Detection of sitting posture using hierarchical image composition and deep learning

Audrius Kulikajevas[1], Rytis Maskeliunas[1] and Robertas Damaševičius[2,3]

[1] Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania
[2] Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania
[3] Faculty of Applied Mathematics, Silesian University of Technology, Gliwice, Poland

ABSTRACT

Human posture detection allows the capture of the kinematic parameters of the human body, which is important for many applications, such as assisted living, healthcare, physical exercising and rehabilitation. This task can greatly benefit from recent development in deep learning and computer vision. In this paper, we propose a novel deep recurrent hierarchical network (DRHN) model based on *MobileNetV2* that allows for greater flexibility by reducing or eliminating posture detection problems related to a limited visibility human torso in the frame, i.e., the occlusion problem. The DRHN network accepts the RGB-Depth frame sequences and produces a representation of semantically related posture states. We achieved 91.47% accuracy at 10 fps rate for sitting posture recognition.

INTRODUCTION

Machine learning and deep learning has shown very good results when applied to various computer vision applications such as detection of plant diseases in agriculture (*Kamilaris & Prenafeta-Boldú, 2018*), fault diagnosis in industrial engineering (*Wen et al., 2018*), brain tumor recognition from MR images (*Chen et al., 2018a*), segmentation of endoscopic images for gastric cancer (*Hirasawa et al., 2018*), or skin lesion recognition (*Li & Shen, 2018*) and even autonomous vehicles (*Alam et al., 2019*).

As our daily life increasingly depends on sitting work and the opportunities for physical exercising (in the context of COVID-19 pandemic associated restrictions and lockdowns are diminished), many people are facing various medical conditions directly related to such sedentary lifestyles. One of the frequently mentioned problems is back pain, with bad sitting posture being one of the compounding factors to this problem (*Grandjean & Hünting, 1977*; *Sharma & Majumdar, 2009*). Inadequate postures adopted by office workers are one of the most significant risk factors of work-related musculoskeletal disorders. The direct consequence may be back pain, while indirectly it has been associated with cervical disease, myopia, cardiovascular diseases and premature mortality (*Cagnie et al., 2006*). One study (*Alberdi et al., 2018*) has demonstrated that body posture is one of the best predictors of stress and mental workload levels. Another study linked postural instability and gait difficulty with a rapid rate of Parkinson's disease progression

(*Jankovic et al., 1990*). Posture recognition is also relevant for disabled people (*Ma et al., 2020*) and elderly people for proper health diagnostics (*Chen et al., 2018b*) as the sedentary behavior has a negative effect on people's wellbeing and health. Therefore, the solutions that would improve the daily living conditions of both healthy and spine pain affected people in the context of assisted living environments (*Maskeliunas, Damaševičius & Segal, 2019*).

While there are existing classical approaches for human posture prediction, unfortunately, they generally assert that entire human skeleton is visible in frame. Even though those assumptions of scene composition are valid, with everyone moving to their home offices, meeting them is simply not feasible. Not everyone is capable of having complex multi-camera setups to track their body posture. For this reason, there is a need for a solution that is able to inform the end-user (or their care provider) of their bad posture with cheaply available consumer sensors in order to improve their wellbeing without real-time supervision. With the renaissance of machine learning, and its application in computer vision tasks, we are able to solve a lot of complex tasks using black box models by shifting the majority of computations from the end-user device into the training stage. For this reason, artificial neural networks have been used in wide variety of applications. In this article, we propose a novel deep recurrent hierarchical neural network approach for tracking human posture in home office environments, where majority of the person sitting at the desk and therefore is occluded from the camera. Additionally, a pilot of 11 test subjects is made in order to validate our approach effectiveness.

RELATED WORK

The existing solutions such as orthopedic posture braces may not be viable solution due to other underlying medical conditions. Computer-aided posture tracking combined with behaviour improvement techniques due to continuous monitoring and self-assessment can contribute to remedy this issue (*Dias & Cunha, 2018*). The most prominent solution to this problem is skeleton based posture recognition (*Jiang et al., 2015*) using commercially available depth sensors such as *Microsoft Kinect* (*Zhang, 2012*) and *Intel Realsense* (*Keselman et al., 2017*). However, these solutions generally depend on some assertions, i.e., camera calibration settings, lightning conditions, expected posture (*Hondori & Khademi, 2014*), often making the results unreliable.

For identifying inadequate posture wearable textile sensors (*García Patiño, Khoshnam & Menon, 2020*), inertial and magnetic sensors attached to human body (*Bouvier et al., 2015*), depth cameras (*Ho et al., 2016*), radio-frequency identification tags (*Saab & Msheik, 2016*), 3D motion analysis (*Perusquía-Hernández et al., 2017*), video surveillance (*Afza et al., 2021*), Kinect sensors (*Ryselis et al., 2020*) and sensors attached to office chairs (*Zemp et al., 2016*; *Bibbo et al., 2019*) were used, while registering body posture parameters such as forward inclination, distance to the computer, and relative skeletal angles. However, the camera-based systems have demanding requirements for distance, proper lighting, calibration and non-occlusion.

Another approach focuses on wiring sensors directly to the human body to acquire data, although it limits the freedom of movement for work activities (*Arnold et al., 2020*). Despite these achievements, it is still quite difficult to recognize posture in real-time or correctly identify transitional activities in real-world environments (*Nweke et al., 2019*) as the recognition of fine-grained activities such as correct or incorrect cases of sitting postures is still a problem (*Chin et al., 2019*).

Currently, the state of the art in non-invasive posture tracking is depth and image processing (*Abobakr, Hossny & Nahavandi, 2018*; *Matthew et al., 2019*; *Camalan et al., 2018*). For example, *Huang & Gao (2019)* reconstruct realistic 3D human poses using the 3D coordinates of joint points captured by the depth camera and employ conformal geometric algebra to improve human limb modelling. *Li, Bai & Han (2020)* used OptiTrack and Kinect v2 to get and transfer data into a human skeleton coordinates. They used random forest regression learn the joint offset regression function, and adjust the skeleton based on the predictions on joint offset. Finally, as a result, they can determine the motions based on predicted posture. *Liu et al. (2020)* suggested 3D Convolutional Neural Network (CNN), called 3D PostureNet, while Gaussian voxel modeling is adopted to represent posture configurations. The method allows to eliminate the coordinate deviations caused by various recording environments and posture displacements. *Pham et al. (2019)* exploit Deep CNNs based on the DenseNet model to learn directly an end-to-end mapping between the input skeleton sequences and their action label for human activity recognition. The network learns spatio—temporal patterns of skeletal movements and their discriminative features for further downstream classification tasks. *Sengupta et al. (2020)* detect skeletal joints using mmWave radar reflection signals. First, the reflected radar point cloud. Next, CNN was trained on radar-to-image representation and used to predict the position of the skeletal joints in 3-D space. The method was evaluated in a single person scenario using four primary motions. The method has shown to be robust in adverse weather conditions and deviations in light conditions.

