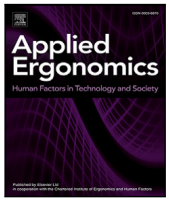
Applied Ergonomics 87 (2020) 103138

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



A novel vision-based real-time method for evaluating postural risk factors associated with musculoskeletal disorders

Li Li, Tara Martin, Xu Xu *

Edward P.Fitts Department of Industrial & Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Keywords:
RULA
Deep learning
MSD risk assessment

ABSTRACT

Real-time risk assessment for work-related musculoskeletal disorders (MSD) has been a challenging research problem. Previous methods such as using depth cameras suffered from limited visual range and wearable sensors could cause intrusiveness to the workers, both of which are less feasible for long-run on-site applications. This document examines a novel end-to-end implementation of a deep learning-based algorithm for rapid upper limb assessment (RULA). The algorithm takes normal RGB images as input and outputs the RULA action level, which is a further division of RULA grand score. Lifting postures collected in laboratory and posture data from Human 3.6 (a public human pose dataset) were used for training and evaluating the algorithm. Overall, the algorithm achieved 93% accuracy and 29 frames per second efficiency for detecting the RULA action level. The results also indicate that using data augmentation (a strategy to diversify the training data) can significantly improve the robustness of the model. The proposed method demonstrates its high potential for real-time on-site risk assessment for the prevention of work-related MSD. A demo video can be found at https://github.com/LLDavid/RULA_2DImage.

1. Introduction

Work-related musculoskeletal disorders (WMSDs) are a leading cause of pain, suffering, and disability in American workplaces (Kang et al., 2014). According to the US Bureau of Labor Statistics, in 2017 there were 344,970 nonfatal occupational injuries and illnesses associated with musculoskeletal disorders (MSDs) involving days away from work (DAFW) across all industries (Bureau of Labor Statistics, 2017b). Especially in manufacturing, MSDs accounted for 31.4% of DAFW cases (Bureau of Labor Statistics, 2017a). It is estimated that in the U.S., the annual direct cost of workers' compensation associated with MSDs is approximately $20 billion (Kang et al., 2014). Indirect costs, such as those associated with hiring and training replacement workers, are as much as five times the direct cost (Kang et al., 2014). Thus, there is a great necessity for developing an efficient and robust assessment tool for the prevention of WMSDs.

Physical risk factors for WMSDs include rapid work pace and repetitive motion, forceful exertions, non-neutral body postures, vibration (Punnett and Wegman, 2004), and lack of rest. Previous studies have proposed several observational tools for performing MSD risk assessment of certain jobs, including rapid upper limb assessment (RULA) (McAtamney and Corlett, 1993), rapid entire body assessment (REBA) (Hignett and McAtamney, 2000), the revised NIOSH lifting equation (RNLE) (Waters et al., 1993), postural loading on the upper body assessment (LUBA) (Kee and Karwowski, 2001), and occupational repetitive actions (OCRA) (Occhipinti, 1998). Among these observational tools, RULA has been widely adopted for safety practitioners in industry practice (Manghisi et al., 2017; Namwongsa et al., 2018; Cao et al., 2019). It requires minimal previous skills in observation technique and is easy to learn (Dockrell et al., 2012). A safety practitioner observes a worker's joint angles, body motion frequency, muscle use level, and carrying load weight, and then uses a series of RULA tables to determine the scores associated with each risk factor. In general, extreme joint angles, highly repetitive motions, and greater muscle use yield larger scores. By summing the scores from each risk factor, the total score of an observational method indicates the overall risk of musculoskeletal disorders. Previous studies (Öztürk and Esin, 2011; Sezgin and Esin, 2015) have shown that greater scores of an observational method are significantly correlated with self-reported MSD symptoms.

However, observational methods suffer from two major weaknesses. In current industrial practice, workers' postures need to be manually observed, coded, and inputted into a standardized form, which is time-consuming and less practical for long-term observation of workers rotating among multiple tasks. Second, scoring accuracy is mainly dependent on the proficiency of raters and viewing angles. Trained raters usually provide more valid scoring.

