IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 August 2022, accepted 18 September 2022, date of publication 6 October 2022, date of current version 11 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210261

RESEARCH ARTICLE

AI Golf: Golf Swing Analysis Tool for Self-Training

CHEN-CHIEH LIAO[1], DONG-HYUN HWANG[2], AND HIDEKI KOIKE[1]
[1]Department of Computer Science, Tokyo Institute of Technology, Tokyo 152-8550, Japan
[2]NAVER CLOVA Voice&Avatar, Seongnam 13561, South Korea

Corresponding author: Chen-Chieh Liao (liao.c.aa@m.titech.ac.jp)

This work was supported by the Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST), Japan, under Grant JPMJCR17A3.

ABSTRACT In the field of the acquisition of sports skills, a common way to improve sports skills, such as golf swings, is to imitate professional players' motions. However, it is difficult for beginners to specify the keyframes on which they should focus and which part of the body they should correct because of inconsistent timing and lack of knowledge. In this study, a golf swing analysis tool using neural networks is proposed to address this gap. The proposed system compares two motion sequences and specifies keyframes in which significant differences can be observed between the two motions. In addition, the system helps users intuitively understand the differences between themselves and professional players by using interpretable clues. The main challenge of this study is to target the fine-grained differences between users and professionals that can be used for self-training. Moreover, the significance of the proposed approach is the use of an unsupervised learning method without prior knowledge and labeled data, which will benefit future applications and research in other sports and skill training processes. In our approach, neural networks are first used to create a motion synchronizer to align motions with different phases and timing. Next, a motion discrepancy detector is implemented to find fine-grained differences between motions in latent spaces that are learned by the networks. Furthermore, we consider that learning intermediate motions may be feasible for beginners because, in this way, they can gradually change their pose to match the ideal form. Therefore, based on the synchronization and discrepancy detection results, we utilize a decoder to restore the intermediate human poses between two motions from the latent space. Finally, we suggest possible applications for analyzing and visualizing the discrepancy between the two input motions and interacting with the users. With the proposed application, users can easily understand the differences between their motions and those of various experts during self-training and learn how to improve their motions.

INDEX TERMS Computer vision, machine learning, motor skill training, golf.

I. INTRODUCTION

In sports, it is difficult for beginners to improve their skills without prior knowledge or assistance from coaches. As a conventional method, people go to lessons to meet experts and learn how to play in the proper form. However, in most sports, players spend considerable time training alone to achieve outstanding results in the field and retain exceptional body conditions. Therefore, it is important to design and implement an effective and accurate self-training process for such situations.

In the field of the acquisition of sports skills, one way to improve sports skills is to replicate professional players' motions. People watch the movements of professional athletes on television or the Internet and try to make their bodies move similarly to professionals. To accelerate this process, many systems have recently been developed to help users understand the movements of professionals [1], [2], [3], [4], [5]. However, in these previous works and systems, users may struggle to refine their movement with no idea which timing of the whole motion, which parts of the body they should focus on, or how they can change their body movements to get their form closer to that of professionals.

With significant advances in machine learning technologies, many systems have been built to recognize different objects, make predictions for decisions, or even predict the future [6]. Researchers have focused on producing self-training systems with neural networks [7]. A recent study [8] introduced a climbing training system in which users can receive recommended poses and movements to

The associate editor coordinating the review of this manuscript and approving it for publication was Jinjia Zhou.

be learned in the future, based on their current poses, after training the network with professional poses. The advantage of incorporating machine learning is that the system can learn without prior knowledge. Therefore, no expert or coach is required to build such a system, and it can be generalized to many other tasks. In addition, apart from image information, multiple sensors have recently been fused to measure human motion, and deep learning technologies can efficiently take advantage of processing such fused data [9]. However, supervised learning requires abundant data from a specific domain, which makes it difficult to design a generalized sensor fusion application.

In this study, we address these gaps in the literature to provide users with interpretable clues so that they can intuitively understand the difference between themselves and professional players. We choose golf as our network learning subject because golf is an individual sport, and the player's standing position is fixed when performing a swing. Recent studies [10], [11] have demonstrated the ability of deep learning to retrieve fine-grained information necessary for golf swing analysis. However, after retrieving the essential factors from human motion, the system designer must know which part of the body is vital in golf and determine which parts must be processed for analysis. Furthermore, the calculation is manually designed for golf swings; thus, it is difficult to generalize the proposed methods to other sports analyses without prior knowledge or the help of experts. In this study, we propose a golf swing analysis tool using deep neural networks to help users recognize the difference between the user's swing motion and an expert's motion obtained from the Internet or broadcasted media. Furthermore, we address the previous issues in an unsupervised manner to encourage the network to learn standard features from professional players without adding domain-specific information. Consequently, the proposed network can be applied to other sports and skill-training processes.

The proposed system consists of three modules: a motion synchronizer, motion discrepancy detector, and motion manipulator. The motion synchronizer matches the two input motions with different timings and speeds. The motion discrepancy detector can recognize the difference between the two motions and find the frame in which the difference is large. The motion manipulator is designed to produce intermediate motions between the two motions to provide more intuitive instructions for users to learn. To evaluate the accuracy and effectiveness of the proposed modules, we collect golf swing data from existing databases and generate a pseudo database containing raw video data, 3D pose data, and labels for the phases during the swing. Next, we examine the accuracy of the motion synchronizer and the capability of motion discrepancy using three types of inputs (raw video, video without background, and 3D human pose). Finally, we discuss the accuracy of the results and explain the correlation among the learned latent space, human motion, and other features. On the other hand, qualitative results are shown to evaluate the ability of the motion manipulator to reproduce motions from the latent space and create new motions unseen in the database.

Furthermore, we discuss the proposed analytical tool and its possible applications in future research. The proposed application visualizes the image frames and human motions where the discrepancy between the expert and the user's swing is large and helps the user quickly recognize the motion that needs to be corrected. Compared to existing methods, the main contributions of this study can be summarized as follows:

- A golf swing analysis method and its applications are introduced.
- The proposed method distinguishes when the difference between two motions is large and small.
- The proposed method helps users understand the difference between themselves and professional players.
- The proposed method provides intermediate poses that are acceptable during the early learning phase of sports.
- Crucial factors that can influence the accuracy of sports analysis are discussed.

