energies

Article

Development of a Deep Neural Network Model for Estimating Joint Location of Occupant Indoor Activities for Providing Thermal Comfort

Eun Ji Choi [1,†], Jin Woo Moon [1,†], Ji-hoon Han [2] and Yongseok Yoo [3,*]

[1] School of Architecture and Building Science, Chung-Ang University, Seoul 06974, Korea; ejjchl77@gmail.com (E.J.C.); gilerbert73@cau.ac.kr (J.W.M.)
[2] Department of Biomedical Engineering, Sungkyunkwan University, Suwon 16419, Korea; lphion96@naver.com
[3] Department of Electronics Engineering, Incheon National University, Incheon 22012, Korea
* Correspondence: yyoo@inu.ac.kr; Tel.: +82-32-835-8453
† These authors contributed equally to this work as co-first author.

Abstract: The type of occupant activities is a significantly important factor to determine indoor thermal comfort; thus, an accurate method to estimate occupant activity needs to be developed. The purpose of this study was to develop a deep neural network (DNN) model for estimating the joint location of diverse human activities, which will be used to provide a comfortable thermal environment. The DNN model was trained with images to estimate 14 joints of a person performing 10 common indoor activities. The DNN contained numerous shortcut connections for efficient training and had two stages of sequential and parallel layers for accurate joint localization. Estimation accuracy was quantified using the mean squared error (MSE) for the estimated joints and the percentage of correct parts (PCP) for the body parts. The results show that the joint MSEs for the head and neck were lowest, and the PCP was highest for the torso. The PCP for individual activities ranged from 0.71 to 0.92, while typing and standing in a relaxed manner were the activities with the highest PCP. Estimation accuracy was higher for relatively still activities and lower for activities involving wide-ranging arm or leg motion. This study thus highlights the potential for the accurate estimation of occupant indoor activities by proposing a novel DNN model. This approach holds significant promise for finding the actual type of occupant activities and for use in target indoor applications related to thermal comfort in buildings.

Keywords: thermal comfort; deep neural network; human joint estimation; indoor activity

check for updates

Citation: Choi, E.J.; Moon, J.W.; Han, J.-h.; Yoo, Y. Development of a Deep Neural Network Model for Estimating Joint Location of Occupant Indoor Activities for Providing Thermal Comfort. *Energies* **2021**, *14*, 696. https://doi.org/10.3390/en14030696

Academic Editor: Boris Igor Palella
Received: 16 December 2020
Accepted: 25 January 2021
Published: 29 January 2021

1. Introduction

The indoor environment quality (IEQ) in buildings is related to the quality of life, health, and productivity [1–3]. The main factors composing the IEQ are classified into thermal comfort, indoor air quality (IAQ), and visual comfort [4,5]. Among these factors, the degree of thermal comfort is decided by predicted mean vote (PMV), one of the thermal comfort indices proposed by Fanger [5]. PMV considers six physical factors and two individual factors of metabolic rate and clothing insulation. While physical factors can be measured simply with sensors, the individual factors are difficult to measure objectively and accurately.

One of these individual factors, the metabolic rate, is the real-time rate of change in the heat production of the human body. The heat production of a human is mostly determined by activity, and the amount of produced heat varies greatly by activity: "Sleeping" (100 W), "light work" (200 W), "walking" (300 W), "jogging" (800 W) [6]. In other words, the metabolic rate can be estimated if the activity that has a decisive influence on the heat production is identified in real-time.

However, unlike other environmental factors, activities of occupants are difficult to measure simply and accurately with sensors. Therefore, the metabolic rate is generally inferred from the measurement of indirect factors that change in response to human activity, such as the body surface temperature measured using an infrared camera [7,8]. However, in order to estimate the PMV for thermal comfort, it is necessary to develop a fundamental measuring method of the metabolic rate. Therefore, to estimate the metabolic rate, a method for identifying human activities in real-time is a promising approach.

A human activity estimation method being developed continuously not only in architecture but also in the computer vision field classifies human activities by learning image patterns [9–12]. Automatic analysis of human behavior from an image is an important but challenging problem. The human activity estimation method estimates the location of each joint of a person from an image and then uses this information to analyze the behavior of the target person. In classical approaches, researchers have proposed hand-crafted features specifically designed for certain attributes of joints in an input image. For example, a histogram of oriented gradients (HOG) [9] is a widely used feature for human pose detection. Because a HOG is based on a gradient, it is robust to unexpected illumination changes in the input image. A classifier is trained to learn the pattern of local features for object recognition [9,10]. In such a framework, features and corresponding classifiers must be carefully chosen for a target application.

By contrast, neural network-based models learn important features on their own from training data rather than requiring a researcher to arbitrarily design the features. Several attempts have been made to improve human activity estimation using neural networks. Toshev and Szegedy [11] proposed DeepPose, a deep neural network (DNN), for human activity estimation from an image, and this model is able to recognize various human activities in images with complex backgrounds. The DeepPose study demonstrated the potential of DNN for improving the activity estimation techniques. Each layer of a DNN learns progressively better representations of the target object in an image. This property allows one to identify features better from training data. Another interesting property of the DNN-based approach is that it facilitates a holistic approach that uses not only local features in local regions of an image but also the context information of the whole image.

However, conventional human activity estimation models are limited to outdoor applications, such as pedestrian detection [9,10,12] and sports activities such as Leeds Sports Pose (LSP) [13] and the Parse dataset [14–16]. These datasets mainly focus on upright postures such as walking, running, and standing, and do not include other postures more typical of indoor activity, such as sleeping, reclining, and sitting. However, because modern people spend on average more than 80% of their time indoors, more accurate indoor activity estimation methods are essential when seeking to provide a comfortable and customized indoor environment that considers the occupants' lifestyle and activity. In addition, based on human–computer interaction (HCI), smart services and technologies by considering indoor activity of an occupant can be provided in real-time. In order to identify indoor postures, it is necessary to develop a dataset and estimation model of indoor activity trained using this data.

In this study, a DNN-based model that estimates the joints location of the occupant for various indoor activities was developed. In the model development process, images of various indoor activities were collected from the laboratory and the internet. The DNN model was developed to estimate the joint coordinates of the person in the image, and the accuracy of the proposed model was assessed. The developed model is a vital technology for being applied to estimate the actual activity [17,18], and it is believed that the metabolic rate and the PMV of occupants can be calculated based on the estimated indoor activity. In addition, this possibility is expected to enable PMV-based environmental control and enhance indoor thermal comfort.

The remainder of this paper is organized as follows. In Section 2, related research on human activity estimation is summarized. Section 3 presents the 10 selected indoor activities and the method for collecting image data. In addition, the structure and parameters of the DNN model are presented. Section 4 presents the results of an analysis of the accuracy of the proposed model. Finally, Section 5 discusses the implications of the study results and future research directions.