* Corresponding author.
E-mail address: xxu@ncsu.edu (X. Xu).

https://doi.org/10.1016/j.apergo.2020.103138
Received 13 January 2020; Received in revised form 31 March 2020; Accepted 21 April 2020
Available online 4 May 2020

Previous studies (Manghisi et al., 2017; Yan et al., 2017a) sought to address these weaknesses by applying different technologies, such as wearable sensors and computer vision, to infer the risk of MSDs through automated observational methods. The wearable sensors, such as inertial measurement units (IMUs) (Yan et al., 2017a; Vignais et al., 2013; Peppoloni et al., 2014; Hsu and Lin, 2019), allow for real-time reconstruction of a human pose for postural risk assessment in 3-D. IMUs enable field studies and provide an objective assessment of working postures (Balaguier et al., 2017). However, IMUs must be fastened tightly to the trunk and limbs with straps to eliminate skin artifact. Mounting IMUs on a worker's body not only interferes with natural motion (Manghisi et al., 2017) but also causes discomfort (Ribeiro and Santos, 2017).

Also used for postural risk assessment are depth cameras, such as the Microsoft Kinect and Intel RealSense (Manghisi et al., 2017; Diego-Mas and Alcaide-Marzal, 2014; Abobakr et al., 2017; Bhatia et al., 2019). Depth cameras can capture RGB-Depth data. Some studies show that using depth cameras for postural risk assessment in human pose detection is both accurate and efficient (Manghisi et al., 2017; Parsa et al., 2019; Umar et al., 2018). One limitation of depth cameras, however, is that the coverage of the infrared emitter limits the working range (e.g., 4.5 m for Kinect V2) (Gonzalez-Jorge et al., 2015). Therefore, the depth camera needs to be placed very close to the worker being observed. In lieu of depth cameras, stereo cameras can also be used for pose reconstruction, where the detected 2-D poses are integrated from two or more cameras placed on-site (Guo and Qian, 2007; López-Quintero et al., 2016). A previous study for ergonomic analysis (Liu et al., 2016) adopted two cameras for a tracking-based pose reconstruction. However, to ensure the effectiveness of this method, the workers needed to limit their movement in the intersection area of the visual field of all of the cameras.

Ideally, a single regular RGB camera could address the above-mentioned difficulties in human pose reconstruction (Yan et al., 2017b; Ding et al., 2019). A regular camera has greater visual field depth and is not intrusive to workers' natural motion. However, pose reconstruction with a single camera has been a challenging research problem in the computer vision community (Toshev and Szegedy, 2014; Shakhnarovich et al., 2003). Difficulties arise from the large feature space of the image and the high level of abstraction of the task. In the past few years, the development of deep learning and more powerful graphical processing units (GPUs) has allowed a great number of researchers to adopt the convolutional neural network (CNN), a form of deep neural structures, for vision-based human pose reconstruction, and greatly improved the accuracy of the reconstructed pose (Newell et al., 2016; Wei et al., 2016). In addition, the advent of Tensorflow Lite makes deploying deep learning-based methods on mobile devices possible (Manning et al., 2018), so that postural risk assessment can be performed on a hand-held cell phone.

This study aims to develop an efficient, robust, and practical method to automate RULA assessment. The proposed method assesses a posture in real-time by applying a deep neural network on streamed images captured with a regular camera (Fig. 1). The first part of the proposed method is a pose detector, which takes a single inspection at a monocular RGB image of a person and predicts the corresponding 2-D pose. The second part is a RULA estimator, which takes the coordinates of the detected 2-D pose as inputs and predicts the RULA action level directly, which is a further division of the RULA grand score (McAtamney and Corlett, 1993). This assessment method was trained with images of lifting tasks taken in a laboratory setting (Xu et al., 2011) and images from Human 3.6 (Ionescu et al., 2011, 2013), which is a public human pose dataset with full-body kinematics marker data. The proposed algorithm achieved results comparable with other recent related studies. Below, the method section introduces the methods including the proposed algorithm and data; the result section shows the test results; and the discussion section discusses the main contributions, limitations, and future work.