II. RELATED WORK

A. SPORTS TRAINING FOR SKILL ACQUISITION

Many recent studies have focused on developing sports training systems to help beginners improve their skills. Studies [13], [14], [15] have proposed multi-modal sports training systems based on sports theories. In their system, users received visual, haptic, and audio feedback when they did not ideally move their bodies or instruments. However, the ideal movement could differ from sport to sport, meaning that it may be difficult to generalize these methodologies to other sports.

On the other hand, another way to learn sports skills is imitating professional players' motions [1], [3], [4], [5]. Ikeda *et al.* [16] proposed a golf swing training system that uses the motions of professional golfers. In their system, a user's motion was synchronized with a selected ideal professional's motion, and the two motions were overlaid and projected onto the ground during training. Sasaki *et al.* [17] also reported the importance of beginners copying expert motions and proposed a climbing training system using pose prediction. Their system predicted and visualized the pose of experts based on the user's current hand and foot positions. While these recent studies have shown the effectiveness of using multi-modal feedback and the potential of applying neural networks to create AI teachers for sports training, it is difficult for users to change their motion forms immediately to match ideal forms. Thus, building a system that can teach users step-by-step to improve their sports skills remains challenging.

B. VIDEO AND MOTION ALIGNMENT

An efficient way to evaluate whether a person is performing a motion correctly is to compare their motion with others whose motion is considered correct. However, owing to the various timings and speeds of motion of different

individuals, we must align the motions to make them comparable. A conventional method for aligning two temporal sequences is dynamic time warping (DTW), which was introduced by Berndt and Clifford [18]. In this method, every index in a sequence is matched with one or more indices from another sequence, and the mapping of the indices from the first sequence to the other sequences must be monotonically increasing. The DTW concept has been introduced in several domains. For example, Ikeda *et al.* [19] proposed a real-time golf swing projecting system that simultaneously visualized professional and user forms with matched timing. To align the two motions in real-time, they measured the DTW only over a short period and penalized the previous cost value at a later time. Halperin *et al.* [20] utilized the concept of DTW in speech and presented an audio-to-video alignment method for matching speech-to-lip movements.

On the other hand, self-supervised neural networks have recently been developed to tackle video alignment tasks using the latent space representation [12], [21], [22]. In this approach, an embedder is used to compress input videos into a latent space. After the embedding process, a loss is designed to find correspondences through time in the latent space, thus encouraging the network to learn a latent space where similar motions should appear to be close. This method helps synchronize high complexity temporal sequences, such as videos. In this study, we base the loss of our network on the temporal cycle-consistency loss used by Dwibedi *et al.* [12], but apply DTW along with the loss for smooth temporal matching.

C. DISCREPANCY DETECTION

By comparing the two synchronized motions, we can determine the difference between them in terms of human postures. In early studies, abnormal detection referred mainly to finding patterns in data that did not match the expected behavior [25]. Recently, two methods have been proposed for detecting abnormalities In the first approach, abnormal information is referred to as prior knowledge. For example, Parra-Dominguez *et al.* [26] trained a binary classifier on annotated data to determine whether abnormal events occurred during a stair descent. In the second approach to abnormal detection, abnormal information is not provided in advance. The research group of Nater *et al.* [27] proposed an unsupervised learning method for learning normal human behavior. They used a hierarchical representation of the appearance and action level of regular movements to detect abnormal events.

While a network may be trained to detect abnormal events, such as falling to the ground, we focus on whether a neural network can be trained to automatically detect fine-grained differences between two regular motions. We call this detection of the fine-grained difference discrepancy detection. The most relevant of these studies is that of Abati *et al.* [28]. They designed a deep autoencoder with a parametric estimator that learned a probability distribution from the latent space to detect discrepancies. The encoder effectively remembered standard samples and could distinguish between normal and abnormal images. However, although the network could easily detect surprise samples, fine-grained differences among standard samples were not discussed. In addition to image-based methods, recent studies have focused on systems that use 3D human pose information [29], [30]. In this study, we apply discrepancy detection to both videos and 3D human poses and discuss the ability of the system to detect fine-grained differences between two input motions.

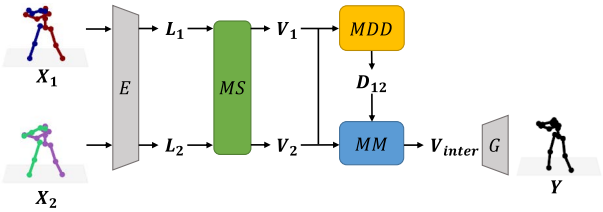

FIGURE 1. System overview. X is the input motion sequence and Y is the output human poses restored from the latent space.

III. METHODS

This study aims to create a system that captures user motions and provides fine-grained feedback to improve users' forms by comparing their motions with those of professionals. To achieve the goal of building such an application, the method proposed in this study is to first train a neural network with professionals' motion data. After training the network, the system compresses the user motions through the network into a latent space and compares their motions with those of professionals in the latent space. Figure 1 shows an overview of the proposed system. The workflow of the approach is divided into three parts: motion synchronization, motion discrepancy detection, and motion manipulation. The system first receives two motion inputs $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ and uses encoder $\boldsymbol{E}$ to embed the input motions into the latent space, where the two motions are represented as $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. The encoder is trained to learn a latent space in which similar motions appear to be close.

Next, using the learned latent space, the motion synchronizer $\boldsymbol{MS}$ matches the timing of the two motions in the latent space by measuring the Euclidean distance between $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. The motion discrepancy detector then captures the two synchronized latent vectors $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$, measures the difference between them, and passes a distance vector $\boldsymbol{D_{12}}$ to the motion manipulator $\boldsymbol{MM}$.

Finally, the $\boldsymbol{MM}$ integrates $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, and $\boldsymbol{D_{12}}$ to specify the key frames where large differences occur and create an intermediate latent vector $\boldsymbol{V_{inter}}$. The system uses decoder $\boldsymbol{D}$ to restore the intermediate human poses $\boldsymbol{Y}$ from $\boldsymbol{V_{inter}}$ for users to gradually improve their motion forms.

