electronics

MDPI

Article

Viewpoint-Aware Action Recognition Using Skeleton-Based Features from Still Images

Seong-heum Kim [1] and Donghyeon Cho [2,*]

[1] School of AI Convergence, Soongsil University, Seoul 06978, Korea; seongheum@ssu.ac.kr
[2] Department of Electronics Engineering, Chungnam National University, Daejeon 34134, Korea
* Correspondence: cdh12242@cnu.ac.kr; Tel.: +82-42-261-3641

Abstract: In this paper, we propose a viewpoint-aware action recognition method using skeleton-based features from static images. Our method consists of three main steps. First, we categorize the viewpoint from an input static image. Second, we extract 2D/3D joints using state-of-the-art convolutional neural networks and analyze the geometric relationships of the joints for computing 2D and 3D skeleton features. Finally, we perform view-specific action classification per person, based on viewpoint categorization and the extracted 2D and 3D skeleton features. We implement two multi-view data acquisition systems and create a new action recognition dataset containing the viewpoint labels, in order to train and validate our method. The robustness of the proposed method to viewpoint changes was quantitatively confirmed using two multi-view datasets. A real-world application for recognizing various actions was also qualitatively demonstrated.

Keywords: still image action recognition; skeleton-based action recognition; viewpoint estimation

Citation: Kim, S.-h.; Cho, D. Viewpoint-Aware Action Recognition Using Skeleton-Based Features from Still Images. *Electronics* **2021**, *10*, 1118. https://doi.org/10.3390/electronics10091118

Academic Editor: Giovanni Dimauro

Received: 23 March 2021
Accepted: 5 May 2021
Published: 9 May 2021

1. Introduction

Human action recognition is useful in analyzing human behaviors and interactions for intelligent surveillance, human–computer interaction, and many other practical user interface/experience applications [1–3]. Robust representations and feature extraction for a human model play an important role in robotics applications. Recently, deep features extracted from a large amount of image data have shown impressive recognition results for practical applications. Although their recent performance on several still-image databases has almost been saturated, the real-world applications are still challenging, because of intra-action variations due to different visual appearances, such as various backgrounds and textures of foreground actors, and camera viewpoints.

To overcome these issues, we utilize a fine-tuned human body detector and pre-trained *Convolutional Neural Networks* (CNNs) for building 2D/3D skeletons [4–8]. This is because the visual variations of foreground actors and background texture can be simplified using skeleton-based representations. Given a static RGB image as an input, we can detect human bodies in 2D regions of interest (ROIs) and their main body orientation. Moreover, we learn a relative camera viewpoint for each person, by using real and synthetic images. After training the view-specific action classifiers, an action label is identified based on the combination of 2D skeleton feature from a selected viewpoint and view-invariant 3D skeleton feature in space.

Specifically, the contributions of our study are summarized as follows: First, we introduce a viewpoint as a latent variable for an action label, which is learned in a supervised manner during the training of real/synthetic images. Our CNN-based detector is fine-tuned with high-quality rendered images using reconstructed avatars. Second, we utilize three different CNN architectures to localize a human body, its 2D/3D joints, and connect them with our work together. Two skeleton features from view-dependent 2D and view-invariant 3D joint sets are computed and concatenated for training and testing view-specific *Random Forest* (RF) classifiers [9], by defining skeletal representations as Euclidean

distances between every pair of joint positions [8]. Hence, the proposed approach is robust to viewpoint changes and foreground/background texture, which is validated with the Human3.6M dataset [10] (fifteen actions from four viewpoints) and a database collected from our system (ten actions across eight viewpoints). Figure 1 shows two examples of the input images and detected actions.

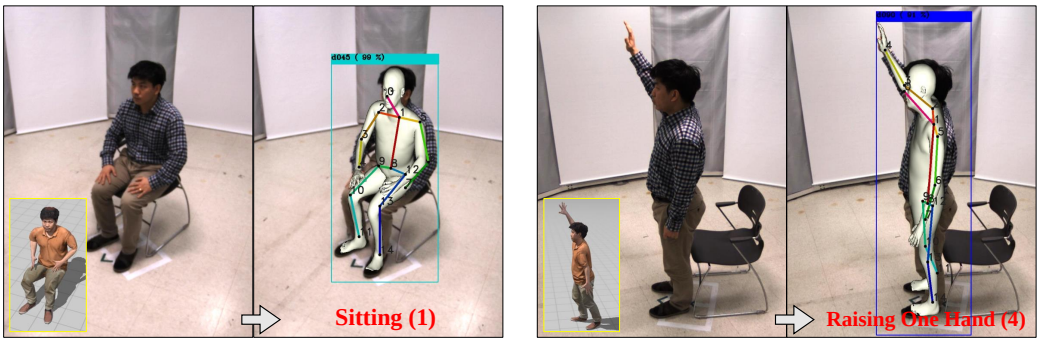


Figure 1. Examples of our view-specific action recognition.

2. Related Work

In this section, we discuss the previous studies related to skeleton-based action recognition using single images, view-invariant action representation, and the generation of synthetic human databases.