However, none of the above mentioned methods are applicable when only the upper part of the body is visible. Some methods tried to tackle this problem by exploiting the temporal relationship between the body parts to deal with the occlusion problem and to get the occluded depth information (*Huang, Hsu & Huang, 2013*) or by recreating a topological character (*Bei et al., 2017*), yet they still require a recreation of a full body skeleton.

To address this problem, we propose a novel approach for human posture classification by using a supervised hierarchical neural network (*Liu et al., 2019*) that uses the RGB-Depth data as input. Our method extends *MobileNetV2* (*Sandler et al., 2018*) neural network to include the recurrent layers. This allows the network to learn from a sequence of images by adding the dimension of time to our feature list. Allowing the network to use the context of what happened in previous frames to make predictions. This is an improvement over existing methods for skeleton prediction as this allows for our approach to predict user posture in more complex environments, for example, when a person is sitting in front of an office desk thus a large portion of his/her body is occluded.
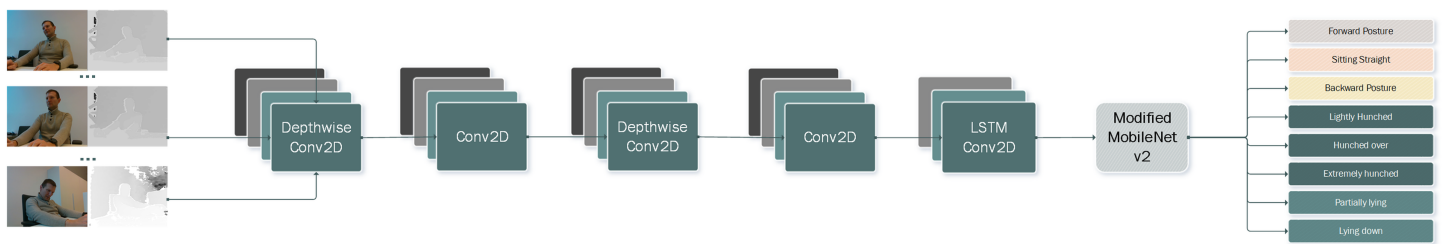
Figure 1 Our recurrent hierarchical ANN architecture using *MobileNetV2* as the main backbone. It takes the RGB-D frame sequence as input and outputs the flattened prediction tree as a result. Full-size ☐ DOI: 10.7717/peerj-cs.442/fig-1

Such position would cause other known skeleton-based posture prediction methods to fail, due to lack of data provided by the sensors to infer the full human skeleton.

Our novelty is the use only a simple depth camera, so the subject does not need to wear any sensors on their body nor have entire body visible in sensor field of view. In fact, only the upper 30% of the body may be visible, whereas when using the Kinect style sensor, the lower legs must be visible or generated artificially to allow the reconstruction of the skeleton or, otherwise, the recognition process fails. Our approach does not rely on a (visual or artificial) reconstruction of the full skeleton and, thus, allows for the detection of posture in advanced scenarios such as sitting at a desk, where a camera often receives very limited information.

METHODS

Network architecture

Our preliminary analysis has shown that it is very hard to predict human posture based on a single shot. For this reason, we opted to use time sequences with $n = 4$ frames. However, during training the input has a variable length of $1 \le n \le 4$, with each frame being about a second apart to reduce the dependence on the previous frames.

We selected to use deep convolutional recurrent neural networks for they have shown some of the best capabilities when it comes to similar tasks requiring to predict sequences of data as with natural speech recognition (*Sundermeyer, Ney & Schlu, 2015*; *Graves, Mohamed & Hinton, 2013*) or even traffic prediction (*Ji & Hong, 2019*).

The input of our network architecture (Fig. 1) is the RGB-D frame sequence that is fed into depth-wise convolutional block (*Zhang et al., 2019*), which reduces the dimensionality of each frame by a factor of two, without losing each individual channel's influence on the output. This is due to depth-wise convolutions applying separate kernels for each channel. This is then followed by a convolutional layer in order to extract the best individual features, which is followed by a second dimensionality reduction layer. We do this because our input frames are captured at 640×480 px resolution, which is the maximum hardware resolution of the *Intel Realsense D435i* device. Reducing the dimensionality twice leaves us with 128 features, each of 160×120 px resolution. At this stage, we use Long Short Term Memory (LSTM) convolutional block (*Xu et al., 2020*; *Li et al., 2020*), which is tasked to extract 32 most useful features in the entire sequence.
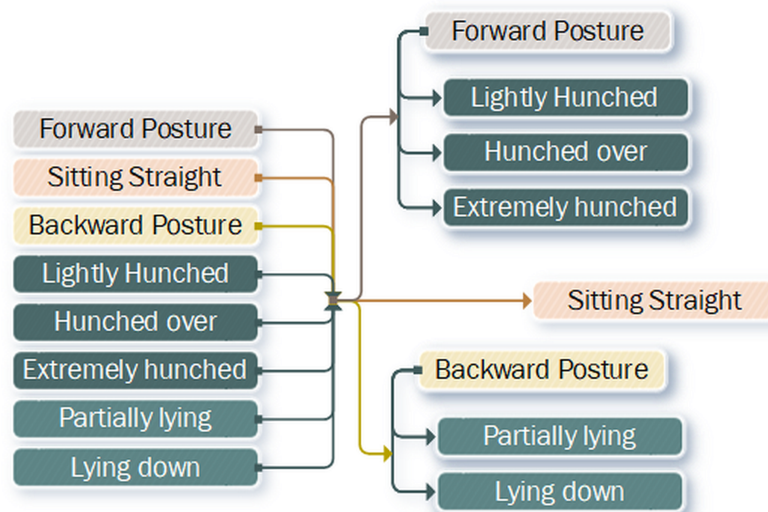
Figure 2 Flattened hierarchy representation of postures expanded into a hierarchical tree.
Full-size ⊡ DOI: 10.7717/peerj-cs.442/fig-2

For our main neural network backbone, we use *MobileNetV2*, which is the extension of *MobileNet*, for it has show to achieve great results in predictive capabilities (*Howard et al., 2017*; *Zhou et al., 2019*), however, the architecture itself is relatively light-weight for it is designed to be used in low power devices such as mobile devices, unlike for example, *YOLOV3*, which while having impressive recall results (*Redmon & Farhadi, 2018*), is much more complex and has a substantially poorer performance. The *MobileNetV2* output is then connected to a global average pooling layer in order to reduce dimensionality and improve generalization rate (*Zhou et al., 2016*). Finally, the output is subsequently connected to a fully-connected layer, which represents the flattened representation of posture state prediction hierarchy, which can be seen in Fig. 2.