A. MOTION SYNCHRONIZER: ALIGNING MOTION SEQUENCES WITH DIFFERENT TIMING

For motion synchronization, we aim to design a network that learns a latent space that shows motion similarity. A common way to achieve this is by constructing an autoencoder and decoder, whose input and output are the same motions. On the other hand, previous studies have

shown that cycle-consistency methods are useful for aligning video inputs with different phasing and timing. Our method is inspired by the temporal cycle consistency (TCC) learning method proposed by Dwibedi *et al.* [12]. The TCC network is designed to allow the network to learn not only the similarity of motion, but also the temporal order of the entire motion. Inspired by this previous study, we implement the TCC algorithm as follows:

- The encoder $\boldsymbol{E}$ compresses two motion sequences to latent vectors $\boldsymbol{L_1}, \boldsymbol{L_2}$.
- For each node of $\boldsymbol{L_1}$, find the nearest node of $\boldsymbol{L_2}$.
- For each identified node of $\boldsymbol{L_2}$, we find the nearest node of $\boldsymbol{L_1}$ (cycle back).
- If the node is cycling back to itself, there is no loss; otherwise, TCC loss is calculated.

For the encoder network, we implement a video-based network and two skeleton-based networks. Each network consists of a base network and an embedder network. While the base network is designed to extract features from a given video or skeleton sequence, the embedder network uses the output of the base network and embeds it into the latent space.

For the embedder network, we first store the features of any given frame together with its context frames along the time dimension. Next, we apply 3D convolutions to aggregate the temporal information and reduce the dimensionality using 3D max-pooling. Finally, we use two fully connected layers and a linear projection to obtain a 128-dimensional embedding for each frame.

Three types of networks are implemented in the base networks: video TCC (V-TCC), skeleton TCC (S-TCC), and skeleton-attention TCC (SA-TCC). V-TCC is a network that uses videos as its input. The original TCC implementation is followed to construct the V-TCC. We use the ResNet-50 [31] architecture pre-trained with ImageNet [32] to extract features from the output of the Conv4c layer. All frames in a given video sequence are resized to 224 × 224, and the extracted convolutional features are 14 × 14 × 1024. The convolutional features produced are then fed into the embedder network.

S-TCC is a straightforward implementation of our baseline method that uses skeletons (human poses) as its input. The S-TCC consists of only fully-connected layers. All frames in a given skeleton sequence have a size of 3 × 16 (joints), and we expand the skeleton input to a single 1 × 48 vector and feed it to the embedder network.

Because the plain implementation of S-TCC may not be able to learn the relationship among the 3D joints, SA-TCC is another skeleton input version of our TCC network that uses the concept of the self-attention mechanism. As shown in Figure 2, in the SA-TCC network, we first expand the 3 × 16 skeleton input to a single 1 × 48 vector $\boldsymbol{x}$ and then turn it to query $\boldsymbol{q(x)}$, key $\boldsymbol{k(x)}$, and value $\boldsymbol{v(x)}$. The output $\boldsymbol{y}$ of the attention block is fed to the embedder network, and the transformation in the attention block is formulated as follows:

$$\beta_{ij} = \frac{exp(s_{ij})}{\sum_i exp(s_{ij})}, \quad \text{where } \boldsymbol{s} = \boldsymbol{q(x)}^T \boldsymbol{k(x)} \tag{1}$$

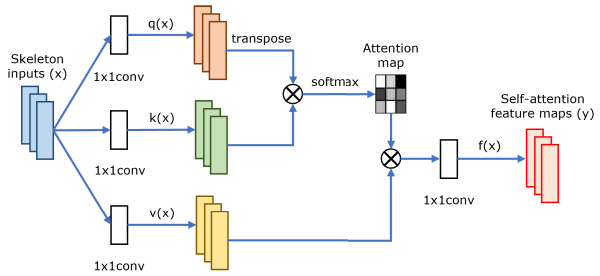

FIGURE 2. Self-attention block. x is the skeleton input, and y is the output. $q(x)$, $k(x)$, and $v(x)$ is the production of the query, key, value respectively. ⊗ is the matrix multiplication.

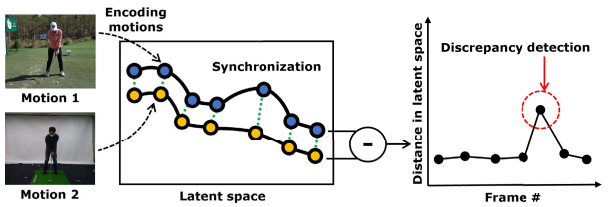

FIGURE 3. Discrepancy detection. The proposed network is encouraged to find a latent space where similar motions appear to be close. After synchronization, frames with large distances in the latent space are considered keyframes where large motion differences occur.

$$\boldsymbol{y_i} = \boldsymbol{f}(\sum_j \beta_{ij} \boldsymbol{v(x)_j}) \tag{2}$$

In the above transformation, the weights to be learned for $\boldsymbol{q(x)}$, $\boldsymbol{k(x)}$, and $\boldsymbol{v(x)}$ are implemented as 1 × 1 convolutions. We aim to enable the network to recognize the relationships between different skeleton joints by learning the attention matrix inside the attention block.

B. MOTION DISCREPANCY DETECTOR: FINDING FINE-GRAINED MOTION DIFFERENCES

After training the network, similar motions must be close together in the latent space. In this section, we focus on the distance between two motions in the latent space to detect and retrieve fine-grained discrepancies and compare two different swing forms, particularly for the differences between beginners and experts. As previously mentioned, the TCC network synchronizes the input sequences by calculating the Euclidean distance between latent vectors. At this point, similar motions appear close to each other in the latent space. As shown in Figure 3, because we assume that the network is trained using the golf swings of advanced golfers, a small difference is computed when the input motion is performed similarly to advanced golfers. On the other hand, if the poses between the two input motions are dissimilar in a specific frame, causing a large distance between the aligned latent vectors, we may find a significant motion difference in that frame. Therefore, we take the latent vectors of the input motions $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$, which are timing-matched by the motion synchronizer, and calculate the frame-by-frame distance vector $\boldsymbol{D}$, which indicates the degree of difference between the motions:

$$\boldsymbol{D_i} = \sum_i ||\boldsymbol{V_{1i}} - \boldsymbol{V_{2i}}||^2 \tag{3}$$

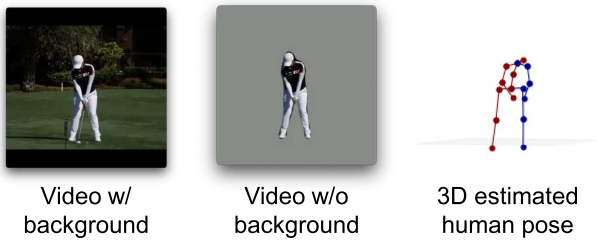

FIGURE 4. Three types of datasets.