Still Image based Action Recognition: Existing methods for human action recognition typically exploit a large number of labeled action images on the web, and they often require either image-level or video-level annotations as well [11,12]. Given image-level training data, still image-based approaches identify the action or behavior of a person from only a single test image [1]. In this case, only single images are sufficient to distinguish some actions (e.g., smiling, sitting, and standing). For real-time applications, one/few-shot action recognition aims at recognizing unseen action categories when a single reference or a small amount of training examples is available (e.g., when the target action categories are not present in the current publicly available datasets) [13,14]. In contrast to temporal action recognition, this study focuses on the skeleton-based representation from a single image to simplify the visual patterns of actions.

Skeleton based Action Recognition: Human action is produced by articulated body movements. While some engineered features with RGB or RGB-D datasets often fail for ambiguous images/videos, skeleton-based representations have become more popular than traditional approaches, because of the development of cost-effective depth sensors and pose estimation algorithms [3,15,16]. In particular, our representation is inspired by the methods in [6,8], where a body pose appearing in an image is expressed by all the relative distances between every pair of joint locations. By extending this metric, we develop a skeleton-based action descriptor and train a classifier in the context of action recognition.

View-invariant Action Representation: In addition, the action can be simultaneously captured as a set of snapshots from multiple viewpoints [17]. When actors are captured by various cameras at different poses, view-invariant features can be used to classify action descriptors effectively. Previous studies have proposed low-level feature based, high-level model based methods, and mostly hybrid approaches for robust cross-view action recognition [18,19]. Similar to the methods in [20,21], our work considers a 3D skeleton feature to recognize human actions across different views. In addition, we learn view-specific classifiers using 2D skeleton data and select them to improve the recognition performance from several canonical viewpoints.

Generation of Synthetic Human Database: Several data augmentation techniques have been utilized to improve the performance of action recognition [22,23]. We refer to selected examples that are most relevant to a synthetic human database. For human 3D pose estimation, Chen et al.rendered a parametric shape completion and animation of

people (SCAPE) model with various viewpoints, clothing, and light sources, and then composited it with real-world backgrounds [24]. Similarly, Varol et al.generated synthetic training images from a 3D motion capture system. The synthetic bodies were parameterized using the skinned multi-person linear (SMPL) model [25]. As a computer graphics pipeline often causes a visual difference between the training and testing domains, we built a multiple DSLR camera system to acquire realistic shapes and textures of actors from arbitrary viewpoints [26] and utilized the Adobe Mixamo motion templates to retarget reconstructed, rigged people for various actions [27].

3. Revisiting Monocular Joint Detectors

In computer vision, the human body is digitalized as quantitative parametric representations from skeleton-based, contour-based, volume-based, and other approaches. Contour-based or volume-based models usually focus on the silhouette of a person and the boundaries in 2D or the geometric meshes and shapes captured from expensive depth-aware devices. In contrast, a skeleton model consists of a set of joints (keypoints), such as ankles, knees, shoulders, elbows, wrists, and limb orientations, comprising the skeletal structure of the human body. Owing to the relatively low dimensional parameters, this model is widely used in both 2D and 3D human pose estimation, which is also related to markerless motion capture techniques.

Human pose estimation refers to the problem of localization of human joints in images or videos. For example, it estimate a pixel location (x,y) for each joint from a RGB image. Most pose estimation methods have been greatly reshaped by CNNs, by replacing hand-crafted features and graphical models. Hence, recent methods require a huge amount of body parts (also known as keypoints) annotations and learn to associate body parts with individuals in an image. The process of monocular 3D pose estimation also involves the prediction and analysis of keypoints. Given a single image, most monocular joint detectors determine the keypoint locations of a 3D human body, based on well-developed deep learning frameworks. For example, we can directly regress the joint locations of a body model in a data-driven manner. In addition to learning 3D joint annotations, state-of-the-art methods exploit parametric human models, such as SMPL, extended SMPL, and *Adam*, to fit the predefined 3D body representation iteratively. To reconstruct 3D joint angles, several types of inputs, such as 2D keypoint score maps, depth heat maps, body part segmentation, and DensePose maps, can be considered. This optimization step is particularly important when pure 3D annotations are not sufficient to train deep representations or mixed annotations are available instead. One example of monocular 3D joint detectors is shown in Figure 2.

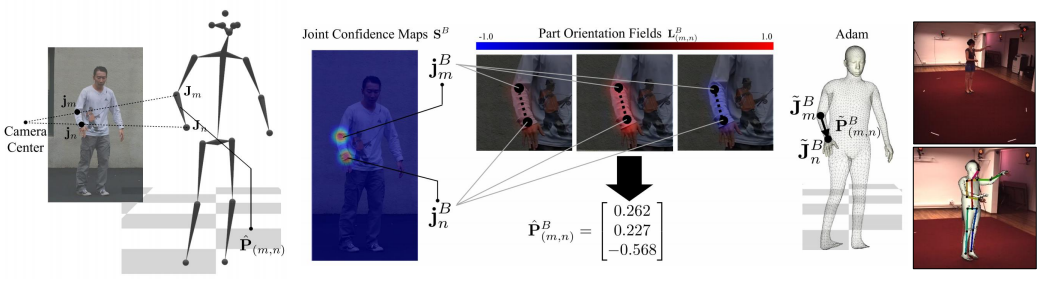


Figure 2. Example of monocular 3D joint detector (Adapted from ref [7]).