Our entire ANN setup can be seen in Table 1. After each of two bottleneck layers we additionally use spatial dropout layers as it is shown to improve generalization during training and reduce the effect of nearby pixels being strongly correlated within the feature maps (*Murugan & Durairaj, 2017*), each with dropout probability of 0.2, whereas the spatial dropout post LSTM cell has a dropout probability of 0.3, because the higher the network is upstream the more dropout layers influence the entire network, therefore high values upstream may cause the network to be more unstable and difficult to train. The dropout layer before the output layer however, has a dropout probability of 50%. Aggressive dropout values reduce the chance that the model is will overfit by training on noise instead of image features. All our previous layers up to this point had used Rectified Linear Unit as our activation function in order to impose non-linearity into our model for it has shown better results and improved performance due to its mathematical simplicity in CNNs (*Hanin, 2019*). However, for the last layer we opted to use the *sigmoid* activation due to our network outputting hierarchical values and acting as multi-label classifier, while the *softmax* activation is more useful for multi-class classification tasks.

Kulikajevas et al. (2021), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.442

5/20

Table 1 Layers of the proposed neural network architecture for human posture recognition.

Type	Filters	Size	Output
Input	–	–	$t \times 640 \times 480$
Depthwise convolution	–	$11 \times 11/2$	$t \times 320 \times 240$
Convolution	64	1×1	$t \times 320 \times 240$
Spatial dropout $P(x) = 0.2$	–		$t \times 320 \times 240$
Depthwise convolution	–	$5 \times 5/2$	$t \times 160 \times 120$
Convolution	128	1×1	$t \times 160 \times 120$
Spatial dropout $P(x) = 0.2$	–		$t \times 160 \times 120$
LSTM convolution	16	3×3	160×120
Spatial dropout $P(x) = 0.3$	–		160×120
MobileNetV2	–	–	4×5
Global average pooling	–	–	1,280
Dropout $P(x) = 0.5$	–		1,280
Fully-connected (sigmoid)	–	–	8

Algorithm of sitting posture detection

Figure 3 depicts algorithm of our enhanced posture detection solution. Process starts with the initialization of model weights for sorting out the RGB-D camera output (both depth and RGB as varying on the condition either modality can provide compensation features). Algorithm then tries to reconstruct intermediate frame for retrieval and analysis of frame semantics, which are then used for stack validity status verification. Assuming the condition, analysis starts in the recurrent layers of our modified MobileNet v2 architecture, with Pareto-Optimal berparameter optimization (*Plonis et al., 2020*). The model then assigns prediction labels and algorithm further tries to improve the quality by firing smart semantic prediction analyzer, checking not only the output value but probable output status for a combined confidence level of <40%, as an improved determinator for further frame semantic analysis. A final validity status is then initiated, depending on condition leading to a majority voting and a very reliable detection of bad posture. Algorithm was designed to work continuously and is able to automatically stop processing to stay compatible with green computing paradigm (*Okewu et al., 2017*) to save energy.

Prediction of posture states

We adopted the semantic matchmaking approach (*Ruta et al., 2014*) to describe the semantic relationship between different postures using an ontological tree for analysis, reasoning and knowledge discovery. In order to extract the specific prediction label we parse the posture hierarchy tree (Fig. 2), first, by checking, which posture state is most likely according to the neural network. Once we know, which posture type is most likely to be represented in the frame sequence, we proceed to the sub-nodes and check their predictions. We continue this search until we find the leaf node, which represents the actual label. This approach allows us to filter out the most likely path that is seen in the frames. This is helpful in cases when the similarities between postures is large.
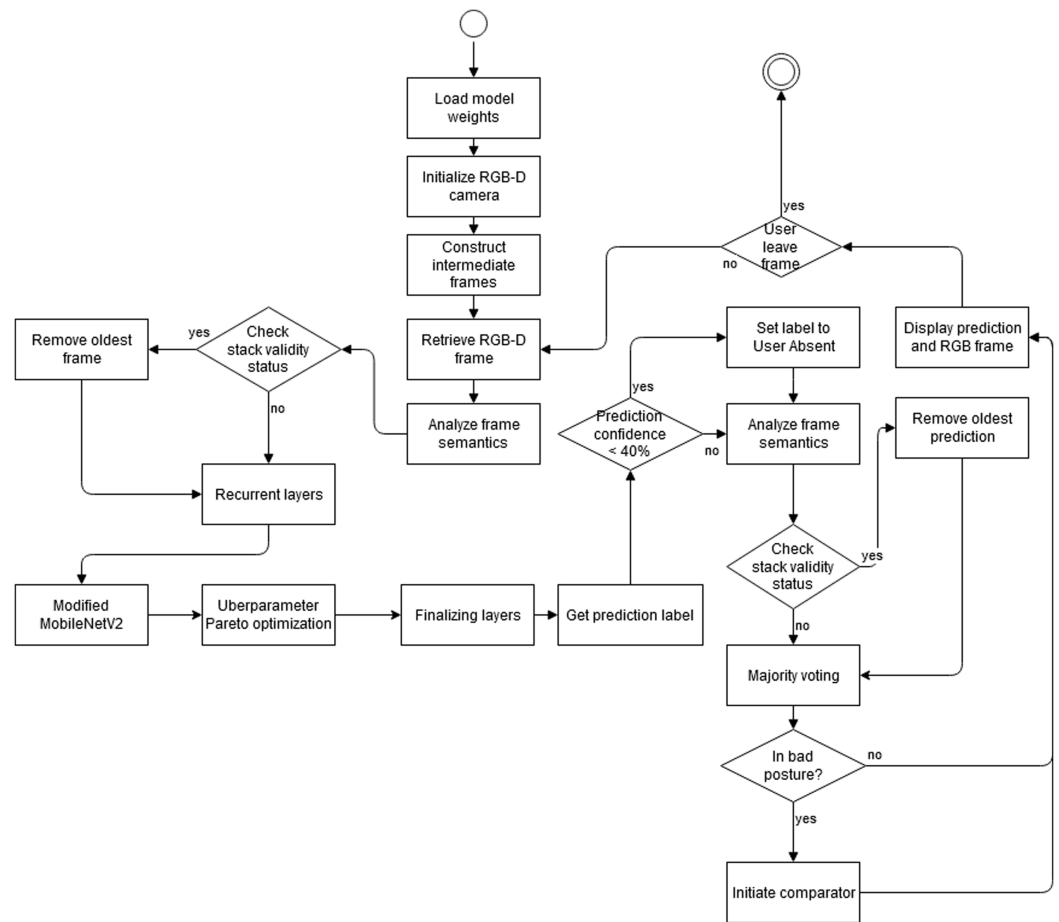
Figure 3 Activity diagram of the proposed method for sitting posture state recognition.
Full-size ◲ DOI: 10.7717/peerj-cs.442/fig-3

For example, all forward postures share the same characteristics—shoulders are not at 90 degree angle, and the head is positioned forward with respect to the body. This allows us to ignore all the weight influences, where for example, the person is lying down. Additionally, the further down the tree the label leaves are the less of an overall recall error is, due to each level of the tree being ontologically similar, for example, predicting *lying down*, instead of *partially lying down* is a smaller error than predicting *hunched over*.