C. MOTION MANIPULATOR: DISCOVERING INTERMEDIATE MOTION BETWEEN HUMAN POSES

During self-training, it is not always simple for beginners to imitate an ideal motion form, which is very different from their current form. In this study, we propose a motion decoder to generate an intermediate motion. As mentioned previously, the TCC network is trained to learn a latent space that shows motion similarity. Therefore, we understand that a high-dimensional data point in latent space can be representative of a human pose. To retrieve the intermediate motion between two points in the latent space, we train a decoder using latent vectors to predict human poses that are the same as the inputs. In particular, we first input the training set data into the trained TCC network to obtain the outputs of the latent vectors. Next, using the latent vectors as inputs, we trained a simple decoder consisting of a single fully-connected layer to produce the outputs of the human poses. For the loss function L_{MSE}, we take the mean square error (MSE) between the output and input poses:

$$L_{MSE} = \sum_{i} ||\boldsymbol{Y_i} - \boldsymbol{X_i}||^2 \tag{4}$$

where $\boldsymbol{X_i}$ is the i^{th} joint of the input human pose $\boldsymbol{X}$ and $\boldsymbol{Y_i}$ is the i^{th} joint of the output human pose $\boldsymbol{Y}$.

After training the motion decoder, we retrieve a new latent vector $\boldsymbol{V_{inter}}$ between the two timing-matched latent vectors using linear interpolation:

$$\boldsymbol{V_{inter}} = (1 - \alpha) \times \boldsymbol{V_1} + \alpha \times \boldsymbol{V_2} \tag{5}$$

where $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ are the two latent vectors synchronized by the motion synchronizer, and $\alpha \in 0.0, 1.0$ is the magnitude parameter. In the above formulation, by increasing the value of α, we can obtain a human pose whose latent vector is closer to $\boldsymbol{V_2}$, and the restored human pose should ideally be more similar to the human poses generated from $\boldsymbol{V_2}$.

IV. EXPERIMENTAL SETUP

To evaluate the accuracy and effectiveness of the three modules introduced in the previous section, we collected golf swing data via the Internet and created a pseudo database consisting of raw video data, without-background-video data, and 3D pose data (Figure 4). Next, we implemented three models of the network (V-TCC, S-TCC, and SA-TCC) utilizing TCC loss and conducted statistical analysis under four different conditions:

- V-TCC using video inputs with backgrounds
- V-TCC using video inputs without backgrounds
- S-TCC using 3D human pose inputs
- SA-TCC using 3D human pose inputs

The following sections introduce the data collection process and the evaluation metrics used in this study.

FIGURE 5. Key event and phase. The impact moment and the top moment are labeled as key events. Frames between them are labeled as swinging down phases.

A. VIDEO DATASET

As the training dataset, we used GolfDB [33], a video dataset collection for all types of golf iron swings and driver swings, consisting of 1400 high-quality golf swing videos of male and female professional golfers. Although the GolfDB provided preprocessed video clips for a frame size of 160 × 160, we rebuilt our pseudo dataset with high-resolution videos. In addition to clean videos, we created another video dataset without background information. This was because, in our hypothesis, the background information, for example, the human shadow, might influence the alignment of the network. Moreover, the network might also learn the motion of props, such as golf clubs. We used Mask R-CNN [34], a generic object detection and segmentation network, to detect the human body in a single image frame. We then removed the background pixels and left only the human body (Figure 4).

B. 3D POSE DATASET

To conduct a more precise analysis of only human poses, we created a new pseudo dataset consisting of 3D point data of human body poses. In this dataset, we first used HRNet [35] to retrieve the time series of 2D human poses from golf-swing videos. While the time series of 2D poses could roughly represent human motion, 2D poses could vary significantly owing to camera poses, and it was difficult to address the normalization problem in 2D space. Therefore, human 3D poses were produced using the simple linear network structure proposed in [36]. The estimated 2D poses from the HRNet were fed to a linear network to retrieve the 3D human poses (Figure 4).

C. EVALUATION METRICS

Because we used a self-supervised learning method, we trained the network until the TCC loss converged. To evaluate how well the network was trained, we applied an accuracy metric showing the precision of the alignment using two label types: key events and phases. A key event is a single

frame showing a particular moment, and the phase is a time series between two key events. For example, as shown in Figure 5, a key event in golf may be the moment when the golf club hits the ball (impact), and the motion before the golf club hits the ball can be considered as the phase of the golf club approaching the ball (swinging down). Note that all the frames in the period between two key events have the same phase label. Following the key event annotation of GolfDB, we labeled our pseudo dataset with eight key events: address, toe-up, mid-backswing, top, mid-downswing, impact, mid-follow-through, and finish. Phases were labeled between every two key events, for a total of seven phases.

The phase classification accuracy was per frame phase classification. To calculate the accuracy, we first used the encoders of our TCC networks to extract latent vectors. We then trained a simple classifier on the latent vectors to predict the labeled phases. The classifier was trained under several conditions by changing the percentage of the given labeled data. After the classifier was trained, we used all labeled data to calculate the phase classification accuracy. In general, the larger the size of the given labeled data, the higher the accuracy of the classifier.

To explore more specifically the effectiveness of the network in detecting discrepant motion, we investigated whether the distance in the latent space could represent the discrepancy between two input sequences by computing Pearson's correlation coefficient. Because 3D poses varied for different camera views, we could not directly compare the two human poses using the results from the 3d pose estimator. Thus, we applied Procrustes analysis to align the two human poses to ensure that they were seen from the same camera direction. In our experiment, we compared the distance in the latent space with the mean per joint point error (MPJPE) and measured the Pearson's correlation coefficient ρ:

$$\rho_{D,E} = \frac{cov(D, E)}{\sigma_D \sigma_E} \tag{6}$$

where cov is the covariance; D is the distance in the latent space; E is the MPJPE; and σ_D and σ_E are the standard deviations of D and E.

V. RESULTS

This chapter presents the results of an early qualitative analysis investigating the potential of the proposed method to detect discrepant motion differences, followed by more detailed results comparing different modules. Finally, the intermediate human poses generated by the motion manipulator are visualized for qualitative studies.