2. Related Research

Extensive research has been conducted on how to recognize and analyze human behaviors from images, beginning with studies on how to detect a person present in an image. The most widely known and used technique for this is based on HOG. This technique was first proposed by Dalal and Triggs [9], who presented a method for calculating features based on the distribution of a gradient by region and detecting a human body by combining the HOG features of adjacent regions. This gradient-based feature has the advantage of being robust to external influences, such as illumination, and it was designed to first calculate the distributions by region and then combine them to recognize complex human behavior. Various algorithms have since been proposed to improve the accuracy and speed of the HOG method [10,19]. As such, there have been ongoing studies into approaches for designing features for partial regions and recognizing an object by combining the features of surrounding regions. For example, approaches such as the “bag-of-features” approach to object recognition have been tried [20]. Felzenszwalb [21] proposed a part-based model for object detection that is a multi-scale approach that detects a whole object by considering the relationships between parts and filtering certain parts of a target object. In this way, the possibility of detecting the entire object and the part constituting the object was confirmed.

Improving such image recognition techniques, extensive studies have been performed to address the human activity estimation problem. Yang and Ramanan [15] proposed a part-based model that simultaneously detects parts of the human body and identifies the relations between adjacent structures. This part-based model has a flexible structure that facilitates the capture of complex human activity information. Compared to conventional human activity models, this model is faster and more accurate. Chen and Alan [16] developed a model that utilizes pairwise spatial relationships between joints and local image patches. Their model is a form of a graphical model combined with a deep convolutional neural network (DCNN) that learns conditional probabilities regarding whether body parts exist in an image patch and detects body parts in a local image patch using the spatial linking of the joint relationships. However, in these studies, most features are based on the characteristics of specific regions of an image; the overall characteristics of an image are not considered. In particular, feature-based methods perform poorly for foreshortening and occluded joints in an image.

To overcome these limitations, several attempts have been made recently to use DNNs. DNNs with a large number of layers can learn progressively complex features effectively from input images. Furthermore, a DNN learns to detect complex patterns for a whole image, not just local patches. The DeepPose model proposed by Toshev and Szegedy [11] was the first model that used these advantages in human activity estimation. In their study, the last layer of a neural network called AlexNet [22] was modified, resulting in significantly higher accuracy than is achievable by conventional image recognition methods, to output the coordinates of joints from an input image. However, the dataset used in this model had a high proportion of poses related to sports, and it was difficult to predict joints that were not visible in the image. State-of-the-art research has also been conducted on the estimation of hands and poses using the learning of three-dimensional (3D) coordinates [23,24]. This approach has the advantage of being able to estimate three-dimensional coordinates; however, estimating the coordinates from a single two-dimensional (2D) image can cause perspective distortion. This limitation needs to be resolved because it increases the possibility of error when predicting 3D values from 2D sources.

3. Methods

In this study, a DNN-based human activity estimation model that estimates human joint coordinates from various types of indoor activity images was developed. The 14 major joints of the human body are as follows: (1) Right ankle, (2) right knee, (3) right hip, (4) left hip, (5) left knee, (6) left ankle, (7) right wrist, (8) right elbow, (9) right shoulder, (10) left shoulder, (11) left elbow, (12) left wrist, (13) neck, and (14) head. The DNN-based model was constructed to output the coordinates of human joints from an input image. Complex behaviors were recognized by increasing the number of layers. The residual deep learning method [25] was used to train the DNN effectively with a large number of layers.

The development process for the model is presented in Figure 1. First, a dataset consisting of images of people in various indoor activities was constructed, and the coordinates of 14 major joints were marked manually for each image. Second, the DNN was trained and optimized with 80% of the dataset. The DNN model took the images as input and produced the location of the 14 joints as output. Third, the estimation accuracy was tested with the remaining 20% of the dataset.

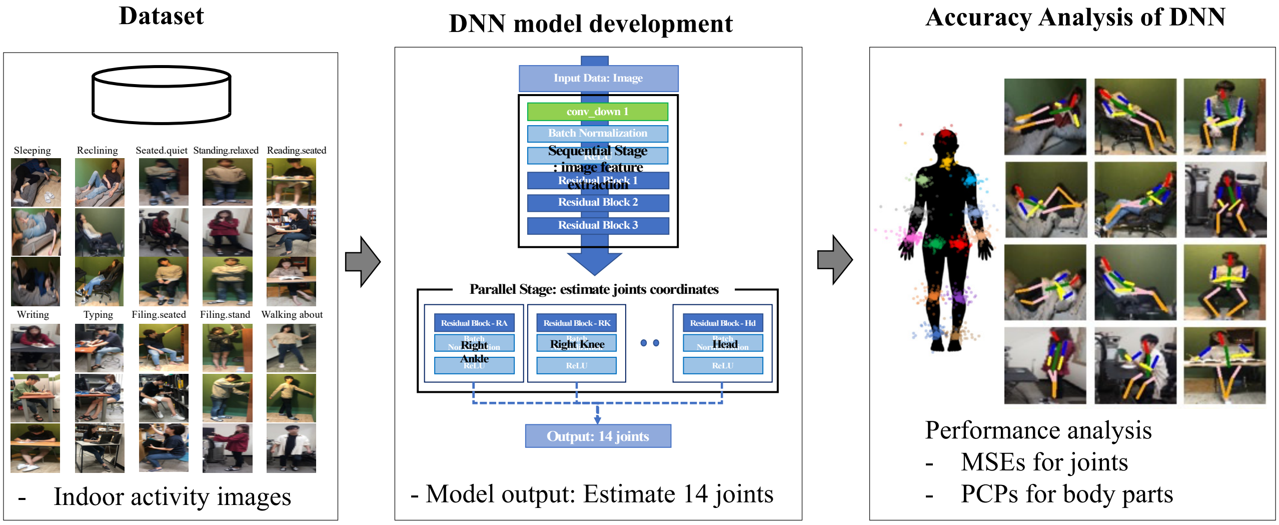


Figure 1. The activity estimation process for indoor activity images.

3.1. Datasets

First, representative indoor activities were selected from Table 1, "Metabolic Rates for Typical Tasks," of ASHRAE standard 55, produced by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) [26]. ASHRAE standard 55 provides a list of various indoor activities. From this list, 10 representative indoor activities within the home and at the office were selected (Table 1): Sleeping, reclining, sitting quietly, standing in a relaxed manner, reading while seated, writing, typing, filing while seated, filing while standing, and walking about. Of the 10 selected activities, five were sitting, three were standing, one was lying, and one was reclining.