Loss function

One of the reasons why we use the flattened final layer to represent our posture hierarchy is because we can represent our problem as multi-label classification (*Wang et al., 2019*). This allows us to use binary cross-entropy (Eq. (1)) in order to calculate the loss between expected output and ground truth:

$$H_p = \sum_{i=1}^{N} \hat{y}_i \cdot log(y_i) + (1 - \hat{y}_i) \cdot log(1 - y_i) \tag{1}$$

Binary cross-entropy classifier is fit for our multi-label classification task (*Wu et al., 2018*) as each of our cells output is a binary one and more than one cell can be positive at a time, depending on how deep the classification is, as opposed to the categorical cross-entropy, which is a solution for multi-class tasks, where the input can yield only a single-class output.

Network training

For training a neural network various optimization methods have been proposed. However, one of the most popular optimization methods due to its computational efficiency allowing training ANN on large datasets more easily on weaker hardware in addition to the ability to achieve faster convergence than other methods is Adam (*Kingma & Ba, 2015*). For these reasons, we had opted to use Adam for training using the initial training rate of $5e^{-4}$, with a batch size of *eight*.

Additionally, we perform data augmentation as it has shown to improve ANN generalization (*Fawzi et al., 2016*). We perform horizontal image flipping in order to increase the view count, and perform random hue and saturation changes, aiming to increase stability against different lightning conditions as all of our video sequence instances were recorded during same time frame, therefore, maintaining nearly identical lightning. In all cases, the identical augmentation values are used for all images in the same series with the same probability of performing image flipping, hue and saturation augmentation being 50% independently. Random hue shift is performed in the range of $h = [0, 2\pi)$ radians, while the saturation has the random range of $s = [0, 2]$.

Data collection

ANNs have the benefit of doing all the heavy work upfront during the training therefore, allowing to improve system runtime by reducing the number of required calculations (*Holden et al., 2019*). However, this approach depends on the quality of the training data, which can be defined in terms of the size of available samples, class balance and even the correctness of the labels. Our approach depends on both color and depth information. Unfortunately, still there are no publicly available labeled human posture dataset that additionally provides depth information. For our experiments, we have devised a methodology to create such dataset. The data collection procedure consists of *two* stages.

Stage I

The person starts by sitting up straight. This position is then filmed for 30 s. Afterwards, the person is instructed to lightly hunch forward, which is followed by another 30 s of sitting in this position. Afterwards, the person is again instructed to hunch more, emulating their average hunching posture. After the filming, the subject is instructed once more to hunch forward in order to emulate the worst possible forward posture. Once the 30 s have been recorded in this posture, the person is then instructed to sit up straight for an additional 30 s to get used to this position. Then, we start emulating the bad backwards posture, i.e., lying down in the chair. The person is instructed to

partially lie in the chair for 30 s, after which he/she is instructed to do it twice more in increasingly bad posture positions giving us three sets of bad forward and backward posture examples.

Stage II

The person is instructed to initially sit up straight. Then the person is instructed to start slowly counting from *one* to *five*, while slowly worsening their forward posture. When the person finishes counting, he/she is expected to be in the worst forward posture they imagine. Afterwards, the subject returns to straight position. This action is repeated for *five* times. Once the forward posture data is recorded, the person is asked to perform the similar action, this time with backwards posture, where once again when they finish counting, they are fully lying in the chair.

Each of the stages is recorded three separate times using different camera perspectives at *10 o'clock, 12 o'clock and 3 o'clock*. The person is filmed in front of the computer desk and during the filming they are asked to interact with the table in their current posture how they imagine they would sit on the table. This can range from drawing on a piece of article, to checking the drawers, using keyboard or even holding their head with their hand.

When collecting our dataset, we asked 11 subjects (seven men and four women) to perform the posture emulation tasks. The informed consent was obtained, while we followed strictly the requirements of the Helsinki declaration. The research was approved by the Institutional Review Board, Faculty of Informatics, Kaunas University of Technology (2018-09-24 No. IFEP201809-2). Further expansion of dataset to include different body types or disabilities may additionally improve the results in more real world cases.

Data labelling

Once the data is collected it must be labeled manually. However, one of the issues when labeling data we have noticed that has caused some of the data points to be thrown out completely is for a person to actually differentiate properly what posture that person is in. Even though the filming took in relatively discrete time intervals, some subjects may take longer/shorter to perform specified actions, they may attempt to *fix* their posture due to it being uncomfortable for them, etc. Additionally, some people have indiscernible *sitting straight* and *lightly hunched* posture, as their normal posture is already biased towards leaning head forward. Therefore, the labeling of such data is a challenge due to its subjectiveness as bad data labels may poison the network and cause it to overfit instead of generalizing.

Using our recorded dataset, we have extracted these labels: *sitting straight*, *lightly hunched*, *hunched over*, *extremely hunched*, *partially lying* and *lying down*. While we have three backwards posture angles, we opted for only two backwards posture labels as it is difficult to objectively measure *lying down* and *extremely lying down* as in multiple cases subjects barely made any movements.

Table 2 Frame count in the dataset.

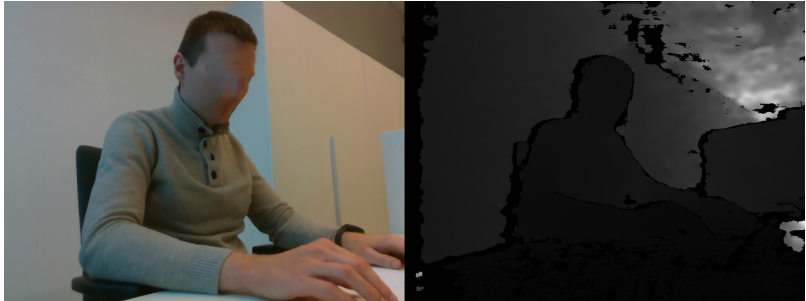| Posture class	Training	Testing	Dataset (%)
Sitting straight	3390	505	21.53
Lightly hunched	2230	200	14.16
Hunched over	2534	321	16.09
Extremely hunched	1918	182	12.18
Partially lying	2053	339	13.04
Lying down	3622	302	23.00

Dataset

Our dataset consists of 66 different captured video sequence instances totaling 133 min of recording, which we split into individual labeled frames. We used 10-fold cross-validation. For training, we had split the data from each individual in 90:10 ratio instead of splitting the frames, as this gives more objective results, because similar frames from the same captured video will not be a part of evaluation, thus artificially increasing the recall rate. We can see the number of frames in training and testing frames in Table 2, additionally we can see that dataset is slightly skewed towards *sitting straight* and *lying down* due to the dataset being not completely balanced. While class imbalance may cause issues in generalization of the network, we believe that the imbalance is not high enough to have a noticeable impact. Finally, the examples of images in the dataset are given in Table 3 (right side view). The subjects presented in the images have provided their informed consent for publication of identifying images in an online open-access publication.