A. CASE STUDY

In our early study, we conducted a qualitative analysis by exploring the latent space to investigate whether the network could trace fine-grained differences. We first used the V-TCC to synchronize the swing motions of professionals and beginners. We then computed the distances between the aligned videos in the latent space and visualized the overlaid 3D human poses for qualitative comparison (Figure 6).

TABLE 1. Phase classification accuracy. This is the accuracy metric showing the ability of the network to classify any given motion frame to its corresponding phase.

Labeled data (%)	5%	10%	30%	80%
V-TCC (with background)	0.718	0.724	0.796	0.840
V-TCC (without background)	0.859	0.839	0.893	0.917
S-TCC	0.895	0.901	0.916	0.929
SA-TCC	0.881	0.902	0.913	0.918

B. PHASE CLASSIFICATION ACCURACY

We trained the V-TCC using the GolfDB video dataset with and without background subtraction. In contrast, S-TCC and SA-TCC were trained using the pseudo skeleton dataset with unit vector normalization. After training the four models, we computed the phase classification accuracy for each trained model by assigning 1%, 5%, 10%, 30%, and 80% of labeled data.

The results are presented in Table 1. As we trained the network properly, the phase classification accuracy was low when the given labeled data were insufficient and rose with an increase in the number of labeled data.

C. CORRELATION

We computed Pearson's correlation coefficient using the four models. For video inputs, a 0.69 Pearson's correlation coefficient was measured for regular video, and a 0.72 Pearson's correlation coefficient was obtained when the background was removed (Figure 7). For skeleton inputs, an over 0.76 Pearson's correlation coefficient was found when using the SA-TCC; however, the lowest Pearson's correlation coefficient with 0.51 was obtained from the S-TCC (Figure 8).

D. MOTION INTERPOLATION

For the qualitative results, we computed and visualized the intermediate human pose between a pair of human poses considering the following three circumstances:

- The two poses were from a single person. The two poses were in different phases (Figure 9 (a)).
- The two poses were from different individuals. The two poses were in the same phase (Figure 9 (b)).
- The two poses were from different individuals. The two poses were in different phases (Figure 9 (c)).

VI. DISCUSSION

A. CASE STUDY

In our early case study, we observed that in most cases, when the distance in the latent space was small, the difference between the two 3D poses was small(Figure 6 c). On the other hand, when the distance was considerable, the 3D poses showed more differences(Figure 6 b, d). This suggested that the network could distinguish between beginners and professionals. Thus, we can use this feature to retrieve motion features that help users target the key moves they have to correct. However, as depicted in Figure 6 (a), the distance in the latent space remained large even in certain circumstances where the

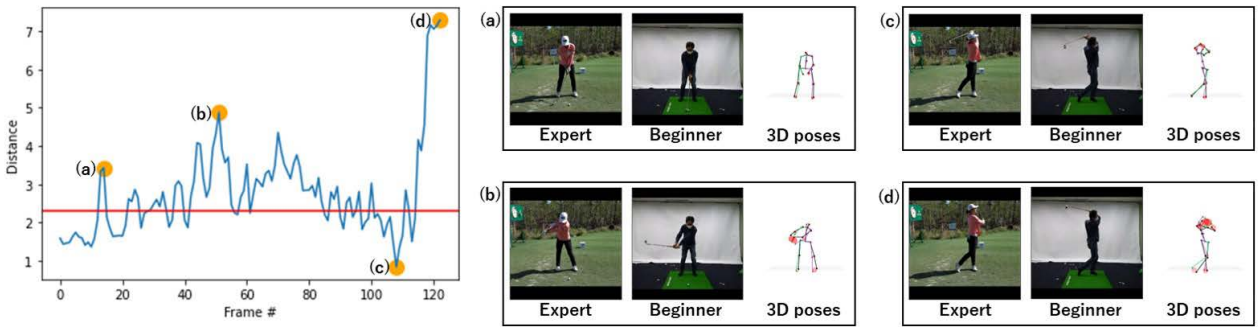

FIGURE 6. Case study with the V-TCC. The line graph shows the distance between two synchronized motions in the latent space. The red line in the graph indicates the threshold for discrepancy detection. The colored skeleton and black skeleton indicate the user's pose and expert's pose, respectively. The density and radius of red spheres indicate the degree of joint position difference between the two skeletons.

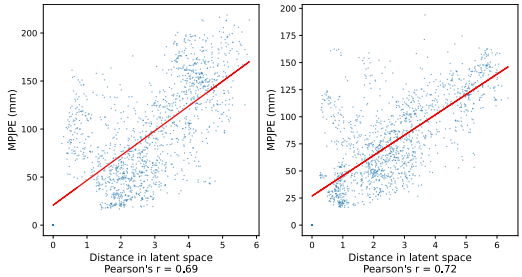

FIGURE 7. Pearson's correlation test for V-TCC. Left: normal videos. Right: videos without background.

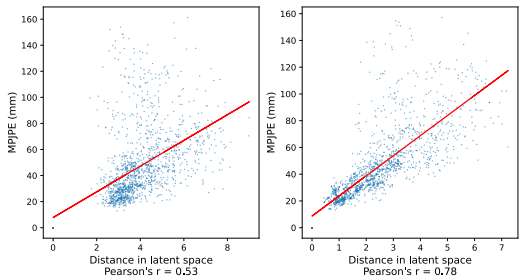

FIGURE 8. Pearson's correlation test with skeleton input. Left: S-TCC. Right: SA-TCC.

joint difference was negligible. This might be explained by the fact that the network focused not only on human motion but also on other motion features. In this case, the golf club's movement, which was also critical during the swing, might be the main factor causing the enormous distance in the latent space. This led us to the following quantitative studies, where we conducted statistical tests to examine the accuracy of the synchronization and the correlation between the latent space and joint difference under various conditions.

B. PHASE CLASSIFICATION ACCURACY

In this quantitative study, we discovered that the previous TCC implementation had limited precision in terms of synchronization. As our hypothesis suggests, synchronization quality might be influenced by various background information and other movements, such as the golf club's and human shadow's motions. When the background was removed, a significant increase in the accuracy of the V-TCC model was observed. This result supported our hypothesis that information outside the human region would affect alignment. Various types of information from the background, the movement of human shadows, and the motion of properties controlled by a human might be learned as features by the network, thus affecting the accuracy of the analysis. Because of the significant superiority of the without-background version, we considered it necessary to apply background subtraction before synchronization when building applications for real usage.