2. Method

2.1. RULA Scoring

The inputs to RULA assessment are body segment angles (upper arm, lower arm, wrist, neck, and upper trunk), muscle use, and external load. The output is the grand score (numbered 1–7). Scores 1–7 are further categorized into four levels of actions needing to be taken McAtamney and Corlett (1993) (illustrated in Fig. 2).

Previous work (Parsa et al., 2019) has revealed that the RULA grand score is highly sensitive to changes in rotation of a single segment, e.g., a minor change of neck position can lead to a significant change in neck score. Thus, postures that land between two scores have low reliability. To alleviate this problem, RULA action level was chosen as the algorithm output instead of the RULA grand score because it is less sensitive to minor changes in rotation. Besides, interventions could be given based on the estimated action level directly.

The RULA scores for training and testing were manually derived by two experimenters with experiences in ergonomics risk assessment. The camera setup in this study does not provide high enough resolution to recognize hand gestures with pixel-level details. Therefore, wrist score, muscle use, and workload were assumed to be uniform among all data. An on-screen ruler was used as reference for measuring segment angles on extracted frames. The scored images were used for training and validating the algorithm.

2.2. Pose detector

A 2-D pose detector was used in this study. The pose estimation can be regarded as a regression problem. The input is an RGB image, represented by an $N \times M \times 3$ matrix. N and M represent the number of rows and columns of pixels, respectively, which are determined by the resolution of the raw image. The color of each pixel is represented by a 3-D vector (RGB). The pose detector outputs detected 2-D pose, which consists of 17 key joint locations (referred as key points, and illustrated in Fig. 3) on the image plane. Each key point is represented by a 2-D vector (x, y), representing its location (width and height) on the input image. The choice of key points was based on two primary considerations. First, the key points must possess certain distinguishable image features that are invariant under different conditions (Lowe, 2004), including scaling, viewing angle rotation, etc. Second, the key points should give an articulated representation of the human pose. Therefore, the shoulder, elbow, wrist, etc., are the most commonly used key points (Yan et al., 2017b; Ding et al., 2019). The 17 key points used in this study follow the choice of OpenPose (Cao et al., 2017), a recently published open-access pose detection library.

Note that it is very challenging to perform 3-D pose reconstruction from an image using deep neural networks. However, the estimated 3-D poses are more prone to error, because pose depth information is missing and must be inferred (Ionescu et al., 2011, 2013). Additionally, in our previous work (Li and Xu, 2019), we used a three-layer neural network to predict RULA score from 2-D poses. The results indicated that an end-to-end 2-D pipeline is less computationally demanding for a real-time RULA assessment. Therefore, in the current study, we used an end-to-end RULA assessment algorithm with an integrated 2-D pose estimation. The proposed algorithm demands less computational power, and so has the potential to be used on mobile devices.

In this study, the pose detector was built on OpenPose (Cao et al., 2017), a convolutional neural network (CNN)-based pose detection approach. CNN is a popular choice for dealing with 2-D graphical input because the convolutional operations are extremely good at extracting low-level spatial information. More sophisticated tasks can be completed by stacking multiple CNN layers. However, one problem of using CNN is that the demand for computational resources snowballs rapidly as the size and number of layers increase.