On top of this, our approach for viewpoint invariant action recognition consists of three main parts. As described in Algorithm 1, we first categorize a viewpoint from an input static image. Then, we extract 2D/3D joints using the state-of-the-art CNNs and analyze the geometric relationship of the joints for computing 2D and 3D skeleton features. Finally, we perform view-specific action classification per person based on viewpoint categorization and extracted 2D and 3D skeleton features. To train and validate our method, we implemented two multi-view data acquisition systems and created a new

action recognition dataset with the main body orientations of actors (viewpoint labels). An overview of the proposed method is presented in Figure 3.

For skeletal analysis in this study, we adopt existing, pre-trained deep networks for predicting and optimizing 2D and 3D joint location. For example, the well-known 3D joint detector, *Monocular Total Capture* (MTC) [7], inputs an image into regression networks. In the subsequent phase, the outputs of the network, namely joint confidence maps and 3D part orientation fields (similar to the 2D part affinity fields in *OpenPose* [6]), are used to reconstruct the initial parameters of the Adam model. The joint locations are optimized by fitting a deformable human model with the image measurements processed by the CNNs. However, the accuracy is severely affected by the 2D heatmap detection. As the observed data-driven results are not consistent with viewing angles, we mainly focus on applying the detected joint locations for action classification.

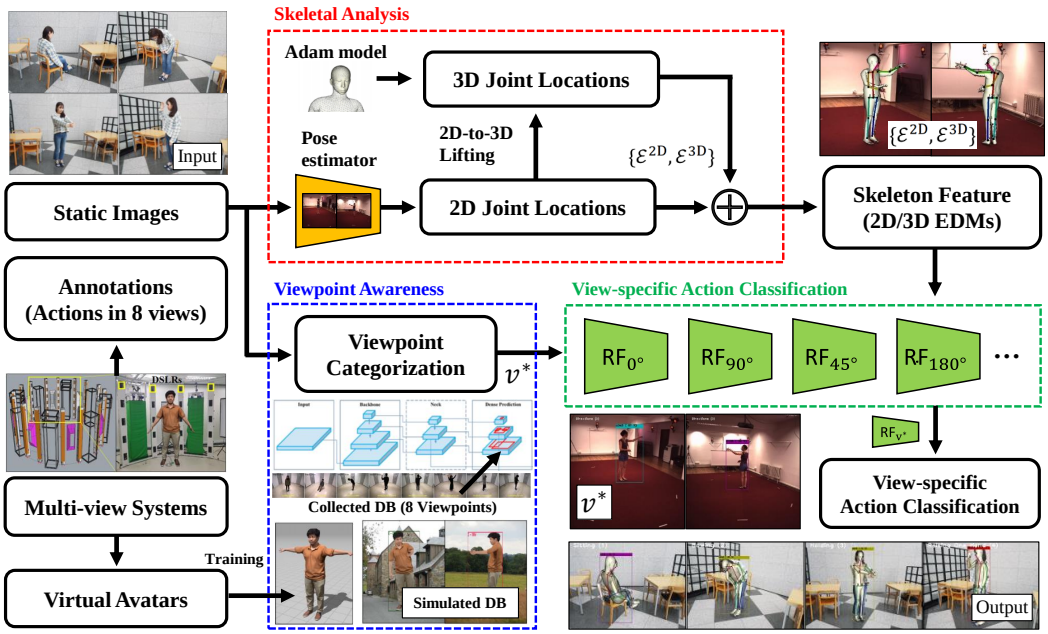


Figure 3. Algorithm overview.

Algorithm 1 Still image based classification

Require: Static image-level viewpoint and action annotations: (v_{gt}, $\mathcal{L}_{gt}$)
1: Synthetic data augmentation for viewpoint awareness (using virtual avatars)
2: Train view-specific RF classifiers with action and viewpoint annotations: RF_v
3: **for** *snapshots* $= 1, 2, \ldots$ **do**
4: 1) Viewpoint classification for each actor [5]: v^*
5: **for** *detected actors* $= 1, 2, \ldots, N$ **do**
6: 2) 2D, 3D joint estimation [6,7]: (p_i^{2D}, p_i^{3D}) (i = joint locations)
7: Skeleton representation [8]: ($\mathcal{E}_{v^*}^{2D}$, $\mathcal{E}^{3D}$)
8: 3) View-specific action classification [9]: $\mathcal{L}^*$
9: **end for**
10: **end for**

4. Viewpoint-Aware Action Recognition

Human body orientation (i.e., human viewpoint with respect to a camera) is an important attribute for resolving the view-specific ambiguity of projected body parts. Depending on the viewpoint, an action of the same person can be projected differently. The performance of action recognition is affected by the viewpoint between the camera and the object. Therefore, considering the viewpoint as a latent variable for the action

label could be a promising solution. For viewpoint awareness, our method contains multiple view-specific action classifiers for each viewpoint. Given an input image, our viewpoint categorization network predicts the viewpoint for choosing the proper the action classifier. Each action classifier uses the 2D/3D skeleton-based features as inputs. As the 2D skeleton-based features are different for each viewpoint, they are trained separately according to the viewpoints. Both 2D/3D skeleton-based features are computed from the 2D/3D joint locations estimated using a human pose estimation algorithm. Finally, the selected skeleton-based features are passed to the selected action classifier to predict the action label.