Next, considering the skeleton version, both S-TCC and SA-TCC outperformed V-TCC under all conditions. Given only 10% of the labeled data, both S-TCC and SA-TCC achieved a phase classification accuracy of over 90%. This might be explained by the fact that the skeleton version had more compressed and precise information than the video version, where the color information might be noisy. Additionally, 3D poses represented high-level human features. Therefore, the features outside the human region were removed when retrieving 3D human poses, resulting in more precise accuracy for further golf swing analysis. Note that the skeleton data were the estimated results from the video data, and the estimation could not be perfect for every human pose. Despite imperfect skeleton inputs, the S-TCC and SA-TCC networks outperformed V-TCC. This result suggested that we could use the skeleton version for more precise implementation in actual usage.

Finally, comparing the two skeleton version models, S-TCC performed better than SA-TCC under most conditions. However, SA-TCC remained competitive in some circumstances (better than S-TCC with 10% labeled data). This led to the following correlation test, in which we discussed the ability of the networks to detect discrepancies between two motions.

C. CORRELATION

Discussing the correlation test, as shown in Figure 7, we observed a correlation between the distance in the latent space and the MPJPE. While the distribution of the data pairs seemed to be scattered, a Pearson's correlation coefficient of over 0.69 showed the potential of the network to distinguish whether the difference between the two motions was small or large. However, as discussed in the case study section, in some circumstances, the network failed to detect discrepant

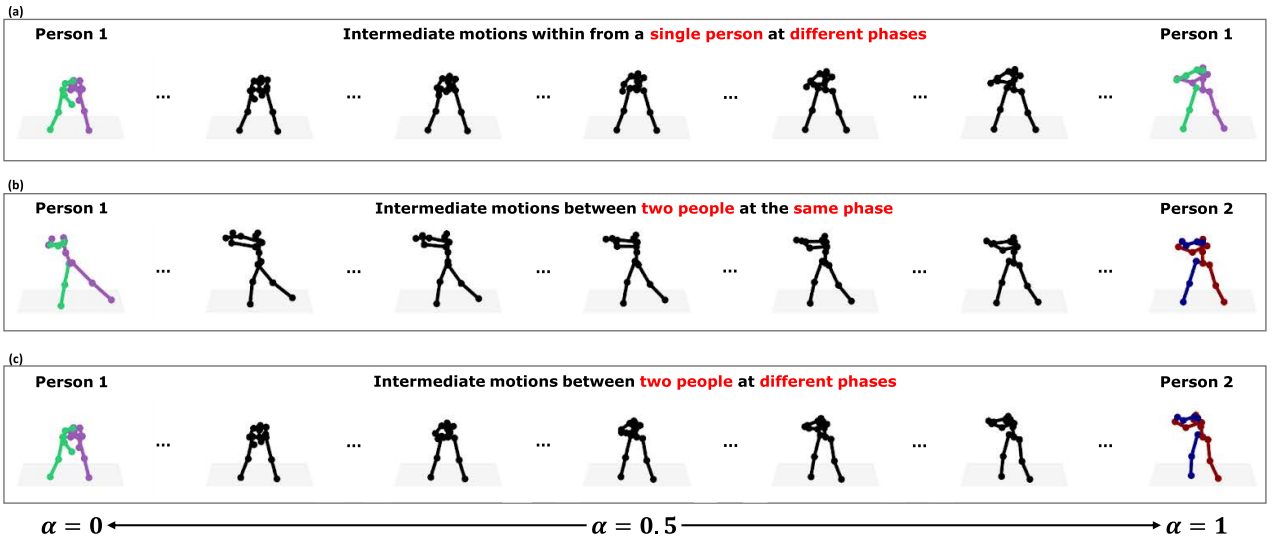

FIGURE 9. Motion manipulation. Unseen intermediate motions between two different 3D human poses are retrieved from the latent space using linear interpolation. α is the blending value. The same color of the skeleton denotes the same person.

motion showing no difference in the latent space, while the joint difference was significant. As shown in Figure 7, a greater correlation and a more clustered distribution were found when the background and golf club were removed from the scene. This indicated that the network could learn the motion features of the club, and that it is essential to analyze the club motion when showing the difference between the two input motions.

For the two skeleton models, an over 0.76 Pearson's correlation coefficient was found when using the SA-TCC. On the other hand, only a 0.51 Pearson's correlation coefficient could be obtained from the S-TCC. Therefore, along with the results shown in Table 1, we found that although the S-TCC network had the most remarkable performance in tackling synchronization, its limited ability to detect discrepant motions might be a drawback in terms of real application implementations. In contrast, SA-TCC had a superior Pearson's correlation coefficient over all other competitors while maintaining an acceptable phase classification accuracy. Overall, the skeleton model with the attention module implementation outperformed the video models in all aspects of evaluation. Therefore, based on the results, we considered the attention module the best choice for actual usage. We then used SA-TCC as our final version of the TCC network for the decoder and application implementation, which will be discussed in the next section.

D. MOTION INTERPOLATION

From the results shown in Figure 9 (a), we observed that the trained decoder could restore the ground truth of human poses. Furthermore, the motion decoder demonstrated its ability to generate new human poses unseen in the training dataset. This suggested that the TCC network could learn fine-grained features of the poses of a single person. Notably, the intermediate motion exhibited a continuous change from one pose to another. From this, We could infer that the human

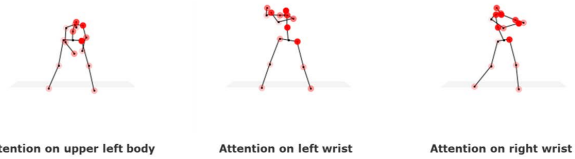

FIGURE 10. Visualization of body parts for revision suggestions. The attention-based network focuses on different body parts at three different phases (address, top, follow-through from left to right). The density and radius of red spheres show the intensity of the attention of the network.

poses of a motion sequence were arranged in orders in the high-dimensional latent space.