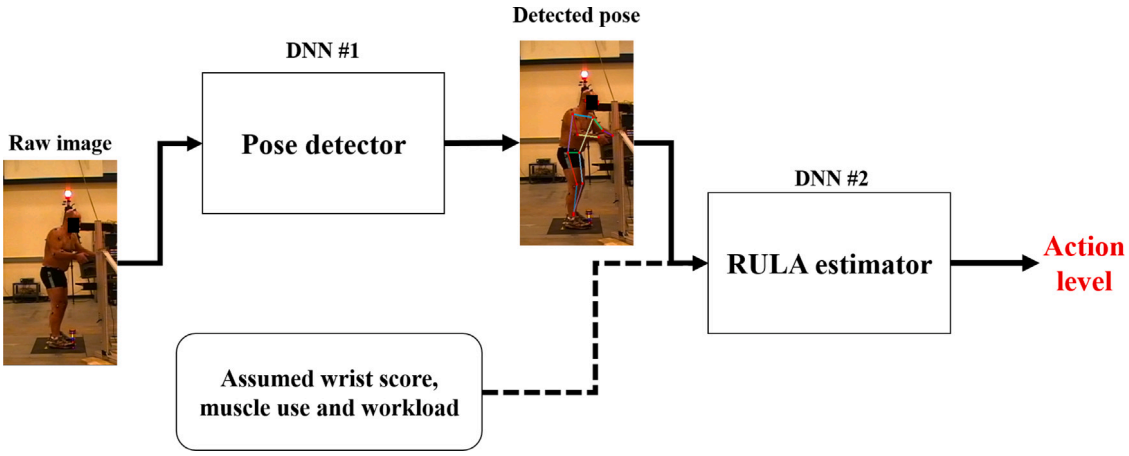


Fig. 1. Method overview. The first deep neural network (DNN #1) takes the raw image as input and outputs the detected pose. The wrist score, muscle use and workload were assumed during data labeling. DNN #2 takes the detected pose as input and estimates the action level, which is a further division of RULA grand score (McAtamney and Corlett, 1993).

Action level	Grand score	Description
L1	1 - 2	Posture is acceptable if it is not maintained or repeated for long periods.
L2	3 - 4	Further investigation is needed and changes may re required.
L3	5 - 6	Investigation and changes are required soon.
L4	7	Investigation and changes are required immediately.

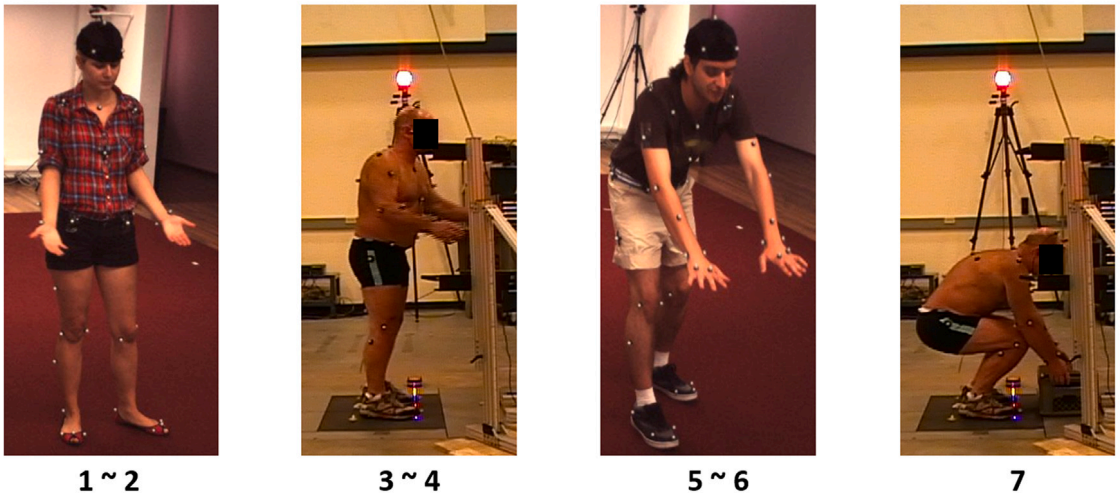


Fig. 2. Requirements for action levels (McAtamney and Corlett, 1993). RULA grand scores are divided into 4 levels according to the actions needing to be taken.

