivity across a broader range remains an unresolved issue. The evaluation of this study focused on Taekwondo Pumsae movements, and the generalizability of the proposed approach to other action recognition tasks remains to be explored. Further validation is necessary for the model's efficacy in more extensive action recognition scenarios. Future research could benefit from integrating a wider array of data sources to capture diverse lighting conditions, backgrounds, and environmental factors. This would ensure a more comprehensive evaluation of the proposed method's robustness. To overcome limitations associated with viewpoint sensitivity and misalignment, the integration of multiple data modalities (such as RGB and depth) might enhance recognition accuracy across different perspectives. Addressing misalignment effects might involve refining human pose estimation algorithms or exploring techniques that explicitly handle pose variations induced by projection and estimation discrepancies.

5. Conclusions

This study constructed a TUAD-3D dataset by employing full-body motion-capture suits to collect accurate 3D skeletal data. This dataset contained data of 7,680 samples and included 16 fundamental techniques performed by 40 Taekwondo experts. The model effectively synthesized 2D skeletal representations from the collected 3D skeletal data and integrated multiple viewpoints during the training process. This approach ensures consistent and reliable model performance, regardless of the observer’s angles and positions. Through a comprehensive evaluation of various action recognition networks, we observed that Models C and D, which were trained using 3D skeleton projection, exhibited higher accuracy. The assessment results demonstrate the superiority of the proposed model over those reported previously, highlighting its effectiveness in Taekwondo poomsae action recognition and classification. Furthermore, analysis across different viewpoints and datasets revealed the significance of optimal camera viewpoint selection for training, which influenced the model performance. However, a decline in accuracy was observed during the model evaluation of image data owing to inherent disparities between the 2D skeletal representation obtained via 3D skeleton projection and that estimated by the algorithm. Despite efforts to align these skeletal representations, their accuracy remained affected. In conclusion, this study contributes to the advancement of action recognition technology in the context of Taekwondo poomsae. The model demonstrates robust performance across various viewpoints and datasets, highlighting its potential for real-world applications. Considering factors such as viewpoint selection and data alignment, further investigation and refinement of action recognition models could enhance their accuracy and performance in the field of human motion analysis.

Funding: This study was supported by Konkuk University in 2019.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Konkuk University (7001355-202004-HR-372).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data cannot be provided due to the security reason.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

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

References

1. Choi, C.-H.; Joo, H.-J. Motion recognition technology based remote Taekwondo Poomsae evaluation system. Multimed. Tools Appl. 2016, 75, 13135–13148.
2. Lee, J.; Jung, H. TUHAD: Taekwondo Unit Technique Human Action Dataset with Key Frame-Based CNN Action Recognition. Sensors 2020, 20, 4871.
3. Andó, B.; Baglio, S.; Lombardo, C.O.; Marletta, V. An Event Polarized Paradigm for ADL Detection in AAL Context. IEEE Trans. Instrum. Meas. 2015, 64, 1814–1825.
4. Hsieh, J.; Chuang, C.; Alghyaline, S.; Chiang, H.; Chiang, C. Abnormal Scene Change Detection From a Moving Camera Using Bags of Patches and Spider-Web Map. IEEE Sens. J. 2015, 15, 2866–2881.