Table 1. 10 representative indoor activities.

Type	Resting	Office Activities
Activities	Sleeping Reclining Seated, quiet Standing, relaxed	Reading, seated Writing Typing Filing, seated Filing, stand Walking about

In total, 870 images for the 10 indoor activities were collected from the internet and by photographing participants directly in the laboratory. Images showing as many human joints as possible were used because, if some of the joints are not visible in the images, suboptimal training will occur. The collected images were resized to 128 × 128 for the training and testing of the proposed model. Figure 2 shows examples of resized laboratory images for the 10 indoor activities.

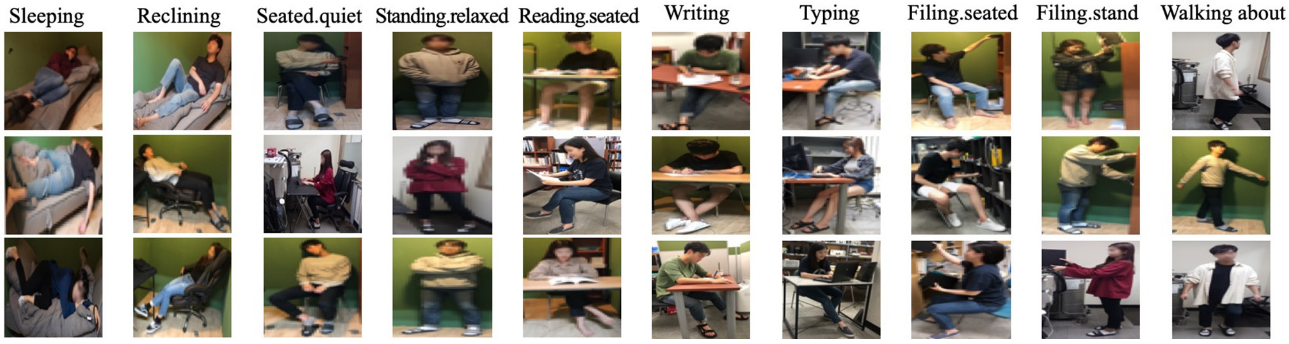


Figure 2. Examples of the images taken in the laboratory.

A human activity was defined by the coordinates of the 14 joints as shown on the left side of Figure 3: (1) Right ankle, (2) right knee, (3) right hip, (4) left hip, (5) left knee, (6) left ankle, (7) right wrist, (8) right elbow, (9) right shoulder, (10) left shoulder, (11) left elbow, (12) left wrist, (13) neck, and (14) head. These are identical to the joints defined in the Leeds Sports Pose (LSP) dataset [13]. For different joint models, the proposed DNN model can be applied by modifying the final output layer only.

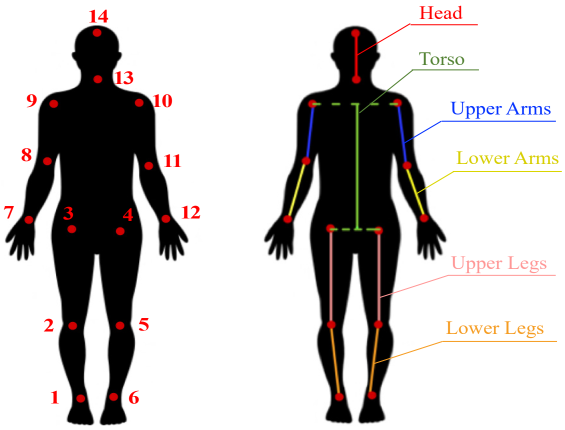


Figure 3. Major joints and body parts of human.

Using the 14 joints, 10 body parts were defined, as shown on the right side of Figure 3. Considering body symmetry, these 10 body parts could be further reduced to six: The head, torso, upper arms, lower arms, upper legs, and lower legs. Each body part, except the torso, was defined by connecting two adjacent joints. The torso (the solid green line) was defined as the centerline connecting the center points of the shoulders (joints 9 and 10) and hips (joints 3 and 4).

In the dataset, the coordinates of some joints that were not visible (because they were occluded or out of the image) were treated separately with a special annotation (−1), and these joints were excluded from the accuracy analysis.

3.2. Residual Deep Learning for Human Activity Estimation

The proposed model consisted of 138 convolutional layers to learn the features of the images and 14 fully connected layers for the output of the joint coordinates. To overcome the problem of vanishing or exploding gradients as the number of layers increases and to improve the overall accuracy, the residual deep learning method proposed by He et al. [25] was used in this study.

A residual block is a unit that has three convolution layers with a shortcut, followed by batch normalization (BN) and a rectified linear unit (ReLU) (see Figure 4). A shortcut connection was added to directly connect the input and output of every three convolution layers. It prevents the gradient vanishing problem in training a very deep neural network. The values of the kernel size (k × k) of each convolution layer and the stride were set to one. Introducing numerous shortcuts prevented the vanishing gradient problem in training the proposed DNN with a very larger number of layers.

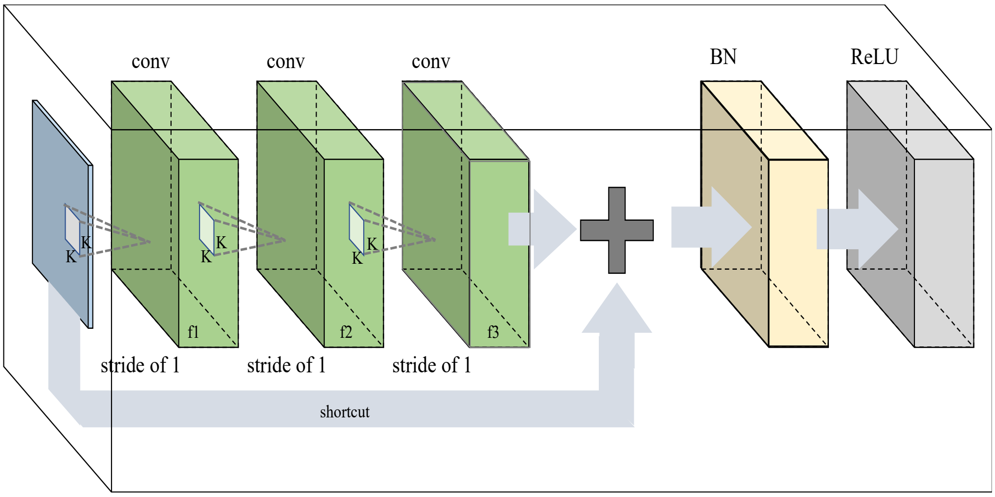


Figure 4. The structure of the residual block unit.