The pre-trained weights using COCO (a large-scale human key points dataset (Lin et al., 2014)) were loaded for the CNN-based pose detector in this study. The image was first reshaped into a fixed $N \times M \times 3$ matrix, which was then fed into the CNN. The CNN not only detects the key points but also learns the associations between those key points. The learned associations further refined the detected key points for a predetermined number of times. The detailed architecture of the network can be found in Cao et al. (2017). The CNN took the maximum value from the confidence map for each key point and outputted the 34 coordinates for each of the 17 key points (Fig. 4).

2.3. RULA Estimator

This study aims to infer the action level from 2-D joint coordinates derived from pose detection. The inference is performed by a second deep neural network.

The input to the RULA estimator is a 34×1 vector generated by the pose detector, and the output is a 4×1 vector corresponding to the four action levels based on the RULA grand score (see Fig. 2).

Fig. 5 shows the architecture of the RULA estimator. It consists of four dense layers. For each layer, batch normalization was added to prevent covariate shift (Ioffe and Szegedy, 2015). Rectified linear units (RELU) were chosen as the activation function (Nair and Hinton, 2010).

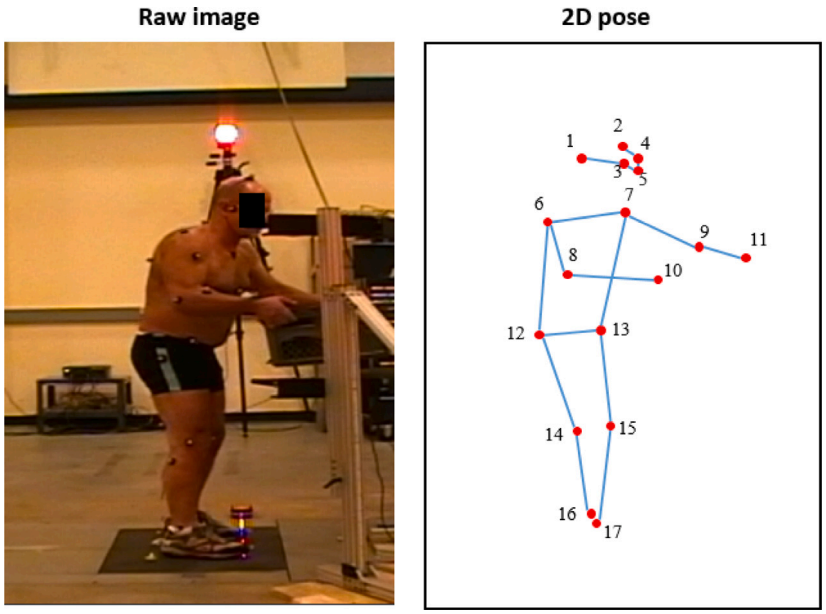


Fig. 3. Indexing for the 17 key points. 1-right ear; 2-left ear; 3-right eye; 4-left eye; 5-nose; 6-right shoulder; 7-left shoulder; 8-right elbow; 9-left elbow; 10-right wrist; 11-left wrist; 12-right hip; 13-left hip; 14-right knee; 15-left knee; 16-right ankle; 17-left ankle. Note that although RULA is designed for upper limb assessment, the lower limb locations are still necessary for determining the upper trunk bending angle and whether legs/feet are supported. The markers attached to the subject were for other research purposes and are irrelevant to this study.

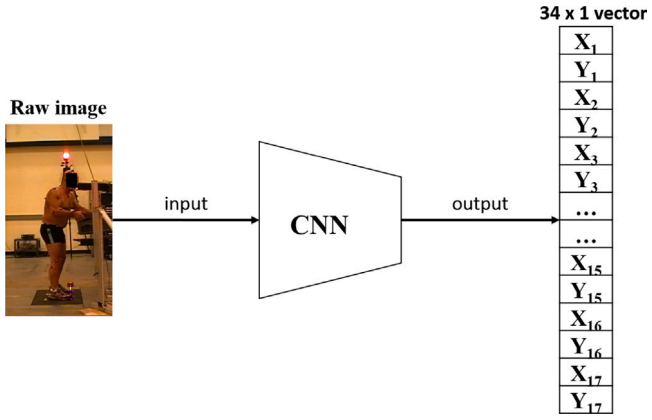


Fig. 4. Pose detector. (X_i, Y_i) represents the ith key point on the X_ith column and Y_ith row of the input image. The details of the convolutional neural network (CNN) structure can be found in Cao et al. (2017).