4.1. Network Architecture for Viewpoint Categorization

For simplicity, we assume that the candidate viewpoint angles are eight bins at 45°. Our viewpoint categorization network is based on *YOLO* [5], already pre-trained over a large scale image collection. This is because the backbone network is one of the state-of-the-art CNN architectures for image classification, which can be easily replaced by deeper architectures together with increasing RGB image datasets for a better recognition performance. The pre-trained CNN model handles 80 categories for cross-subject object detection; hence, we changed the last loss layer to map linearly to the PERSON category with eight viewpoint labels. Subsequently, the modified YOLO network with a new loss layer is fine-tuned with viewpoint annotation, and the remainder of the post-processing is the same as in [5]. The estimated viewpoint, v^*, from the categorization network is used to select a view-specific action classifier with view-dependent 2D skeleton-based features. The output of our viewpoint categorization network is the detection of human actors from the estimated viewpoint, v^*.

4.2. Avatar-Based Data Augmentation

In general, data augmentation has been proven to benefit the training of deep architectures. It typically consists of applying a set of transformations in either data or feature spaces, or even both. In the proposed framework, we further improve the viewpoint categorization performance by adding an adequate number of viewpoint samples from a virtual avatar-based simulation. To the best of our knowledge, this study is the first attempt to use synthetic human samples of DSLR-quality for viewpoint categorization and action recognition; thus, there is a smaller synthetic–real domain gap than the existing dataset. With data augmentation by other standard transformations, new samples are added to pre-existing real data. The most common augmentations generate new samples by applying certain transformations to the pre-existing data, and synthetically producing new samples and annotation data at hand. As a regularizer, the proposed framework avoids overfitting and increases generalization capabilities. It is well-known that deep learning can benefit from a large amount of annotated data. The problem of learning a categorical camera viewpoint is not an exception.

4.3. Skeleton-Based Features from Still Images

As our view-specific action classifiers take 2D skeleton-based features as input, we first compute 2D body skeleton joints from a static RGB image. The skeleton information in single RGB images is extracted from human pose estimators based on deep learning techniques [6]. After deriving view-dependent 2D joint positions, we exploited another third-party software package and its 2D-to-3D database to derive its corresponding 3D joint positions [7]. In practice, our system integrates two submodules for efficient and simple inference. Using the spatial coordinates of the selected 15 joints, we calculated the Euclidean distances between the 2D or 3D joint locations p_i and p_j to consider the 15×15 *Euclidean Distance Matrix* (EDM) as a relationship of the body parts [8].

$$EDM^{U}_{i,j} = \sqrt{(p_i - p_j)^2}, \quad s.t. \quad 1 \leq i < j \leq 15. \tag{1}$$

Only the non-diagonal, upper triangular elements in this affinity matrix are expressed as a 105-dimensional skeleton feature, $\mathcal{E}$. Depending on the extracted 2D or 3D joints, we can compute the 2D and 3D skeleton features for a specific spatial structure of the human body. All the skeleton features were normalized with the total length of all the relative distances between the visible joints. The distance involving a non-visible joint was zero. The 2D skeleton-based features and 3D skeleton-based features are denoted as $\mathcal{E}_v^{2D}$ and $\mathcal{E}^{3D}$, respectively. Note that v is a viewpoint, and that the 2D skeleton features are affected by viewpoints. The set of viewpoints that we consider as human body orientations for view-specific feature fusion consists of eight angles, $\mathbf{v} = \{0°, \pm45°, \pm90°, \pm135°, 180°\}$. Figure 4 shows an example of the visualization of skeleton-based features.

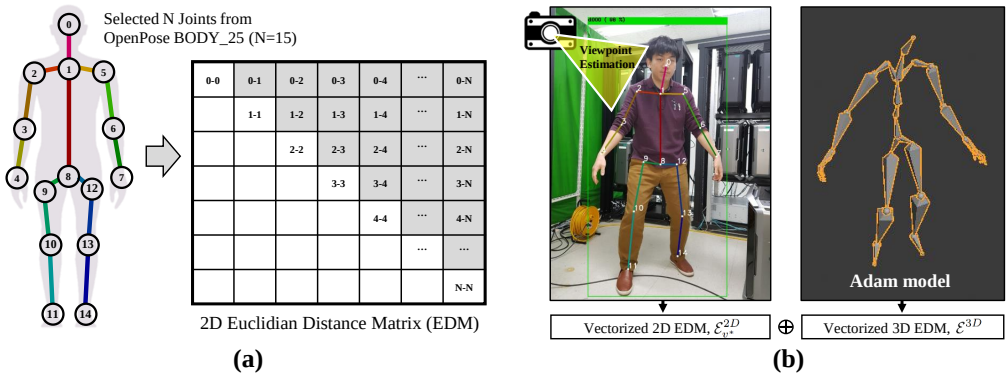


Figure 4. Skeletal representation: (**a**) skeletal representations derived from [6,7]; and (**b**) example of skeleton-based features. The vectorized EDMs were concatenated to form an action descriptor. Note that 2D EDM is highly affected by the camera viewpoint.