RESULTS

Accuracy

We evaluate the prediction correctness against ground truth in two stages: using final prediction labels (*sitting straight, lightly hunched, hunched over, extremely hunched, partially lying, lying down*); and intermediate branch predictions (*forward posture, backward posture, straight posture*). This provides us better insight on the prediction results as this will show both absolute error and intermediate branch error. The confusion matrix for the first case can be seen in Fig. 4. In the first case, we achieved an accuracy of 68.33%, sensitivity of 0.6794, specificity of 0.9372 and *f*-score of 0.6789. Note that the network has achieved a high specificity rate, which means that it can effectively recognize the subjects, who do not have the posture problems. As we can see from the confusion matrix (Fig. 4), the biggest issues arise in prediction regarding *hunched over* and *extremely hunched* labels. The proposed network model had a hard time discerning between these two values. This indicates that either our dataset for these two labels have little variation and the positions are very similar, or that one of the labels has been mislabeled and has poisoned the predicted values. This suggests that further investigation in our dataset is definitely needed.

Table 3 Examples of images in dataset (right side view).

Posture class	RGB and Depth images
Sitting straight	
Lightly hunched forward	
Hunched over forward	
Extremely hunched forward	
Partially lying down in the chair	

(Continued)

Table 3 (continued)

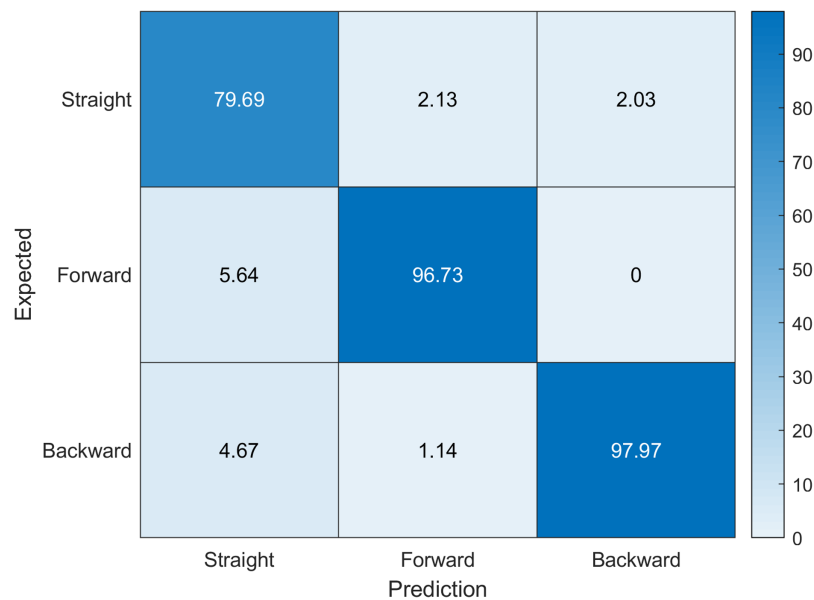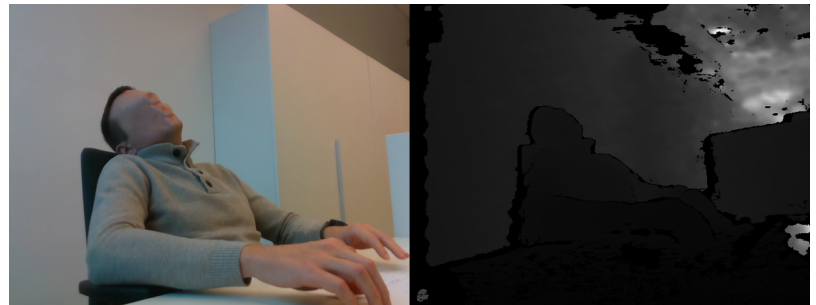| Posture class	RGB and Depth images
Lying down in the chair	

Figure 4 Confusion matrix indicating expected labels versus network predictions. Accuracy values are given in percents. Diagonal values indicate correct predictions.

Full-size ⬛ DOI: 10.7717/peerj-cs.442/fig-4

However, all largest misclassification values occur between neighbouring classes (*extremely hunched* vs *hunched over*—49.5%), (*hunched over* vs *extremely hunched*—40.66%), (*partially lying* vs *lying down*—28.15%), (*lying down* vs *partially lying*—19.47%), suggesting that perhaps the need for some fuzzification of class definitions and interpretation of results, or that these posture classes should be combined.

Our dataset depends on the expert interpretation of what they are seeing in the camera, which may be the cause of this disparity. Performing data labeling by more experts may improve the results as this would reduce the ambiguity in our dataset that we have due to a limited number of experts labeling the data. However, the network is accurate enough that it can suggest the labels in further labeling processing. This would change our solution from being supervised machine learning into semi-supervised or even completely
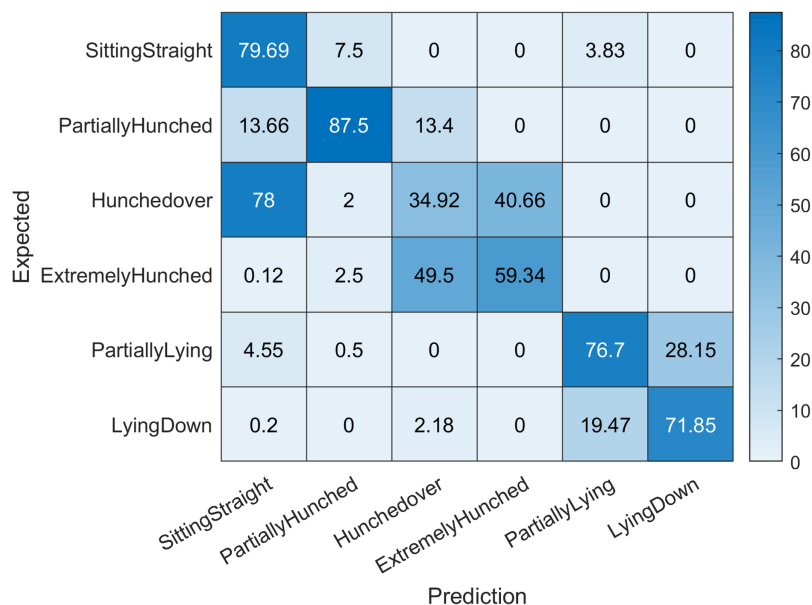
Figure 5 Confusion matrix indicating bottom level expected labels versus network predictions. Accuracy values are given in percents. Diagonal values indicate correct predictions.

unsupervised machine learning approach. Notwithstanding, this is beyond the scope of our research. However, if we investigate further we can see from Fig. 5 that the root posture prediction has better results, where the network model manages to generalize between *forward posture*, *backward posture* and *straight posture* cases.

This partial confusion matrix (Fig. 5) makes it clear that while some finer detail in our dataset is less objective and is difficult for the network to generalize, the neural network itself is adept in solving the classification of the base postures with the mean accuracy rate of 91.47% (sensitivity 0.9185, specificity 0.9595, *f*-score 0.9132 and kappa 0.8081). The bottom level (root) labels are more than enough in a lot of cases when it comes to posture recognition tasks that do not require precise user angle extraction. Additionally, when comparing partial and full confusion matrices we can see that the deeper levels additionally have lower false negative results, indicating the addition of the hierarchical structure for prediction can inherently improve the prediction results in deeper levels due to the semantic connections between labels.