The deep residual network obtained using a combination of residual blocks in two stages is shown in Figure 5. The sequential stage learns the whole image features sequentially, and the parallel stage independently estimates the coordinates of each of the 14 joints with 14 residual blocks. The sequential stage consists of four convolutional down-sampling (conv-down) layers and 16 residual blocks, and the sequence includes a conv-down layer and repeated multiple residual blocks. Parameters of the individual residual blocks and convolutional layers are denoted on the right side of Figure 5. An input image passes through three residual blocks (1–3) after its spatial resolution is lowered by passing through the conv-down 1, BN, and ReLU stages. After passing through conv-down2, it goes through four residual blocks (4–7). Then, after passing through conv-down3, it goes through six residual blocks (8–13). Finally, after passing through conv-down 4, it goes through three residual blocks (14–16). The coordinates are then refined by going through residual blocks in parallel for each joint. As a result, the total parameter size is 1,734,490,260.

For the purpose of training, the proposed model was initialized with the Xavier initializer, and the Adadelta optimizer [27] was used to minimize the Euclidean distances between the estimated joint coordinates and the true coordinates. The training was performed for 100 epochs, using a batch size of 45 and a learning rate of 0.01.

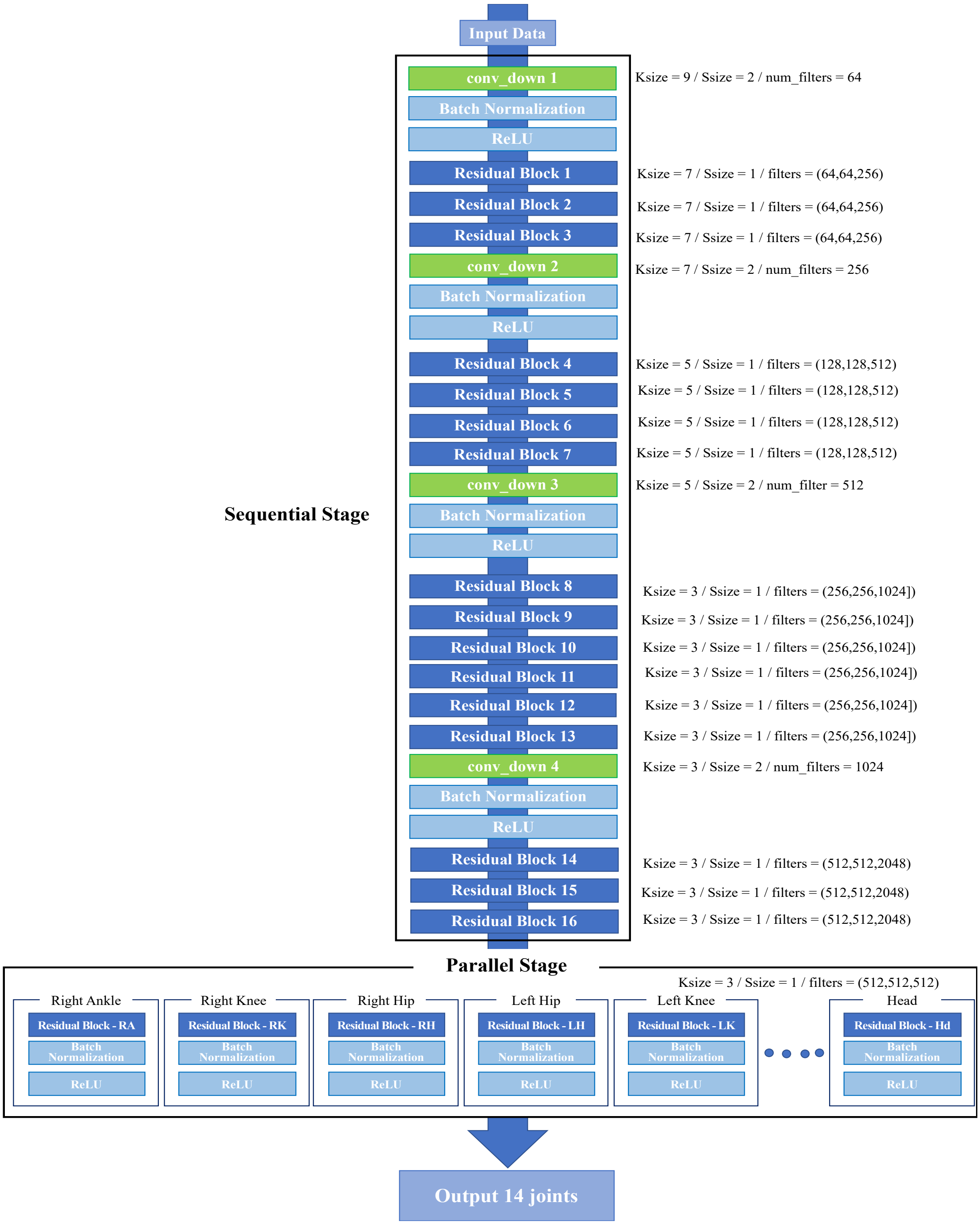

Input Data
Sequential Stage
conv_down 1
Ksize = 9 / Ssize = 2 / num_filters = 64
Batch Normalization
ReLU
Residual Block 1
Ksize = 7 / Ssize = 1 / filters = (64,64,256)
Residual Block 2
Ksize = 7 / Ssize = 1 / filters = (64,64,256)
Residual Block 3
Ksize = 7 / Ssize = 1 / filters = (64,64,256)
conv_down 2
Ksize = 7 / Ssize = 2 / num_filters = 256
Batch Normalization
ReLU
Residual Block 4
Ksize = 5 / Ssize = 1 / filters = (128,128,512)
Residual Block 5
Ksize = 5 / Ssize = 1 / filters = (128,128,512)
Residual Block 6
Ksize = 5 / Ssize = 1 / filters = (128,128,512)
Residual Block 7
Ksize = 5 / Ssize = 1 / filters = (128,128,512)
conv_down 3
Ksize = 5 / Ssize = 2 / num_filter = 512
Batch Normalization
ReLU
Residual Block 8
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
Residual Block 9
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
Residual Block 10
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
Residual Block 11
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
Residual Block 12
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
Residual Block 13
Ksize = 3 / Ssize = 1 / filters = (256,256,1024])
conv_down 4
Ksize = 3 / Ssize = 2 / num_filters = 1024
Batch Normalization
ReLU
Residual Block 14
Ksize = 3 / Ssize = 1 / filters = (512,512,2048)
Residual Block 15
Ksize = 3 / Ssize = 1 / filters = (512,512,2048)
Residual Block 16
Ksize = 3 / Ssize = 1 / filters = (512,512,2048)
Parallel Stage
Ksize = 3 / Ssize = 1 / filters = (512,512,512)
Right Ankle
Residual Block - RA
Batch Normalization
ReLU
Right Knee
Residual Block - RK
Batch Normalization
ReLU
Right Hip
Residual Block - RH
Batch Normalization
ReLU
Left Hip
Residual Block - LH
Batch Normalization
ReLU
Left Knee
Residual Block - LK
Batch Normalization
ReLU
Head
Residual Block - Hd
Batch Normalization
ReLU
Output 14 joints

Figure 5. The structure of proposed deep neural network (DNN) model for activity estimation.