4.4. View-Specific Action Classification

Our view-specific action descriptor is defined as differently concatenated 2D and 3D skeleton features. Given a viewpoint v^* estimated via viewpoint categorization, we consider a 2D skeleton-based feature $\mathcal{E}_{v^*}^{2D}$. Here, the viewpoint for encoding the action descriptor is a latent variable for an action label. Our action classifiers RF_{v^*} are based on the Random Forest classifier trained over images in the viewpoint, respectively, and the selected classifier predicts the final action label as

$$\mathcal{L}^* = \mathrm{RF}_{v^*}\left(\mathcal{E}_{v^*}^{2D}, \mathcal{E}^{3D}\right), \tag{2}$$

where $\mathcal{L}^*$ is the predicted action label. As the 2D skeleton features are heavily affected by viewpoints $\mathbf{v}$ $(0°, \pm45°, \pm90°, \pm135°, 180°)$, RF classifiers $RF_{\mathbf{v}}$ are trained and tested with the concatenated 2D/3D skeleton feature $\mathcal{E}_v^{2D}$, $\mathcal{E}^{3D}$, according to the five body orientations. Here, the viewpoint for encoding a skeleton feature is a latent variable for our action label $(0°, \pm45°, \pm90°, \pm135°, 180°)$. Through cross-entropy loss, $\mathcal{L}^*$ is compared with the human annotated ground truth action label $\mathcal{L}_{gt}$ for training. Each view-specific action classifier was trained separately for each view.

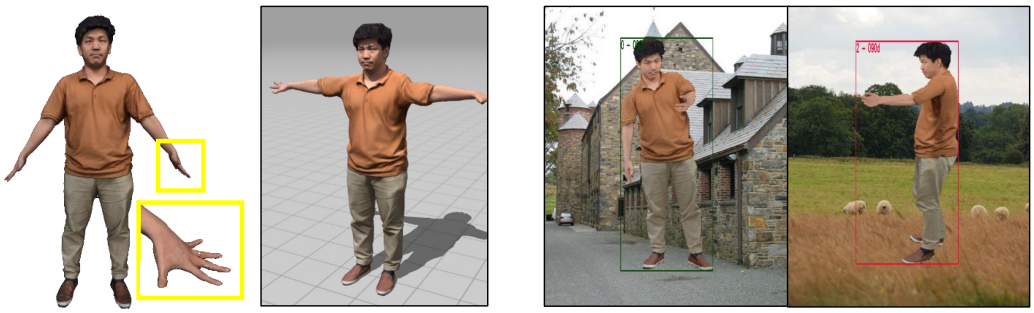
5. Experimental Results

In this section, we describe the procedures and results of our experiments, and discuss the benefits of the skeleton-based feature with viewpoint categorization, the lifted 3D skeleton feature, and our viewpoint-aware representation for view-invariant action recognition. A practical application of a mobile platform is also demonstrated.

5.1. System Setup

As few multi-view datasets with static image-level action annotations are currently available, we built our experimental setup using synchronized eight RGB cameras (2048 × 1536 pixels, The Imaging Source) and nine LED lighting sources. For each camera viewpoint,

we manually confirmed image-level action annotations. We also utilized synthetic human models reconstructed using 80 Nikon DSLR cameras and eight additional strobe lighting sources, to improve the performance of viewpoint categorization. With synchronization hardware, the first system provides an adequate amount of training and testing data for still image action recognition, and the second system accumulates realistic body templates with different body orientations for view-invariant action recognition. Figure 5 shows our avatar-based data augmentation for viewpoint categorization.



(a) **(b)**

Figure 5. Avatar-based data augmentation: (**a**) example of a reconstructed, and rigged virtual avatar; and (**b**) motion retargeting for viewpoint simulation

5.2. Data Collection

We captured eight human subjects and simulated ten virtual avatars to collect training and testing samples from eight camera viewpoints. During acquisition, all subjects were asked to perform ten types of static postures (a state of doing something). The categories of actions $\mathcal{L}$ are STANDING (0), SITTING (1), BOWING (2), HOLDING (3), RAISING ONE HAND (4), ONE HAND ON FACE (5), RAISING TWO HANDS (6), TWO HANDS ON HEAD (7), LEFT POINTING (8) and RIGHT POINTING (9). Figure 6 shows the examples of captured images. For these postures, a single static image is sufficient to distinguish the pre-defined actions. Note that there are eight body orientation annotations, according to the camera viewpoints, $\mathbf{v} = \{0°, \pm 45°, \pm 90°, \pm 135°, 180°\}$. However, the information for action recognition, according to each viewpoint, is not identical.

The *Human3.6M* (H3.6M) dataset consists of videos containing 3.6 million images, and they were recorded from four camera viewpoints [10]. Each video shows a human subject performing a target action without setting the exact action time. For the evaluation in Table 1, we collected video-level samples without image-level action annotations. Among the 11 subjects with 15 actions, we used five subjects (S5–S9) for the training and two subjects (S1 and S11) for the testing. We selected seven video samples to obtain a balanced number of images for fifteen static postures, but randomly split the train and test samples for evaluation. Manual viewpoint annotations are required for randomly selected frames for training and testing. In contrast, our database using eight cameras provides not only static image-level action annotation but also eight groups of viewpoint annotations. We were able to train eight viewpoint-specific classifiers with all the action descriptors according to the annotated main orientations.