Performance

Due to fact that our approach uses *MobileNetV2* as the backbone for our ANN that means that is lightweight and can be used in real-time applications. Our method performs a posture prediction on average in 94 ms (which corresponds to 10 fps rate) on a workstation with the following specifications: *Intel i7-4790* CPU with 16GB of RAM, *nVidia 1070* GPU with 8GB of GDDR5 VRAM.

Comparison

We compare our results with the results of other authors in Table 4.

Table 4 Comparison of posture recognition methods.

Method	Frame resolution, px	Frame rate, fps	Accuracy, %	Task	Reference
Real-time deformable detector	320 × 240	10	75.33	Hand posture recognition	*Hernandez-Belmonte & Ayala-Ramirez (2016)*
Ensemble of InceptionResNetV2	640 × 480	n/a	95.34	Four postures (standing, sitting, lying, and lying crouched)	*Byeon et al. (2020)*
LVQ (learning vector quantization) neural network	640 × 480	333	99.01	Five full-skeleton postures (standing, sitting, stooping, kneeling, and lying)	*Wang et al. (2016)*
Multi-stage convolutional neural network (M-CNN)	n/a	5	98.70	Two postures for fall detection	*Zhang, Wu & Wang (2020)*
LVQ neural network	48 × 16	10	99.95	Eight postures (stand, hand raise, akimbo, open wide arms, squat, toe touch, crawl, and lie)	*Gochoo et al. (2018)*
Deep CNN	24 × 8	9	99.99	26 yoga postures	*Gochoo et al. (2019)*
D CNN	n/a	n/a	98.16	Detection of 10 standstill body poses.	*Liu et al. (2020)*
Deep recurrent hierarchical network	640 × 480	10	91.47	Spine posture recognition while sitting	This paper

Note:
n/a, data is not available.

Our method allows to achieve the real time sitting posture recognition with the same or better recognition accuracy and video resolution than other similar state-of-the-art methods. For example, *Wang et al. (2016)* achieved higher accuracy and recognition rate, however their method require the visibility of full skeleton detected by Kinect sensors and not occluded by any obstacles. *Gochoo et al. (2018, 2019)* achieved a very high recognition rate using three low resolution thermal sensors placed around the subject to recognize eight postures and 26 yoga postures, respectively, but no occlusions were allowed either. *Tariq et al. (2019)* used Kinect and additional motion sensors from smartwatch to achieve the required level of accuracy.

DISCUSSION

The training of neural network depends on hardware used for recording. We used *Intel Realsense D435i*, but the results may be worse when using different hardware, for example, *KinectV2* as these two devices produce different noise in their depth fields. This may cause the network to have poorer results when compared to the one that it has been trained on. However, we are not able to validate this claim. Additionally, when testing the network using the real-time camera feed we had noticed that while relatively similar and their mirror image angles work it may have lower precision rates with something more extreme like placing sensor very high or very low relative to the table or user.

Finally, when using in real world application, one of the measures to improve prediction stability is to use majority voting on the preceding 10 video frames. This is performed by taking the prediction label that had appeared the most times in the previous recorded 10 frames. This technique can improve the stability of the predictions as a single video frame will no longer change the prediction results. However, the predictions will have a delay, due to previous video frames influencing the result for a short period of time.

Another limitation of this study is a small number of subjects (11), all healthy, which may have influenced the validity of the results. The age range and gender diversity of the subject group was limited. In future, we will have to extend the subject group to include various professional/occupational groups as well as school children and adolescents as well as people with different body types and disabilities in order to improve the results for real world cases.

CONCLUSION

We have proposed an extension of the *MobileNetV2* neural network, which allows the use of sequential video data as an input, therefore, allowing for the deep neural network to extract important temporal features from video frames, which would otherwise be lost when compared to a single-frame classification while still being capable of the single-frame prediction due to being biased towards the last frame. We have improved the top-layer of the *MobileNetV2* architecture by adding the hierarchical data representation, which acts as a semantic lock for top-level label classification by filtering out the invalid class labels early. Additionally, we have performed a pilot study based in which we suggest the methodology required to collect the training dataset and validation datasets. Further improvements in dataset collection methodology can be made in order to account for different body shapes, disabilities and removing labeling ambiguities. The proposed posture classification approach is highly extensible due to its flattened tree representation, which can be easily adapted to the already existing posture classification tasks with the depth of the ontological semantic posture model being one of the driving factors for classification quality. Based on our validation data giving us a classification accuracy of 91.47% in predicting three main sitting posture classes (*backward posture*, *forward posture* and *straight posture*) at a rate of 10 fps. Finally, unlike in related work, our method does not depend on the skeletal predictors, therefore we can perform the sitting human posture prediction when only as low as 30% of the human torso is visible in the frame. For these reasons, we believe that our approach is more robust for real-time human posture classification tasks in the real-world office environment.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Competing Interests

Robertas Damaševičius is an Academic Editor for PeerJ.

Author Contributions

- Audrius Kulikajevas conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.
- Rytis Maskeliunas conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Robertas Damaševičius analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

The research was approved by the Institutional Review Board, Faculty of Informatics, Kaunas University of Technology (No. IFEP201809-2).

Data Availability

The following information was supplied regarding data availability:

Data and code are available on GitHub:

https://github.com/realratchet/SitStraightNet

REFERENCES

Abobakr A, Hossny M, Nahavandi S. 2018. A skeleton-free fall detection system from depth images using random decision forest. *IEEE Systems Journal* **12(3)**:2994–3005 DOI 10.1109/JSYST.2017.2780260.

Afza F, Khan MA, Sharif M, Kadry S, Manogaran G, Saba T, Ashraf I, Damaševičius R. 2021. A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. *Image and Vision Computing* **106**:104090.

Alam F, Mehmood R, Katib I, Altowaijri SM, Albeshri A. 2019. TAAWUN: a decision fusion and feature specific road detection approach for connected autonomous vehicles. *Mobile Networks and Applications* **15(5)**:50 DOI 10.1007/s11036-019-01319-2.

Alberdi A, Aztiria A, Basarab A, Cook DJ. 2018. Using smart offices to predict occupational stress. *International Journal of Industrial Ergonomics* **67(3)**:13–26 DOI 10.1016/j.ergon.2018.04.005.

Arnold D, Li X, Lin Y, Wang Z, Yi W, Saniie J. 2020. Iot framework for 3d body posture visualization. *IEEE International Conference on Electro Information Technology* **2020**:117–120.

Bei S, Xing Z, Taocheng L, Qin L. 2017. Sitting posture detection using adaptively fused 3d features. In: *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference.* Piscataway: IEEE, 1073–1077.

Bibbo D, Carli M, Conforto S, Battisti F. 2019. A sitting posture monitoring instrument to assess different levels of cognitive engagement. *Sensors (Switzerland)* **19(3)**:455 DOI 10.3390/s19030455.