Next, as depicted in Figure 9 (c), we found that the trained decoder could perform fair interpolation between different poses. Specifically, intermediate poses had the characteristics of two different input human poses. For example, we observed continuous changes in the distance between limbs from one human pose to another. This suggested that similar poses which pass through the TCC network were embedded in nearby points in the latent space regardless of the poses from different individuals.

Finally, as shown in Figure 9 (b), the motion decoder could generate meaningful intermediate poses at an aligned time. We observed that the human pose gradually changed from an abnormal to a standard form. This suggested that instead of teaching beginners directly with professionals' swing forms, we could provide them with a preliminary pose that is more acceptable for beginners to imitate.

VII. APPLICATIONS

The proposed method can be used for self-training systems that detect discrepant motion frames using the distance between golf swings in the latent space and compare 3D human poses at the detected frames. As an application prototype, we combine the motion synchronizer, motion discrepancy detector, and motion manipulator into a single graphical

user interface where users can select any professional's form from the database and compare the difference between their forms and the professional's. In addition, instead of directly imitating the selected motion, users can gradually imitate intermediate human poses using a motion manipulator. Furthermore, we visualized the attention maps along with the self-attention module to demonstrate the importance of each body part (Figure 10). This may be a clue for discovering body parts that should be focused on and what can be ignored. To go a step further, this might indicate the importance of the body parts that trainees should follow to revise their poses optimally. In the future, we plan to conduct user studies to evaluate the training effectiveness of the proposed system and determine whether the crucial points shown by the attention maps can be used in real training scenarios. Moreover, we consider that using a high-quality motion capture system can enhance the solidity of the proposed system with high-precision human poses despite the high speed of motion, such as golf swings.

As this work focuses on proposing a novel flow for constructing a sports analysis tool, and because user studies may vary from domain to domain and are flexible for the system designer, further user studies evaluating the proposed system's efficiency have not been addressed in practice. As previously mentioned, there are many ways to design a proper way to provide feedback to users for training. Related works that use multi-modal feedback to alert users when performing the wrong way compared to professionals have shown their significance during training. In the future, we plan to combine our approach with other feedback systems and evaluate the effectiveness of the system. Furthermore, we plan to apply the proposed approach to other sports and skill training processes to explore the generality of the proposed method.

VIII. CONCLUSION

We propose a golf swing analysis tool that uses neural networks to help users intuitively understand the difference between themselves and professional players. We divide our work into three parts: synchronization, discrepancy detection, and manipulation. First, the motion synchronizer aligns motions with different phases and timings. The experiment shows that our implementation using skeleton inputs can achieve a better performance than state-of-the-art video implementations.

Second, using the proposed networks, we use a motion discrepancy detector to find fine-grained differences between golf swings in the latent space. By applying comparative analysis, such as comparing 3D human poses in those detected frames, we conclude that the motion discrepancy detector can distinguish whether the difference between the two motions is small or large. The proposed SA-TCC network outperforms the previous TCC network in terms of phase classification accuracy and has the best ability to show the correlation between the distance in the latent space and the MPJPE.

Third, based on the synchronization and discrepancy detection results, we introduce a decoder structure called a motion manipulator to restore motion from the latent space. The restoration results suggest that the network can retrieve intermediate human poses between two motions that do not exist in the original dataset. Furthermore, instead of teaching beginners directly about professional forms, the motion manipulator can provide them with an intermediate pose that is more acceptable for beginners to start with.

Finally, with the above three main contributions of this work, we create an application for analyzing and visualizing the discrepancy between two input golf swing motions. For user interaction, users can quickly grasp the difference between their swings and those of various experts during self-training. In addition, by understanding the continuous changes step-by-step between the two selected human poses, we aim to help users efficiently learn an ideal form in a gradual manner instead of directly imitating ideal motion. Using the proposed system, users can choose an ideal form to imitate and learn to play sports without the help of coaches during self-training. We intend to refine the proposed prototype application and conduct user studies to investigate its effectiveness.

REFERENCES

[1] T. N. Hoang, M. Reinoso, F. Vetere, and E. Tanin, "Onebody: Remote posture guidance system using first person view in virtual environment," in *Proc. 9th Nordic Conf. Hum.-Comput. Interact.*, New York, NY, USA, Oct. 2016, pp. 1–10, doi: 10.1145/2971485.2971521.

[2] C.-C. Liao, D.-H. Hwang, and H. Koike, "How can i swing like pro?: Golf swing analysis tool for self training," in *Proc. SIGGRAPH Asia Posters*, New York, NY, USA, Dec. 2021, pp. 1–2, doi: 10.1145/3476124.3488645.

[3] P.-H. Han, Y.-S. Chen, Y. Zhong, H.-L. Wang, and Y.-P. Hung, "My Tai-Chi coaches: An augmented-learning tool for practicing Tai-Chi Chuan," in *Proc. 8th Augmented Hum. Int. Conf.*, New York, NY, USA, Mar. 2017, pp. 1–4, doi: 10.1145/3041164.3041194.

[4] I. Kuramoto, Y. Nishimura, K. Yamamoto, Y. Shibuya, and Y. Tsujino, "Visualizing velocity and acceleration on augmented practice mirror self-learning support system of physical motion," in *Proc. 2nd IIAI Int. Conf. Adv. Appl. Informat.*, Aug. 2013, pp. 365–368, doi: 10.1109/IIAI-AAI.2013.28.

[5] M. Kyan, G. Sun, H. Li, L. Zhong, P. Muneesawang, N. Dong, B. Elder, and L. Guan, "An approach to ballet dance training through MS Kinect and visualization in a CAVE virtual reality environment," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, pp. 1–37, May 2015, doi: 10.1145/2735951.

[6] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, Jun. 2019, doi: 10.1109/JSAC.2019.2904363.

[7] E. Wu and H. Koike, "FuturePose—Mixed reality martial arts training using real-time 3D human pose forecasting with a RGB camera," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1384–1392, doi: 10.1109/WACV.2019.00152.

[8] K. Shiro, K. Egawa, T. Miyaki, and J. Rekimoto, "InterPoser: Visualizing interpolated movements for bouldering training," in *Proc. IEEE Conf. Virtual Reality 3D User Inter. (VR)*, Mar. 2019, pp. 1563–1565, doi: 10.1109/VR.2019.8798366.

[9] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Inf. Fusion*, vol. 80, pp. 241–265, Apr. 2022, doi: 10.1016/j.inffus.2021.11.006.