In addition, Figure 5a shows our multi-view camera system that takes synchronized, high-resolution images from 80 DSLR cameras. From the captured DSLR images, we acquired high-quality human avatar templates using the multi-view stereo reconstruction software (e.g., Reality Capture), as shown in Figure 5b. We automatically performed rigging (Figure 5c) and retargeting (Figure 5d) by using free action templates from the Adobe Mixamo for reconstructed avatar model. In the retargeting stage, we gathered more than 50,000 backgrounds by excluding images in the PERSON category. Then, we randomly rotated the animated actors by 45° and composited multiple actors over one of the collected background images. Our annotation for the categorical viewpoints and bounding boxes

of actions can be automatically computed. In Figure 5d, two examples and viewpoint annotations are provided with random backgrounds.

5.3. Quantitative Comparison

Our focus in this paper is to combine viewpoint categorization and 2D/3D skeleton-based features for action recognition. Hence, we evaluated the quantitative performance of our feature representations with ten different configurations using two multi-view datasets. For the comparison in Table 1, the skeleton features can be obtained via two methods: EDM and EigenJoint. We report the results from two different baseline methods and three view-invariant skeleton feature combinations. For viewpoint invariance, one method is to utilize a 3D skeleton instead of directly using the detected 2D joints, and the other is to combine viewpoint-specific 2D and view-invariant 3D skeleton features. In the last method, we further scale up the volume of training datasets by adding the synthetic data. In our experiment, the proposed design choices were effective for skeleton-based action recognition.

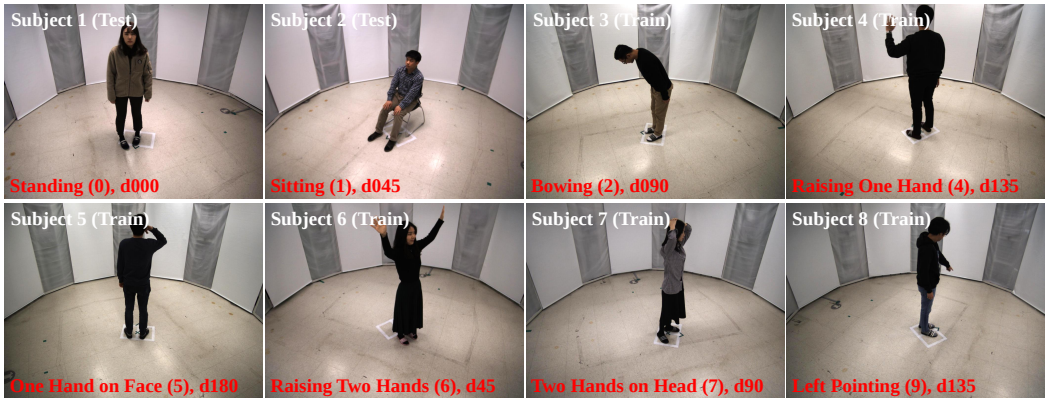


Figure 6. Data collection. Ten action categories from eight viewpoints were defined in the database.

Table 1. Quantitative comparisons.

	Feature Representations	Human3.6M [10]	8-View DB (Ours)
k-NN	2D Skeleton [6] & EDM [8]	25.97%	72.14%
	EigenJoint [16]	26.13%	72.29%
	3D Skeleton EDM [7]	31.8%	74.35%
	View specific 2D + 3D EDMs (No sim.)	43.56%	81.01%
	View specific 2D + 3D EDMs (Ours)	47.32%	85.57%
RF	2D Skeleton [6] & EDM [8]	25.63%	75.08%
	EigenJoint [16]	25.52%	74.79%
	3D Skeleton EDM [7]	32.75%	78.63%
	View specific 2D + 3D EDMs (No sim.)	45.44%	86.21%
	View specific 2D + 3D EDMs (Ours)	50.91%	89.67%

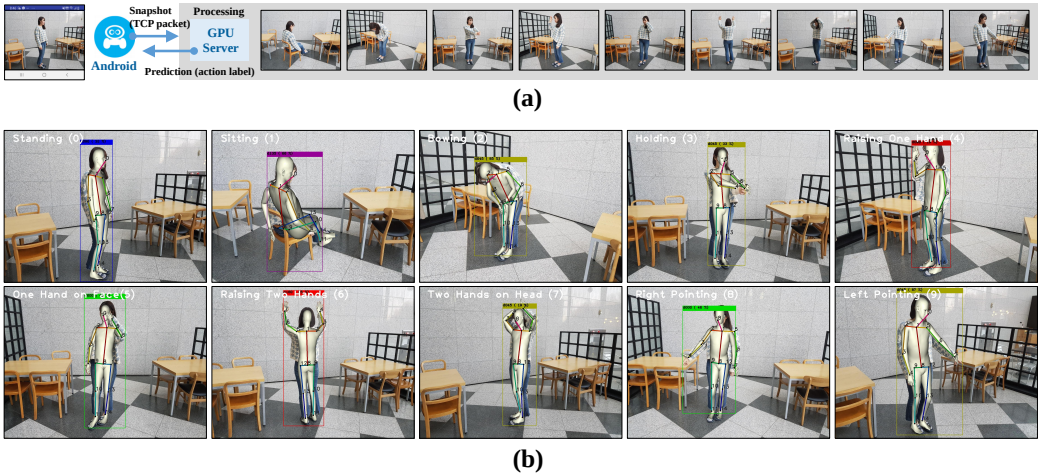
Specifically, the first experiment for factor analysis was to share a common classifier for the detected 2D features across all the viewpoints. As shown in Table 1, the dimension-reduced action descriptor in [16] for the predicted 2D skeleton is similar to the EDM method for measuring the position differences of the joints of interest. There was only a marginal difference between our EDM feature and the PCA-based feature, particularly when handling single frames. The third experiment utilized a 3D skeleton instead of simply