Bouvier B, Duprey S, Claudon L, Dumas R, Savescu A. 2015. Upper limb kinematics using inertial and magnetic sensors: comparison of sensor-to-segment calibrations. *Sensors* **15(8)**:18813–18833 DOI 10.3390/s150818813.

Byeon Y-H, Lee J-Y, Kim D-H, Kwak K-C. 2020. Posture recognition using ensemble deep models under various home environments. *Applied Sciences* **10(4)**:1287 DOI 10.3390/app10041287.

Cagnie B, Danneels L, Tiggelen DV, Loose VD, Cambier D. 2006. Individual and work related risk factors for neck pain among office workers: a cross sectional study. *European Spine Journal* **16(5)**:679–686 DOI 10.1007/s00586-006-0269-7.

Camalan S, Sengul G, Misra S, Maskeliunas R, Damaševičius R. 2018. Gender detection using 3d anthropometric measurements by kinect. *Metrology and Measurement Systems* **25(2)**:253–267.

Chen H, Dou Q, Yu L, Qin J, Heng P. 2018a. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage* **170**:446–455 DOI 10.1016/j.neuroimage.2017.04.041.

Chen Y, Yu L, Ota K, Dong M. 2018b. Robust activity recognition for aging society. *IEEE Journal of Biomedical and Health Informatics* **22(6)**:1754–1764 DOI 10.1109/JBHI.2018.2819182.

Chin LCK, Eu KS, Tay TT, Teoh CY, Yap KM. 2019. A posture recognition model dedicated for differentiating between proper and improper sitting posture with kinect sensor. In: *HAVE, 2019—IEEE International Symposium on Haptic, Audio-Visual Environments and Games, Proceedings.*

Dias D, Cunha JPS. 2018. Wearable health devices—vital sign monitoring, systems and technologies. *Sensors* **18(8)**:2414 DOI 10.3390/s18082414.

Fawzi A, Samulowitz H, Turaga D, Frossard P. 2016. Adaptive data augmentation for image classification. In: *2016 IEEE International Conference on Image Processing (ICIP).* Piscataway: IEEE, 3688–3692.

García Patiño A, Khoshnam M, Menon C. 2020. Wearable device to monitor back movements using an inductive textile sensor. *Sensors* **20(3)**:905.

Gochoo M, Tan T-H, Batjargal T, Seredin O, Huang S-C. 2018. Device-free non-privacy invasive indoor human posture recognition using low-resolution infrared sensor-based wireless sensor networks and DCNN. In: *2018 IEEE International Conference on Systems, Man, and Cybernetics.* Piscataway: IEEE.

Gochoo M, Tan T-H, Huang S-C, Batjargal T, Hsieh J-W, Alnajjar FS, Chen Y-F. 2019. Novel IoT-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning. *IEEE Internet of Things Journal* **6(4)**:7192–7200 DOI 10.1109/JIOT.2019.2915095.

Grandjean E, Hünting W. 1977. Ergonomics of posture—review of various problems of standing and sitting posture. *Applied Ergonomics* **8(3)**:135–140 DOI 10.1016/0003-6870(77)90002-3.

Graves A, Mohamed A, Hinton G. 2013. Speech recognition with deep recurrent neural networks. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* Piscataway: IEEE, 6645–6649.

Hanin B. 2019. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics* **7(10)**:992 DOI 10.3390/math7100992.

Hernandez-Belmonte U, Ayala-Ramirez V. 2016. Real-time hand posture recognition for human-robot interaction tasks. *Sensors* **16(1)**:36 DOI 10.3390/s16010036.

Hirasawa T, Aoyama K, Tanimoto T, Ishihara S, Shichijo S, Ozawa T, Ohnishi T, Fujishiro M, Matsuo K, Fujisaki J, Tada T. 2018. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* **21(4)**:653–660 DOI 10.1007/s10120-018-0793-2.

Ho ESL, Chan JCP, Chan DCK, Shum HPH, Cheung Y, Yuen PC. 2016. Improving posture classification accuracy for depth sensor-based human activity monitoring in smart

environments. *Computer Vision and Image Understanding* **148(3)**:97–110
DOI 10.1016/j.cviu.2015.12.011.

Holden D, Duong BC, Datta S, Nowrouzezahrai D. 2019. Subspace neural physics: fast
data-driven interactive simulation. In: *SCA '19: Proceedings of the 18th annual ACM
SIGGRAPH/Eurographics Symposium on Computer Animation.*

Hondori HM, Khademi M. 2014. A review on technical and clinical impact of microsoft kinect on
physical therapy and rehabilitation. *Journal of Medical Engineering* **2014**:1–16.

**Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H.
2017.** Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv.
Available at http://arxiv.org/abs/1704.04861v1.

Huang X, Gao L. 2019. Reconstructing three-dimensional human poses: a combined approach of
iterative calculation on skeleton model and conformal geometric algebra. *Symmetry* **11(3)**:301
DOI 10.3390/sym11030301.

Huang J, Hsu S, Huang C. 2013. Human upper body posture recognition and upper limbs motion
parameters estimation. In: *2013 Asia-Pacific Signal and Information Processing Association
Annual Summit and Conference.* 1–9.

**Jankovic J, McDermott M, Carter J, Gauthier S, Goetz C, Golbe L, Huber S, Koller W, Olanow
C, Shoulson I, Stern M, Tanner C, Weiner W. 1990.** Variable expression of parkinson's
disease: a base-line analysis of the datatop cohort. *Neurology* **40(10)**:1529–1534
DOI 10.1212/WNL.40.10.1529.

Ji B, Hong EJ. 2019. Deep-learning-based real-time road traffic prediction using long-term
evolution access data. *Sensors* **19(23)**:5327 DOI 10.3390/s19235327.

Jiang M, Kong J, Bebis G, Huo H. 2015. Informative joints based human action recognition using
skeleton contexts. *Signal Processing: Image Communication* **33(2)**:29–40
DOI 10.1016/j.image.2015.02.004.

Kamilaris A, Prenafeta-Boldú FX. 2018. Deep learning in agriculture: a survey. *Computers and
Electronics in Agriculture* **147(2)**:70–90 DOI 10.1016/j.compag.2018.02.016.

Keselman L, Woodfill JI, Grunnet-Jepsen A, Bhowmik A. 2017. Intel(r) realSense(TM)
stereoscopic depth cameras. In: *IEEE Conference on Computer Vision and Pattern Recognition
Workshops (CVPRW).* Piscataway: IEEE.

Kingma DP, Ba J. 2015. Adam: a method for stochastic optimization. *CoRR. Available at*
http://arxiv.org/abs/1412.6980.

Li B, Bai B, Han C. 2020. Upper body motion recognition based on key frame and random forest
regression. *Multimedia Tools and Applications* **79(7–8)**:5197–5212
DOI 10.1007/s11042-018-6357-y.

Li Y, Shen L. 2018. Skin lesion analysis towards melanoma detection using deep learning network.
Sensors **18(2)**:556 DOI 10.3390/s18020556.