5. Cosar, S.; Donatiello, G.; Bogorny, V.; Garate, C.; Alvares, L.O.; Brémond, F. Toward Abnormal Trajectory and Event Detection in Video Surveillance. IEEE Trans. Circuits Syst. Video Technol. 2017, 27, 683–695.
6. Ismail, S.J.; Rahman, M.A.A.; Mazlan, S.A.; Zamzuri, H. Human gesture recognition using a low cost stereo vision in rehab activities. In Proceedings of the 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS); 2015; pp. 220–225. 10.1109/IRIS.2015.7451615.
7. Rafferty, J.; Nugent, C.D.; Liu, J.; Chen, L. From Activity Recognition to Intention Recognition for Assisted Living Within Smart Homes. IEEE Trans. Hum. Mach. Syst. 2017, 47, 368–379.
8. Zolfaghari, S.; Keyvanpour, M.R. SARF: Smart activity recognition framework in Ambient Assisted Living. In Proceedings of the 2016 Federated Conference on IEEE Computer Science and Information Systems (FedCSIS), Gdansk, Poland, 11–14 September 2016; pp. 1435–1443.
9. Zhang, L.; Hsieh, J.-C.; Ting, T.-T.; Huang, Y.-C.; Ho, Y.-C.; Ku, L.-K. A Kinect based golf swing score and grade system using GMM and SVM. In Proceedings of the 5th International Congress on Image and Signal Processing (CISP 2012), Chongqing, China, 16–18 October 2012; pp. 711–715.
10. Zhu, G.; Xu, C.; Huang, Q.; Gao, W.; Xing, L. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 431–440.
11. Martin, P.-E.; Benois-Pineau, J.; Péteri, R.; Morlier, J. Sport Action Recognition with Siamese Spatio-Temporal Cnns: Application to Table Tennis. In Proceedings of the 2018 International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, France, 4–6 September 2018.
12. Wang, S. A Deep Learning Algorithm for Special Action Recognition of Football. Mobile Information Systems 2022, 2022
13. Leo, M., D'Orazio, T., Spagnolo, P., Mazzeo, P.L., Distante, A. (2009). Multi-view Player Action Recognition in Soccer Games. In: Gagalowicz, A., Philips, W. (eds) Computer Vision/Computer Graphics CollaborationTechniques. MIRAGE 2009. Lecture Notes in Computer Science, vol 5496. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01811-4_5
14. Lin, C.-H.; Tsai, M.-Y.; Chou, P.-Y. A Lightweight Fine-Grained Action Recognition Network for Basketball Foul Detection. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW); 2021; pp. 1–2.
15. Ji, R. Research on Basketball Shooting Action Based on Image Feature Extraction and Machine Learning. IEEE Access 2020, 8, 138743–138751.
16. Mora, S.V.; Knottenbelt, W.J. Deep Learning for Domain-Specific Action Recognition in Tennis. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 170–178.
17. Rahmad, N.; As'ari, M. The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognition on vision based data. *J. Phys. Conf. Ser.* **2020**, *1529*, 022021.
18. Rahmad, N.; As'ari, M.; Soeed, K.; Zulkapri, I. Automated badminton smash recognition using convolutional neural network on the vision based data. In Proceedings of the IOP Conference Series: Materials Science and Engineering; IOP Publishing: Putrajaya, Malaysia, 2020; Volume 884, p. 012009.
19. Ijjina, E.P.; Chalavadi, K.M. Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognit. 2017, 72, 504–516.
20. Wang, P.; Wang, S.; Gao, Z.; Hou, Y.; Li, W. Structured Images for RGB-D Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 1005–1014.
21. Trivedi, N.; Kiran, R.S. PSUMNet: Unified Modality Part Streams are All You Need for Efficient Pose-based Action Recognition. arXiv 2022, arXiv:2208.05775.
22. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
23. Xia, H.; Gao, X. Multi-Scale Mixed Dense Graph Convolution Network for Skeleton-Based Action Recognition. IEEE Access 2021, 9, 36475–36484.
24. Gupta, P.; Thatipelli, A.; Aggarwal, A.; Maheshwari, S.; Trivedi, N.; Das, S.; Sarvadevabhatla, R.K. Quo vadis, skeleton action recognition? Int. J. Comput. Vis. 2021, 129, 2097–2112.
25. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 45, 1474–1488.
26. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. arXiv 2021, arXiv:2107.12213.
27. Wang, M.; Ni, B.; Yang, X. Learning Multi-View Interactional Skeleton Graph for Action Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 45, 6940–6954.
28. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1110–1118.

doi:10.20944/preprints202308.1432.v1

29. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv, 2018; arXiv:1801.07455.
30. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
31. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1110–1118.
32. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv, 2018; arXiv:1801.07455.
33. Roetenberg, D.; Luinge, H.; Slycke, P. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV. Tech. Rep. 2009, 1.
34. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6202–6211.
35. Duan, H.; Wang, J.; Chen, K.; Lin, D. Pyskl: Towards good practices for skeleton action recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7351–7354.
36. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans. Image Process. 2020, 29, 9532–9545.

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.