using the detected 2D joints. The benefit of 3D information was confirmed, as observing the result obtained using 3D skeleton feature was better than the obtained using the 2D skeleton only. Our method for skeleton representation combines viewpoint categorization with a fine-tuned human body detector and 2D/3D skeleton-based feature fusion. Not surprisingly, all the 2D-based approaches with viewpoint categorization and viewpoint-specific classifiers were better than the cases with only one common appearance model.

In practice, the extracted skeleton features can be classified using any other machine learning methods, e.g., *k-Nearest Neighbors* (kNN) and Random Forest (RF). In general, the RF classifier handles the missing values and maintains the accuracy of a large proportion of data with higher dimensions. On top of that, the benefit of using simulated images for data augmentation is also observed. The use of synthetic training images increases the diversity in terms of visual appearance of viewpoints, and achieves a better performance in action recognition for both classifiers. As long as the interpolated simulation was bounded to real-world samples, we verified the improved performances in both datasets. Based on view-specific 2D and view-invariant 3D skeleton representations, the proposed method outperformed all the other methods under all conditions. We compared the performance of the proposed methods using two multi-view datasets, respectively.

5.4. Application

The proposed skeleton-based action recognition with viewpoint categorization was effective even in a new real-world situation. We define the action of a person using only a single RGB image and predict an action label at every input frame, which is suitable for low-cost, real-time applications. Using collected viewpoint annotation from real and synthetic data and still-image action labels, we implemented our software for a GPU server, a mobile app for any device with the Android platform, and a communication module with the TCP socket interface. As shown in Figure 7, we demonstrated the effectiveness of our system for real-world applications. With multi-threading in the GPU server, we also estimated the viewpoint of the camera relative to an actor and view-specific 2D skeleton feature at 10–15 frames per second. For this operation mode, our system worked well in the real-world and real-time. This was sufficiently good for our mobile app to perform skeleton-based action recognition using viewpoint categorization. Our viewpoint-aware method qualitatively shows better action classification performances for all eight viewpoints than that without viewpoint categorization.



(a) **(b)**

Figure 7. Application: (**a**) an example of inputs for mobile applications; and (**b**) our qualitative results.

6. Discussion

Currently, no other multi-view action databases such as NTU-RGB+D [28], IXMAS [29], and i3DPost multi-view human action and interaction datasets [30] have viewpoint

labels for actors. Thus, we chose the Human3.6M dataset, action videos collected from different viewpoints, and manually obtained image-level action and viewpoint labels with five annotators. In a video sample, the actors can freely rotate their main direction and often show various undefined actions. In contrast, our system using eight synchronized cameras carefully captures static-image-level actions and eight groups of viewpoint annotations. By digging camera viewpoints, our study validates the importance of a viewpoint to clarify human actions better.

The use of skeleton data is intended for robustness to illumination and visual appearance in solving the action classification problem. Several skeleton-based methods that use graph convolutional networks to benefit from the graph structure of the skeleton data have been proposed. In the proposed method, all 2D/3D joint features, which are the Euclidean distances between body joints, are concatenated and a feature vector is used for action classification. As a practical application, this method using simple (dis-)similarities to represent actions was successfully demonstrated on a mobile platform.

However, it is not sufficient to use only one input image for some actions. In addition, images presented from different perspectives can be often different owing to other optical factors. To this end, we also observed some failure cases when a target person was significantly occluded (self-occlusion due to their pose, occlusion by other people or objects, etc.). In practice, an additional viewpoint is required when crucial joints are not visible from the input viewpoint. Solving this problem with image-based representation and a proper mode selection scheme that cannot be handled by our current skeleton-based representation is an interesting research direction in the future. For example, image-based deep features from multiple viewpoints can be combined to handle occluded joints [17]. Since we use a 2D detector, action-annotated images can also be used to directly fine-tune 2D CNNs for a deep feature. Combined with our 2D skeleton feature, it may show better performance than the current approach.

The key idea of the proposed skeleton-based action recognition approach is to consider a viewpoint as a latent variable for an action label. Hence, we first categorized the viewpoint of the input image, then extracted the 2D/3D joints using the CNN methods, and finally built the skeleton data considering the relationship of the joints. Performing skeleton-based action recognition through viewpoint classification is a practical, effective, and pragmatic problem-solving strategy. We used simulated images to categorize the main body orientations as well. Based on the improved viewpoint categorization, we performed viewpoint-specific action classification in a test image. The robustness of our system to viewpoint changes was quantitatively validated using two multi-view datasets. Overall, we think the proposed method is a standard, practical approach for using CNN/RFs. With more large-scale datasets in the future, we expect to utilize the viewpoints implicitly in their higher layers in larger and deeper networks for action classification. All types of human analyses, including 2D/3D joint feature extraction, can be improved by adopting a concept of viewpoints.