Li Y, Xu H, Bian M, Xiao J. 2020. Attention based CNN-convLSTM for pedestrian attribute
recognition. *Sensors* **20(3)**:811 DOI 10.3390/s20030811.

Liu C, Chen L-C, Schroff F, Adam H, Hua W, Yuille AL, Fei-Fei L. 2019. Auto-deeplab:
hierarchical neural architecture search for semantic image segmentation. In: *The IEEE
Conference on Computer Vision and Pattern Recognition (CVPR).* Piscataway: IEEE.

Liu J, Wang Y, Liu Y, Xiang S, Pan C. 2020. 3D posturenet: a unified framework for skeleton-
based posture recognition. *Pattern Recognition Letters* **140(8)**:143–149
DOI 10.1016/j.patrec.2020.09.029.

Ma C, Li W, Cao J, Du J, Li Q, Gravina R. 2020. Adaptive sliding window based activity recognition for assisted livings. *Information Fusion* **53(1)**:55–65 DOI 10.1016/j.inffus.2019.06.013.

Maskeliunas R, Damaševičius R, Segal S. 2019. A review of internet of things technologies for ambient assisted living environments. *Future Internet* **11(12)**:259 DOI 10.3390/fi11120259.

Matthew RP, Seko S, Bailey J, Bajcsy R, Lotz J. 2019. Estimating sit-to-stand dynamics using a single depth camera. *IEEE Journal of Biomedical and Health Informatics* **23(6)**:2592–2602 DOI 10.1109/JBHI.2019.2897245.

Murugan P, Durairaj S. 2017. Regularization and optimization strategies in deep convolutional neural network. *CoRR. Available at* http://arxiv.org/abs/1712.04711.

Nweke HF, Teh YW, Mujtaba G, Al-garadi MA. 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: review and open research directions. *Information Fusion* **46(Part 1)**:147–170 DOI 10.1016/j.inffus.2018.06.002.

Okewu E, Misra S, Maskeliunas R, Damasevicius R, Fernandez-Sanz L. 2017. Optimizing green computing awareness for environmental sustainability and economic security as a stochastic optimization problem. *Sustainability* **9(10)**:1857 DOI 10.3390/su9101857.

Perusquía-Hernández M, Enomoto T, Martins T, Otsuki M, Iwata H, Suzuki K. 2017. Embodied interface for levitation and navigation in a 3d large space. In: *ACM International Conference Proceeding Series.* New York: ACM.

Pham HH, Salmane H, Khoudour L, Crouzil A, Zegers P, Velastin SA. 2019. Spatio—temporal image representation of 3D skeletal movements for view-invariant action recognition with deep convolutional neural networks. *Sensors* **19(8)**:1932 DOI 10.3390/s19081932.

Plonis D, Katkevicius A, Gurskas A, Urbanavicius V, Maskeliunas R, Damasevicius R. 2020. Prediction of meander delay system parameters for internet-of-things devices using pareto-optimal artificial neural network and multiple linear regression. *IEEE Access* **8**:39525–39535 DOI 10.1109/ACCESS.2020.2974184.

Redmon J, Farhadi A. 2018. Yolov3: an incremental improvement. *CoRR. Available at* http://arxiv.org/abs/1804.02767.

Ruta M, Scioscia F, di Summa M, Ieva S, Sciascio ED, Sacco M. 2014. Semantic matchmaking for kinect-based posture and gesture recognition. In: *2014 IEEE International Conference on Semantic Computing.* Piscataway: IEEE.

Ryselis K, Petkus T, Blažauskas T, Maskeliūnas R, Damaševičius R. 2020. Multiple kinect based system to monitor and analyze key performance indicators of physical training. *Human-Centric Computing and Information Sciences* **10(1)**:51.

Saab SS, Msheik H. 2016. Novel RFID-based pose estimation using single stationary antenna. *IEEE Transactions on Industrial Electronics* **63(3)**:1842–1852 DOI 10.1109/TIE.2015.2496909.

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Piscataway: IEEE, 4510–4520.

Sengupta A, Jin F, Zhang R, Cao S. 2020. Mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal* **20(17)**:10032–10044 DOI 10.1109/JSEN.2020.2991741.

Sharma M, Majumdar PK. 2009. Occupational lifestyle diseases: an emerging issue. *Indian Journal of Occupational and Environmental Medicine* **13(3)**:109–112 DOI 10.4103/0019-5278.58912.

Sundermeyer M, Ney H, Schluter R. 2015. From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23(3)**:517–529 DOI 10.1109/TASLP.2015.2400218.

Tariq M, Majeed H, Beg MO, Khan FA, Derhab A. 2019. Accurate detection of sitting posture activities in a secure IoT based assisted living environment. *Future Generation Computer Systems* **92(4)**:745–757 DOI 10.1016/j.future.2018.02.013.

Wang W-J, Chang J-W, Haung S-F, Wang R-J. 2016. Human posture recognition based on images captured by the kinect sensor. *International Journal of Advanced Robotic Systems* **13(2)**:54 DOI 10.5772/62163.

Wang J, Zhang J, Cai Y, Deng L. 2019. DeepMiR2GO: inferring functions of human microRNAs using a deep multi-label classification model. *International Journal of Molecular Sciences* **20(23)**:6046 DOI 10.3390/ijms20236046.

Wen L, Li X, Gao L, Zhang Y. 2018. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics* **65(7)**:5990–5998 DOI 10.1109/TIE.2017.2774777.

Wu G, Shao X, Guo Z, Chen Q, Yuan W, Shi X, Xu Y, Shibasaki R. 2018. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing* **10(3)**:407 DOI 10.3390/rs10030407.

Xu S, Guo J, Zhang G, Bie R. 2020. Automated detection of multiple lesions on chest x-ray images: classification using a neural network technique with association-specific contexts. *Applied Sciences* **10(5)**:1742 DOI 10.3390/app10051742.

Zemp R, Tanadini M, Plüss S, Schnüriger K, Singh NB, Taylor WR, Lorenzetti S. 2016. Application of machine learning approaches for classifying sitting posture based on force and acceleration sensors. *BioMed Research International* **2016(1)**:1–9 DOI 10.1155/2016/5978489.

Zhang Z. 2012. Microsoft kinect sensor and its effect. *IEEE Multimedia* **19(2)**:4–10 DOI 10.1109/MMUL.2012.24.

Zhang J, Wu C, Wang Y. 2020. Human fall detection based on body posture spatio-temporal evolution. *Sensors* **20(3)**:946 DOI 10.3390/s20030946.

Zhang T, Zhang X, Shi J, Wei S. 2019. Depthwise separable convolution neural network for high-speed sar ship detection. *Remote Sensing* **11(21)**:2483 DOI 10.3390/rs11212483.

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. 2016. Learning deep features for discriminative localization. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway: IEEE, 2921–2929.

Zhou J, Tian Y, Yuan C, Yin K, Yang G, Wen M. 2019. Improved uav opium poppy detection using an updated yolov3 model. *Sensors* **19(22)**:4851 DOI 10.3390/s19224851.