In the training of the model, the training data were augmented using various geometric transforms. Of the total 870 images, 696 images (80%) and joint coordinates were used for training, using stratified sampling that maintained the same proportion of each activity. After shuffling and pixel normalization, augmentation was performed with (1) rotation, (2) vertical flipping, and (3) random cropping. The rotation process increased the amount of training data by a factor of 11 by rotating an image by 5° at a time within the range of −30° to 30°, increasing twice using a vertical flip (flipping the left-hand and right-hand sides of an image around the center line of the image), and performing random cropping, by which an image with a person was scaled to various sizes and translations. To apply random cropping, a bounding box was set-up in an image including all human joint coordinates tightly. The four-sided gaps around the tight bounding box from the original image frame were divided into 50 equal intervals and defined as padding. A padded bounding box was generated by selecting a padding size between 1 and 25 intervals. Finally, a bounding box (Bbox) was applied to the augmented data by positioning the padded bounding box randomly between the space of the tight bounding box and the padded bounding box. Random cropping was performed before the training started, and the number of images was increased by a factor of 16. By using data augmentation, the number of training images was increased 352 times. Therefore, the augmented dataset included 696 × 352 = 244,992 images.

3.3. Analysis Metrics

To quantify the accuracy of the joint estimations, the mean squared error (MSE) and percentage of correct parts (PCP) were used to evaluate the joint coordinates and body part estimation, respectively. The MSE is defined as the mean of the squares of the distances between the estimated joint coordinates and the true joint coordinates, as shown in Equation (1):

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2, \tag{1}$$

where $(\hat{x}_i, \hat{y}_i)$ and (x_i, y_i) are the estimated and true coordinates of the i_{th} joint, respectively, and n is the number of joints which is 14 in this model.

PCP is a method proposed by Eichner et al. [28] that is used as a metric to evaluate the estimation accuracy of body parts such as the torso and limbs. The estimated position of a body part is considered correct when the errors for the joint coordinates at both ends of the body part are smaller than a predetermined threshold. The blue dots and lines in Figure 6 represent the true joints and body parts, respectively. Orange dots and lines represent estimated joints and body parts, respectively. Typically, the threshold is set to half of the length of the body part of interest. For example, in Figure 6, the upper arm is correctly estimated, because the estimated right shoulder and elbow positions are within the threshold ranges (shown by circles). Thus, in this study, the PCP was assumed to indicate that an estimated coordinate was correct if ∣Estimated joint coordinates—ground-truth joint coordinates∣ < (length of limb) / 2.

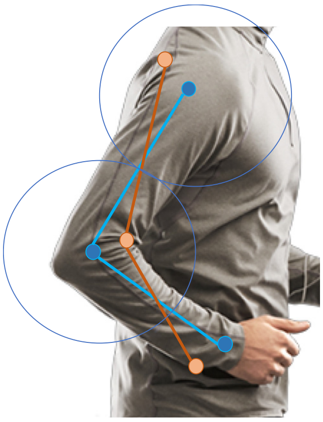


Figure 6. The concept of the percentage of correct parts (PCP).

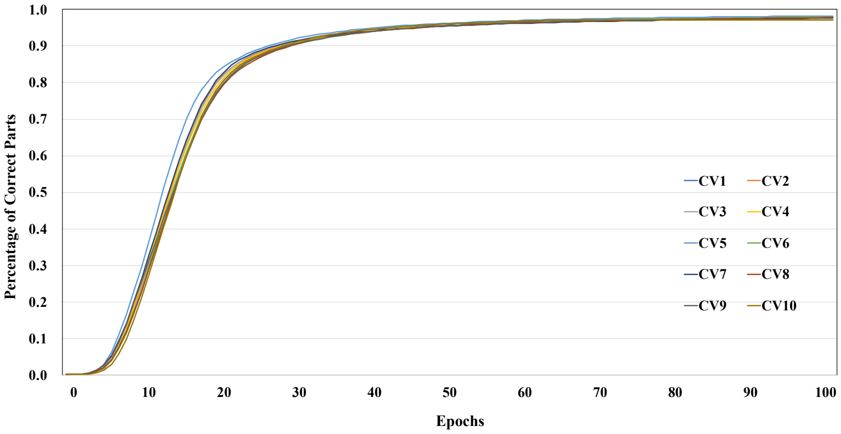
The MSE and PCP were measured using 10 cross-validations. For each cross-validation, the indoor dataset was randomly partitioned into training (80%) and test sets (20%) for the fixed activity ratios.

4. Results and Discussion

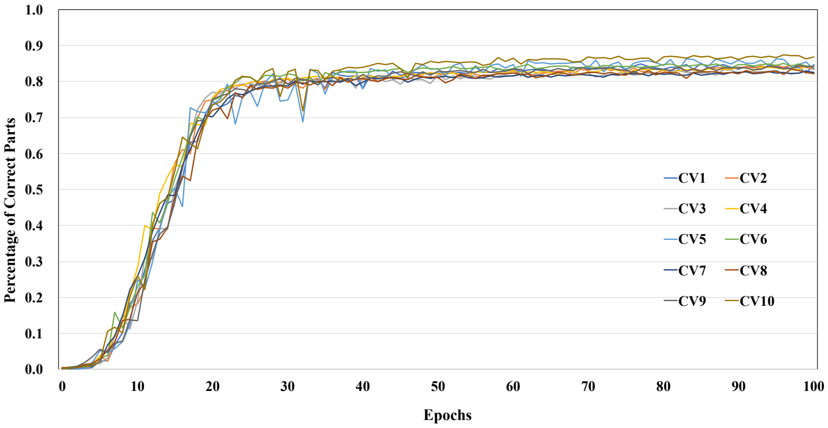
4.1. Training the Deep Residual Network for Activity Estimation

Despite having many layers and a large number of parameters, the model developed in this study converged within only 100 epochs when trained with the augmented dataset.