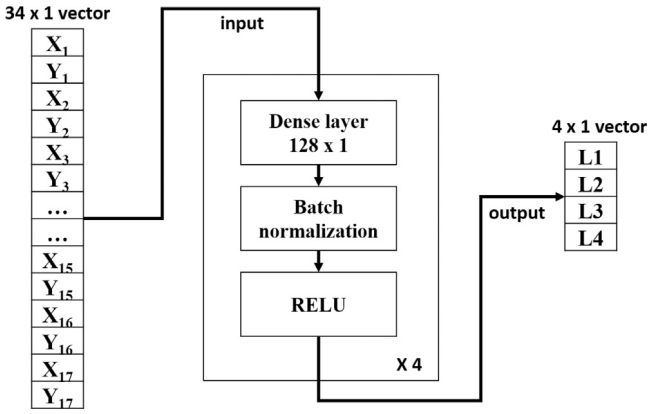


Fig. 5. RULA estimator: The dense layer is also called the fully connected (FC) layer. Batch normalization is one of the normalization methods used to prevent covariate shift. RELU represents rectified linear units, which is one of the most popular activation functions in deep learning. L1 to L4 denote four corresponding action levels (Fig. 2).

The equation is given as follows:

$$RELU(x) = \begin{cases} x, & if\ x > 0 \\ 0, & if\ x < 0 \end{cases} \quad (1)$$

where x is the output from the previous layer. Note that the RELU function sets all negative values to zero, which could create non-linearities to the network, and it is efficient to implement the RELU function during training because the gradient is a constant (0 if the input is negative). However, the limitation of using RELU is that some information encoded in the negative values could be lost during training.

2.4. Data preparation and augmentation for RULA estimator training

The RULA estimator was trained by combining data from our previous lifting study (Xu et al., 2011) with existing human motion data, including the data from Human 3.6 (Ionescu et al., 2011, 2013), which is one of the largest human motion databases (Fig. 6). Data used in this study is available at https://github.com/LLDavid/RULA_2DImage.

To make the algorithm more generalizable, the selected lifting postures contain a wide range of upper extremity motions, including upper arm flexion (approximately 0° to 110°), forearm flexion (approximately 0° to 150°), and upper trunk twisting. The Human 3.6 motion data is mostly daily postures, including more complex and diverse upper extremity motions (Fig. 6). Combining the lifting postures with the Human 3.6 motion data can enhance the generalizing abilities of the trained RULA estimator. The generalizing abilities are especially enhanced, considering that subjects in the lifting tasks were half-naked, and the subjects in Human 3.6 were wearing varied clothing. Therefore, a combined dataset with different clothing conditions can enhance the algorithm generalizability during the automated RULA assessment. On the other hand, the data from the previous lifting study provides specificity to the RULA estimator regarding a variety of RULA risk levels. Lifting tasks are also strongly associated with low-back MSDs and commonly analyzed by RULA. The original source for lifting postures were video clips. The images were extracted every five frames