[10] T. T. Kim, M. A. Zohdy, and M. P. Barker, "Applying pose estimation to predict amateur golf swing performance using edge processing," *IEEE Access*, vol. 8, pp. 143769–143776, 2020, doi: 10.1109/ACCESS.2020.3014186.

[11] K.-R. Ko and S. B. Pan, "CNN and bi-LSTM based 3D golf swing analysis by frontal swing sequence images," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 8957–8972, Mar. 2021, doi: 10.1007/s11042-020-10096-0.

[12] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1801–1810, doi: 10.1109/CVPR.2019.00190.

[13] R. Sigrist, G. Rauter, L. Marchal-Crespo, R. Riener, and P. Wolf, "Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning," *Exp. Brain Res.*, vol. 233, no. 3, pp. 909–925, Mar. 2015, doi: 10.1007/s00221-014-4167-7.

[14] T. Sueishi, C. Miyaji, M. Narumiya, Y. Yamakawa, and M. Ishikawa, "High-speed projection method of swing plane for golf training," in *Proc. Augmented Humans Int. Conf.*, New York, NY, USA, Mar. 2020, pp. 1–3, doi: 10.1145/3384657.3385330.

[15] P. Mikolaj Wozniak, J. Dominiak, M. Pieprzowski, P. Ladonski, K. Grudzien, L. Lischke, A. Romanowski, and W. P. Wozniak, "Subtletee: Augmenting posture awareness for beginner golfers," in *Proc. ACM Hum.-Comput. Interact.*, vol. 4, Nov. 2020, Art. no. 204, doi: 10.1145/3427332.

[16] A. Ikeda, Y. Tanaka, D.-H. Hwang, H. Kon, and H. Koike, "Golf training system using sonification and virtual shadow," in *Proc. ACM SIGGRAPH Emerg. Technol.*, New York, NY, USA, Jul. 2019, pp. 1–2, doi: 10.1145/3305367.3327993.

[17] K. Sasaki, K. Shiro, and J. Rekimoto, "ExemPoser: Predicting poses of experts as examples for beginners in climbing using a neural network," in *Proc. Augmented Humans Int. Conf.*, New York, NY, USA, Mar. 2020, pp. 1–9, doi: 10.1145/3384657.3384788.

[18] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining (AAAIWS)*. Palo Alto, CA, USA: AAAI Press, 1994, pp. 359–370.

[19] A. Ikeda, D.-H. Hwang, and H. Koike, "AR based self-sports learning system using decayed dynamic time warping algorithm," in *Proc. Int. Conf. Artif. Reality Telexistence Eurographics Symp. Virtual Environ.*, 2018, pp. 171–174, doi: 10.2312/egve.20181330.

[20] T. Halperin, A. Ephrat, and S. Peleg, "Dynamic temporal alignment of speech to lips," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3980–3984, doi: 10.1109/ICASSP.2019.8682863.

[21] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *Computer Vision ECCV 2016* (Lecture Notes in Computer Science), vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46448-0_32.

[22] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1134–1141, doi: 10.1109/ICRA.2018.8462891.

[23] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1801–1810, doi: 10.1109/CVPR.2019.00190.

[24] S. Haresh, S. Kumar, H. Coskun, S. N. Syed, A. Konin, M. Z. Zia, and Q.-H. Tran, "Learning by aligning videos in time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5544–5554, doi: 10.1109/CVPR46437.2021.00550.

[25] H. Seki and Y. Hori, "Detection of abnormal action using image sequence for monitoring system of aged people," *IEEJ Trans. Ind. Appl.*, vol. 122, no. 2, pp. 182–188, 2002, doi: 10.1541/ieejias.122.182.

[26] G. S. Parra-Dominguez, B. Taati, and A. Mihailidis, "3D human motion analysis to detect abnormal events on stairs," in *Proc. 2nd Int. Conf. 3D Imag., Modeling, Process., Vis. Transmiss.*, Oct. 2012, pp. 97–103, doi: 10.1109/3DIMPVT.2012.34.

[27] F. Nater, H. Grabner, and L. Van Gool, "Exploiting simple hierarchies for unsupervised human behavior analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2014–2021, doi: 10.1109/CVPR.2010.5539877.

[28] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490, doi: 10.1109/CVPR.2019.00057.

[29] R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Deformation-based abnormal motion detection using 3D skeletons," in *Proc. 8th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2018, pp. 1–6, doi: 10.1109/IPTA.2018.8608143.

[30] Y. Tang, L. Zhao, Z. Yao, C. Gong, and J. Yang, "Graph-based motion prediction for abnormal action detection," in *Proc. 2nd ACM Int. Conf. Multimedia Asia*, New York, NY, USA, Mar. 2021, pp. 1–7, doi: 10.1145/3444685.3446316.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[33] W. McNally, K. Vats, T. Pinto, C. Dulhanty, J. McPhee, and A. Wong, "GolfDB: A video database for golf swing sequencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2553–2562, doi: 10.1109/CVPRW.2019.00311.

[34] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[35] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696, doi: 10.1109/CVPR.2019.00584.

[36] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668, doi: 10.1109/ICCV.2017.288.

CHEN-CHIEH LIAO received the B.E. and M.S. degrees from the Tokyo Institute of Technology, Japan, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree with the School of Computing. His research interests include human–computer interaction, mixed/augmented reality, and machine vision.

DONG-HYUN HWANG received the B.S. degree from the School of Computer Science and Engineering, Korea University of Technology and Education, Cheonan, South Korea, in 2017, and the M.S. and Ph.D. degrees from the School of Computing, Tokyo Institute of Technology, Tokyo, Japan, in 2019 and 2022, respectively. He was a Young Research Fellow at the Japan Society for the Promotion of Science (JSPS), from 2020 to 2022. Currently, he is a machine learning researcher at CLOVA Voice&Avatar, NAVER Corporation. His research interests include mixed/augmented reality, computer vision, and machine learning.

HIDEKI KOIKE received the B.E., M.E., and Dr.Eng. degrees from the University of Tokyo, in 1986, 1988, and 1991, respectively. He was at the University of Electro-Communications, Tokyo. In 2014, he joined Tokyo Institute of Technology, Japan, where he is a Professor with the School of Computing. His research interests include vision-based human–computer interaction, human augmentation, information visualization, and usable security.