This was a quite promising result, considering the larger number of layers and the size of the training dataset. The PCPs present the accuracy of the model by performing 10 cross-validations with different partitions of the dataset. Cross-validation was performed prior to testing to verify that the model was trained regardless of the data sampling impact. After 100 epochs, the training PCP converged close to 1.0 (Figure 7a) and the test PCP converged to 0.86 (Figure 7b). The model did not exhibit overfitting to the test dataset, and the mean PCP value converged to values greater than 0.80 in all 10 cross-validations.



(**a**) The training PCPs



(**b**) the test PCPs

Figure 7. Training results of the proposed DNN model.

The MSE of joint coordinate estimation was confirmed using the converged learning model. In addition, the accuracy of the PCP for each body part and for each indoor activity was analyzed using the joints estimated in the output of the model.

4.2. *Accuracy Analysis of the Estimated Joints*

To evaluate the accuracy of the model, the MSE was measured for each of the estimated joints. As shown in Figure 8, the MSE value of each joint shows the error of the estimated coordinates in the images with 128 × 128 pixels, according to the numbering of the 14 joints shown in Figure 3.

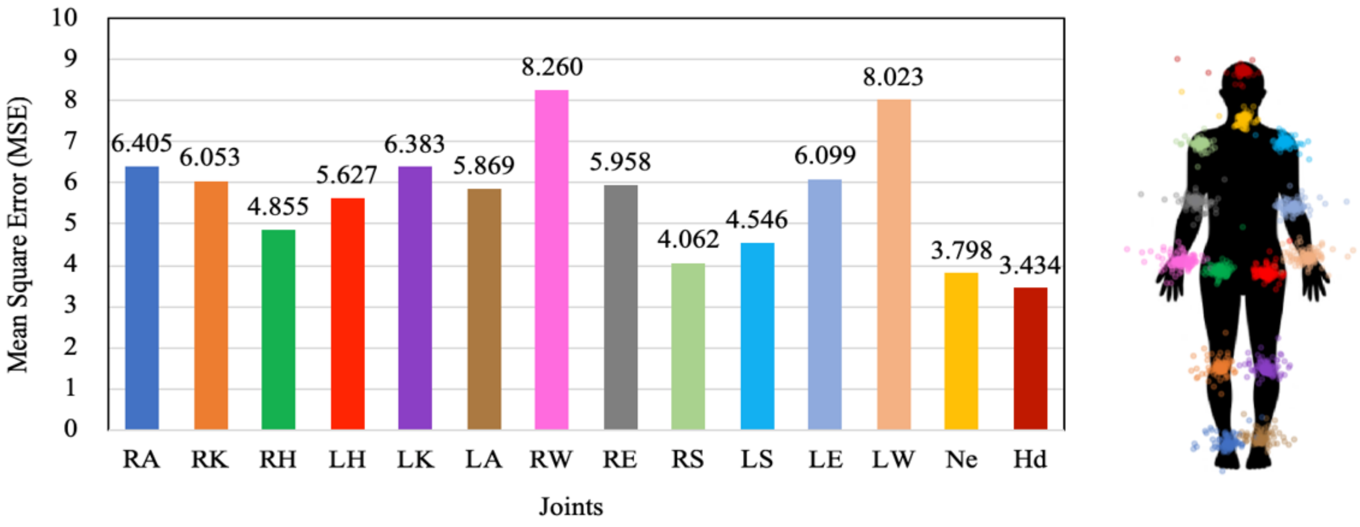


Figure 8. Mean squared errors (MSEs) of the estimated joints with the test dataset.

The further a joint is from the torso, the higher the MSE tends to be. The head (Hd) and neck (Ne) had the lowest MSEs, and the wrists (RW and LW) had the highest MSEs. Four coordinates (RH, LH, RS, and LS) connected to the torso had lower MSEs than the other remaining joints. The MSEs of the knees (RK and LK) were higher than those of the ankles (RA and LA), and the MSEs of the elbows (RE and LE) were higher than those of the wrists (RW and LW). This is because the wrists and ankles had a larger and more varied range of movement than did the torso. Therefore, for joints connected to the torso, because the motion radius is relatively small, the estimation error of the model was low.

The estimated joints are visualized around a set of fixed joints on the left side of Figure 8. Each dot represents the estimated joint location of individual joints. First, estimation errors for individual joints were measured for the test dataset as two-dimensional vectors. Then, these error vectors were plotted relative to the set of fixed joints, as shown on the left side of Figure 8. The colors of the estimated dots were identical to the respective colors of the joints shown in the graph. Estimation errors for the torso were smallest, and the errors increased with the joint distance from the torso. As with the MSE results, there was a greater variance in the coordinates of the right and left wrists.

4.3. Accuracy Analysis of the Estimated Body Parts

The accuracy of each body part was analyzed using the estimated joint coordinates. The estimation accuracies for body parts measured by PCP exhibited similar patterns as Figure 9. Each color in the graph matches the body part's color in Figure 3. A higher PCP represents more accurate estimates of the shape and location of the body parts. The mean PCP for each body part, including all activities, is shown in Figure 9. The PCP of the torso was highest (0.93). The body parts directly connected to the torso (the upper arms and legs) had relatively high PCPs (0.84 and 0.87, respectively). This is because the joints close to the trunk were more accurately estimated, as shown in Figure 8.

The lower arms, which are further from the torso, had the lowest PCP (0.77). This is because the estimated joint accuracy of both wrists in the lower arms in Figure 8 was low. However, the lower legs had a PCP value of 0.89, which was 0.02 higher than that of the upper legs. The distance from the knee to the ankle may appear longer than that from the hip to the knee in a sitting position depending on the image angle, as the depth information in the two-dimensional image is difficult to train. As a result, the PCP area was larger for the lower legs (i.e., from the knee to the ankle) due to this dimensional error. In addition, the PCPs of the upper and lower arms were lower than those of the legs, indicating that legs are easier to localize than arms.

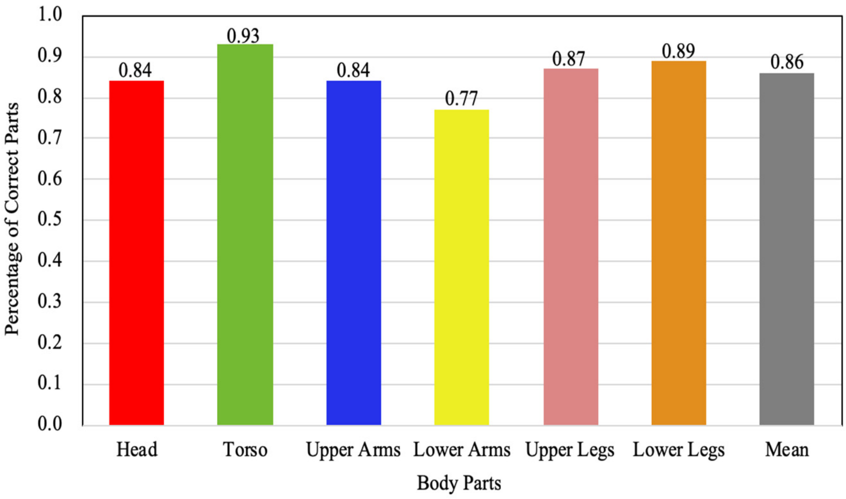


Figure 9. The mean PCPs of each body part.