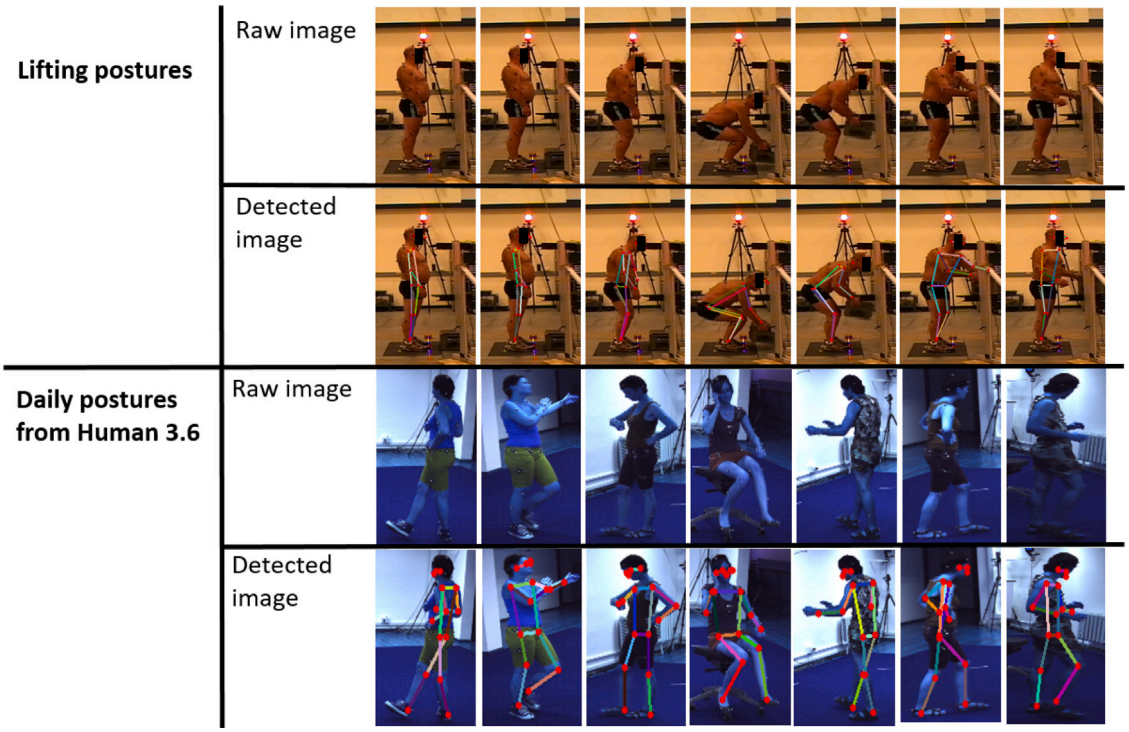


Fig. 6. Examples of lifting postures and Human 3.6 postures. Red dots are the detected key points using the pose detector. Limbs are painted with random colors.

to reduce redundancy, and a total of 423 images were extracted from the video clips. For asymmetric body postures, the left and right side of the body were scored separately, and the maximum of the two scores was taken as the final score since the higher score is of greater concern. The distribution of the action levels based on the RULA grand score from the two datasets is shown in Fig. 7.

Note that the distribution of the raw data is skewed, i.e., scores '2' and '3' take up the majority. To avoid overfitting to the majority and underfitting to the minority in training, two data augmentation techniques were adopted (Fig. 8). First, each key point was randomly 'hidden' on purpose. This is because, in practice, the algorithm will 'miss' some key points due to the occlusion by lifted objects or self-occlusion, as shown in Fig. 6. Manually creating such cases in the training data can enable the algorithm to deal with these occlusion cases and thus improve the robustness of the algorithm. The coordinates of the 'missed' or 'hidden' key points were set to $(-1, -1)$ since the convolutional layers are sensitive to negative values. The RELU function can easily filter out the negative values Eq. (1) so that the network can better distinguish these entries from the others.