7. Conclusions

In this paper, we propose an approach for static-image action recognition using human body viewpoint categorization and skeleton-based features. During the training stage, we utilize a virtual avatar-based simulation to create new human samples of DSLR-quality to improve the performance of viewpoint categorization. In addition, based on the Euclidean distances of the detected 2D and 3D joint locations, the proposed method combines view-specific 2D skeleton features and lifted 3D skeleton features for view invariance. We evaluated our algorithm using the public dataset Human3.6M and our own database from eight synchronized RGB cameras. The concatenated action descriptor of the selected 2D EDM and 3D EDM features showed a better performance compared with the two separate features independently. The proposed method outperformed all other methods by combining viewpoint-specific RF classifiers. Based on real/simulated DBs, our results indicate that the simulated data with viewpoint labels gave a boost in action recognition

accuracy. Finally, a real-world application for practical action recognition was successfully demonstrated.

Author Contributions: Conceptualization, S.-h.K.; formal analysis, S.-h.K.; investigation, S.-h.K. and D.C.; data curation, S.-h.K.; writing–original draft preparation, S.-h.K.; writing–review and editing, D.C.; supervision, S.-h.K. and D.C.; funding acquisition, S.-h.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1G1A1009828).

Acknowledgments: The first author sincerely appreciates Youngbae Hwang at Chungbuk Univ., and all the researchers at Korea Electronics Technology Institute (KETI) (especially Hanna Ryu, Jungho Kim, Ju Hong Yoon, Min-Gyu Park, Han-mu Park, Min-ho Lee, and Ju-mi Kang) for valuable discussion. We also appreciate Je-hyeong Kim, Heon-ki Kim at IOYS, Geun-tae Ryu and Ka-yeong Kim at IPL, and lastly Ki-san Hwang at Yujin Robot for system implementation and mobile demonstration.

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

References

1. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit. (PR)* **2014**, *47*, 3343–3361.
2. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21.
3. Presti, L.; Cascia, M.L. 3D skeleton-based human action classification: A survey. *Pattern Recognit. (PR)* **2016**, *53*, 130–147.
4. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
5. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
6. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
7. Xiang, D.; Joo, H.; Sheikh, Y. Monocular total capture: Posing face, body, and hands in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10965–10974.
8. Moreno-Noguer, F. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2823–2832.
9. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
10. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2014**, *36*, 1325–1339.
11. Zhao, Z.; Ma, H.; You, S. Single image action recognition using semantic body part actions. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3391–3399.
12. Tsai, J.-K.; Hsu, C.-C.; Wang, W.-Y.; Huang, S.-K. Deep Learning-Based Real-Time Multiple-Person Action Recognition System. *Sensors* **2020**, *20*, 4758.
13. Fanello, S.R.; Gori, I.; Metta, G.; Odone, F. One-shot learning for real-time action recognition. In Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Madeira, Portugal, 5–7 June 2013; pp. 31–40.
14. Bo, Y.; Lu, Y.; He, W. Few-Shot Learning of Video Action Recognition Only Based on Video Contents. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 595–604.
15. Rahmani, H.; Bennamoun, M. Learning Action Recognition Model from Depth and Skeleton Videos. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5833–5842.
16. Yang, X.; Tian, Y. L. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 14–19.
17. Keceli, A.S. Viewpoint projection based deep feature learning for single and dyadic action recognition *Expert Syst. Appl.* **2018**, *104*, 235–243.
18. Wang, J.; Nien, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 2649–2656.
19. Rahmani, H.A.; Shah, M. Learning a deep model for human action recognition from novel viewpoints. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2017**, *40*, 667–681.
20. Xia, L.; Chen, C.-C.; Aggarwal, J. K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27
21. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3D skeletons as points in a lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 588–595.

22. Crasto, N.; Weinzaepfel, P.; Alahari, K.; Schmid, C. MARS: Motion-augmented RGB stream for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7882–7891.
23. Liu, J.; Rahmani, H.; Akhtar, N.; Mian, A. Learning human pose models from synthesized data for robust RGB-D action recognition. *Int. J. Comput. Vis. (IJCV)* **2019**, *127*, 1545–1564.
24. Chen, W.; Wang, H.; Li, Y.; Su, H.; Wang, Z.; Tu, C.; Lischinski, D.; Cohen-Or, D.; Chen, B. Synthesizing training images for boosting human 3D pose estimation. In Proceedings of the International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 479–488.
25. Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.; Laptev, I.; Schmid, C. Learning from synthetic humans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 109–117.
26. Realtity Capture. Available online: https://www.capturingreality.com (accessed on 30 November 2018).
27. Adobe Mixamo. Available online: https://www.mixamo.com (accessed on 30 November 2018).
28. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
29. Weinland, D.; Ronfard, R.; Boyer, E. Free Viewpoint Action Recognition using Motion History Volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257.
30. Gkalelis, N.; Kim, H.; Hilton, A.; Nikolaidis, N.; Pitas, I. The i3DPost Multi-view and 3D Human Action/interaction Database. In Proceedings of the 2009 Conference for Visual Media Production, London, UK, 12–13 November 2009; pp. 159–168.