To comprehend the PCP result of body parts more precisely, the estimation accuracies were further analyzed based on indoor activities. Table 2 shows PCPs calculated for the individual activities. The estimation accuracy of each activity can be expressed in terms of its mean PCP. The activities that involve relatively consistent poses exhibit high estimation accuracies. For example, the standing relaxed, typing, and walking activities had notably high PCP values of 0.92, 0.92, and 0.91, respectively, because these activities involve postures with less movement. A common characteristic of these activities is that the torso, as seen in the collected image dataset, is in an almost vertical orientation. Furthermore, in general, the movements of other body parts, such as lower arms and lower legs, do not change significantly.

Table 2. PCPs of body parts by indoor activities.

Body Parts Activities	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs	Mean
Sleeping	0.76	0.90	0.81	0.62	0.86	0.83	0.80
Reclining	0.88	0.94	0.82	0.71	0.88	0.94	0.86
Seated.quiet	0.88	0.88	0.82	0.79	0.88	0.94	0.87
Standing.relaxed	0.82	1.00	0.91	0.85	0.97	0.97	0.92
Reading.seated	0.94	0.94	0.82	0.82	0.88	0.91	0.89
Writing	0.82	0.94	0.70	0.68	0.74	0.85	0.80
Typing	0.94	0.94	0.94	0.82	0.94	0.94	0.92
Filing.seated	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Filing.stand	0.71	0.88	0.65	0.68	0.71	0.65	0.71
Walking about	0.82	0.94	0.91	0.88	0.97	0.94	0.91
Mean	0.84	0.93	0.84	0.77	0.87	0.89	0.86

However, the mean PCPs of sleeping (0.80), writing (0.80), and filing standing (0.71) were lower than those of the above three activities by more than 0.1. The reason for this is that these activities involve more diverse human poses and larger movements of body parts than the above three activities with high PCPs. In this study, activities that mostly involve standing or sitting motionlessly have high accuracy in estimating body parts. On the other hand, accuracy estimates for body parts were relatively low for activities involving large ranges of arm or leg motions.

Figure 10 shows examples of activities estimated using the proposed model. Estimated body parts are shown for some images in the test dataset. The torso, which had the highest PCP of the body parts listed in Table 2, was accurately estimated for all activities. By contrast, the positions of the lower arms, which had low PCPs, as shown in Table 2, were more difficult to estimate. For example, the bottom row, of “Filling stand” in Figure 10, shows a larger error for estimating the right arm, which is extended further to pick up an object. Interestingly, in case of the legs, the position of the joints is not constant, due to various sitting positions, but they are estimated with high accuracy.

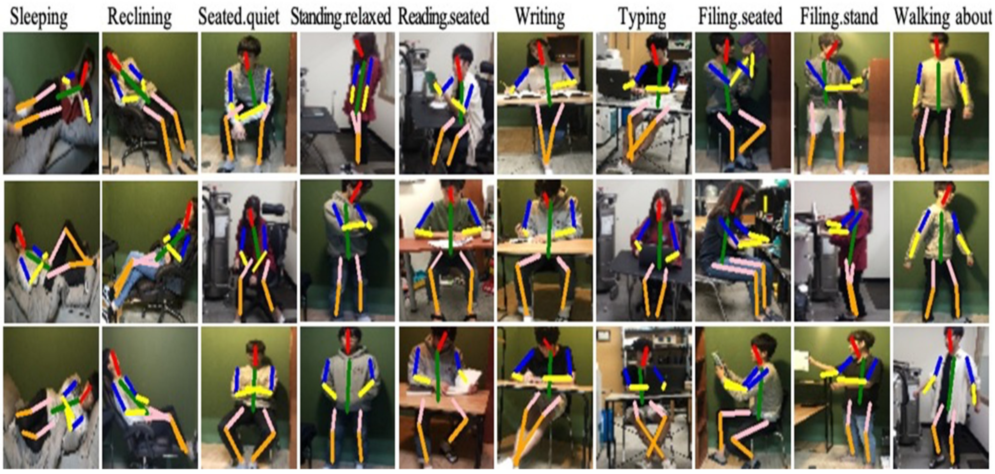


Figure 10. Examples of estimated body parts of test dataset.

Accurate estimation of the body parts position allows you to estimate a person's indoor activity [29]. Estimates of indoor activities can be used as information for an indoor environment and system control in various building types such as offices, houses, and hospitals. In particular, it is possible to provide information on the occupant indoor activities, which is essential for the environment control of systems, energy, health, and safety in spaces involving long periods of time spent indoors. As a result, classifying indoor activities makes it possible to create personalized environments for individuals.

5. Conclusions

In this study, a DNN model for estimating joints location in various indoor activities was developed and analyzed. The model was trained with images of indoor activities and estimated human joint coordinates. The accuracy of the proposed model was then assessed, and the following conclusions could be drawn.

1. The proposed DNN uses a large number of layers to learn complex human poses images effectively. Shortcut connections in the residual block make it possible to efficiently train the model within only 100 epochs. In the first stage of the model, residual blocks are connected sequentially to learn progressively more complex features of indoor activities. In the second stage, 14 branches of residual blocks independently estimate 14 individual joints, encouraging fine-tuning of joint estimation.
2. The accuracy of joint estimation indicated that the MSEs tended to increase as the joint's distance from the torso increased. The MSE of the neck was lowest, while that of the left wrist was highest. The MSEs of the arms were higher than those of the legs. Because the arms exhibit more diverse poses, the range of movement is large; by contrast, legs exhibit smaller variations in poses, which results in a lower estimation error.
3. PCPs were calculated for body parts. The PCP of the torso was highest (0.93). The PCPs of the arms were low because the range of movement was larger than that of the legs. The lower legs had larger PCP values than the upper legs for sitting activities. Activity estimation accuracy varied for different activities. While the overall average PCP for the 10 activities was 0.86, the PCPs of individual activities ranged from 0.77 to 0.93. Accuracies were higher for relatively still activities but lower for activities involving wide ranges of arm or leg motions.

In this study, it was confirmed that joint coordinates can be estimated from indoor activity images by the developed DNN model, which could be applied for indoor activity prediction. The identification of indoor activity is applied to the metabolic rate estimation, which enables PMV-based thermal environment control. As a result, it not only provides a comfortable thermal environment, but also improves the health and quality of life of the occupants.

The developed model is currently trained with only 10 types of indoor activities; follow-up studies, however, will be conducted to expand the range of indoor activities and increase the size of the dataset in order to estimate various activities that might occur in the real environment. In addition, the DNN model can be re-trained with activity image data of occupants in the actual building, so that a personalized environment can be provided by adapting to the new environment.

In order to improve the estimation performance of the current model, which instantly outputs joint information from the image, an activity determination algorithm [18] will be combined with the current model. The algorithm determines the representative activity and the activity intensity by computing the frequency of the estimation outputs from the model over a certain period of time. From this, the performance of estimating the metabolic rate in the actual building can be improved by compensating errors from the model and recognizing the intensity of activity. Another interesting direction for the future work would be to integrate an object detection model with the DNN model to improve the estimation accuracy by recognizing objects and inferring the human pose obscured by objects. Moreover, continuous efforts for model improvement should be carried out such as using information on three-dimensional images or estimating activities of multiple people. By supplementing these techniques for the performance enhancement of the model, it will be improved into an advanced model with high applicability to actual buildings.

Author Contributions: Conceptualization, E.J.C., J.W.M., J.-h.H. and Y.Y.; methodology, J.-h.H. and Y.Y.; software, E.J.C., J.-h.H. and Y.Y.; validation, E.J.C., J.W.M., J.-h.H. and Y.Y.; formal analysis, E.J.C., J.-h.H. and Y.Y.; investigation, E.J.C., J.W.M., J.-h.H. and Y.Y.; resources, J.W.M. and Y.Y.; data curation, E.J.C. and J.-h.H.; writing—original draft preparation, E.J.C. and J.-h.H.; writing—review and editing, E.J.C., J.W.M., J.-h.H. and Y.Y.; visualization, E.J.C. and J.-h.H.; supervision, J.W.M. and Y.Y.; project administration, J.W.M. and Y.Y.; funding acquisition, J.W.M. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2019R1A2C1084145), and by Incheon National University Research Grant in 2018.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: https://dbcollection.readthedocs.io/en/latest/datasets/leeds_sports_pose.html, http://people.csail.mit.edu/torralba/shortCourseRLOC/.

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

References

1. Tham, K.W.; Willem, H.C. Room air temperature affects occupants' physiology, perceptions and mental alertness. *Build. Environ.* **2010**, *45*, 40–44. [CrossRef]
2. Wyon, D.P. The effects of indoor air quality on performance and productivity. *Indoor Air* **2004**, *14*, 92–101. [CrossRef]
3. Frontczak, M.; Schiavon, S.; Goins, J.; Arens, E.; Zhang, H.; Wargocki, P. Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design. *Indoor Air* **2012**, *22*, 119–131.
4. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). *Ventilation and Acceptable Indoor Air Quality in Low-Rise Residential Buildings*; ASHRAE Standard 62.2; ASHRAE: Atlanta, GA, USA, 2003.
5. Fanger, P.O. *Thermal Comfort*; Danish Technical Press: Copenhagen, Denmark, 1970.
6. Lechner, N. *Heating, Cooling, Lighting: Sustainable Design Methods for Architects*, 4th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; p. 67.
7. Lee, K.; Choi, H.; Kim, H.; Kim, D.D.; Kim, T. Assessment of a Real-Time Prediction Method for High Clothing Thermal Insulation Using a Thermoregulation Model and an Infrared Camera. *Atmosphere* **2020**, *11*, 106. [CrossRef]

8. Pavlin, B.; Pernigotto, G.; Cappelletti, F.; Bison, P.; Vidoni, R.; Gasparella, A. Real-time monitoring of occupants' thermal comfort through infrared imaging: A preliminary study. *Buildings* **2017**, *7*, 10. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005. [CrossRef]
10. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006. [CrossRef]
11. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 25 June 2014. [CrossRef]
12. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef] [PubMed]
13. Johnson, S.; Everingham, M. Leeds Sports Pose Dataset. 2019. Available online: https://dbcollection.readthedocs.io/en/latest/datasets/leeds_sports_pose.html (accessed on 28 January 2020).
14. Ramanan, D. Learning to Parse Images of Articulated Bodies. *Proc. Adv. Neural Inf. Process. Syst.* **2007**, *1*, 7.
15. Yang, Y.; Ramanan, D. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2878–2890. [CrossRef] [PubMed]
16. Chen, X.; Yuille, A.L. Articulated pose estimation by a graphical model with image dependent pairwise relations. *arXiv* **2014**, arXiv:1407.3399.
17. Choi, E.J.; Yoo, Y.; Park, B.R.; Choi, Y.J.; Moon, J.W. Development of Occupant Pose Classification Model Using Deep Neural Network for Personalized Thermal Conditioning. *Energies* **2020**, *13*, 45. [CrossRef]
18. Park, B.R.; Choi, E.J.; Choi, Y.J.; Moon, J.W. Accuracy Analysis of DNN-Based Pose-Categorization Model and Activity-Decision Algorithm. *Energies* **2020**, *13*, 839. [CrossRef]
19. Dollár, P.; Tu, Z.; Perona, P.; Belongie, S. *Integral Channel Features*; BMVC Press: London, UK, 2009; pp. 91.1–91.11. [CrossRef]
20. Fei-Fei, L.; Fergus, R.; Torralba., A. Recognizing and Learning Object Categories, ICCV 2009 Short Course. 2009. Available online: http://people.csail.mit.edu/torralba/shortCourseRLOC/ (accessed on 28 January 2020).
21. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, 1097–1105. [CrossRef]
23. Moon, G.; Yong Chang, J.; Mu Lee, K. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5079–5088.
24. Iskakov, K.; Burkov, E.; Lempitsky, V.; Malkov, Y. Learnable triangulation of human pose. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7718–7727.
25. He, K.; Xiangyu, Z.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
26. ANSI/ASHRAE. *ANSI/ASHRAE Standard 55 Thermal Environmental Conditions for Human Occupancy*; ASHRAE: Atlanta, GA, USA, 2010.
27. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, arXiv:1212.5701.
28. Eichner, M.; Marin-Jimenez, M.; Zisserman, A.; Ferrari, V. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *Int. J. Comput. Vis.* **2012**, *99*, 190–214. [CrossRef]
29. Choi, E.J.; Park, B.R.; Choi, Y.J.; Moon, J.W. Development of a Human Pose Classifying Model to Estimate the Metabolic Rate of Occupant. *KIEAE J.* **2018**, *18*, 93–98. [CrossRef]