Second, the raw output of the pose detection network is the probability heatmap. It is assumed that the predicted key points are normally distributed and thus centered around the actual location (illustrated in Fig. 8). The formula is given as follows:

$$\begin{bmatrix} x_{aug} \\ y_{aug} \end{bmatrix} = \begin{bmatrix} x_{raw} \\ y_{raw} \end{bmatrix} + \epsilon \tag{2}$$

where ϵ follows a normal distribution with zero mean. x_{aug} and y_{aug} are the augmented coordinates. x_{raw} and y_{raw} are the original coordinates. The actual locations were manually moved to neighboring locations to create more training data, and the shiftings of key points are mutually independent. The shifted points do not compromise the feasibility of the original pose because more than 90% of them are within two neighboring pixels of the actual locations. Rather, this augmentation method can enhance the robustness of the algorithm under noise. As shown in Fig. 7, the augmented data is less skewed than the original data. Note that in Human 3.6, there are zero training samples (or postures represented by joint coordinates and the corresponding image) with a RULA score equal of 1, so data augmentation is unnecessary.

2.5. Network training

The pre-trained weights using COCO dataset (Lin et al., 2014) were loaded into the pose detector. Note that the markers attached to the subjects were for other research purposes and were not used to train the pose detector. Thus, the markers would not have impacted the pose inference in this study. For the RULA estimator, the network was trained from scratch using pairs of 2-D poses and corresponding action levels in a fully supervised learning manner.

K-fold (k=5) cross-validation was used to validate the model. In each fold, the training data included 80% of the total data, and the remaining 20% was used as the test data. Adam Grad was adopted as the optimizer (Kingma and Ba, 2014), and the batch size was set to 10 (i.e., 10 training samples were used to optimize the parameter for each iteration). The model was trained for 50 epochs. In each epoch, the network was trained with 10 training samples at a time until iterating over all the data. Categorical cross-entropy (Goodfellow et al., 2016) was calculated as the training loss. It measures the difference between the distribution of the predicted and the actual action levels. The training was performed on a workstation with Intel Xeon CPU E5-1560 v4 @ 3.60 GHz and two Titan Vs.

3. Results

The average of the k-fold (k=5) validation was calculated as the final result. The dataset was first split into five groups. During each 'fold', one group was selected as the validation set, and the remaining four groups were used for the training. Fig. 9 reports the training loss and validation loss under different data preparation and augmentation conditions. The difference between the training loss and validation loss reflects the level of overfitting. In the first case, where only the lifting postures were included in the training set, the model is slightly overfitted (around 0.75 for loss difference). After implementing the data augmentation introduced in Fig. 8, the overfitting issue was mitigated. In the end, the training loss and validation loss were very close to each other. In the last case, the augmented lifting data was mixed with the data from Human 3.6, and the results show slight overfitting (approx.

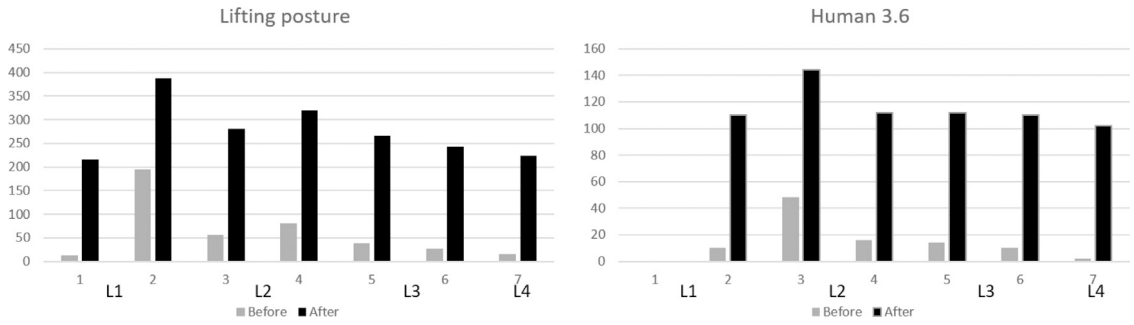


Fig. 7. Histogram of RULA score distribution before and after data augmentation. The horizontal axis represents the RULA score (1–7). L1 to L4 denote the four corresponding action levels (Fig. 2), e.g., RULA score 1 and 2 belong to L1.

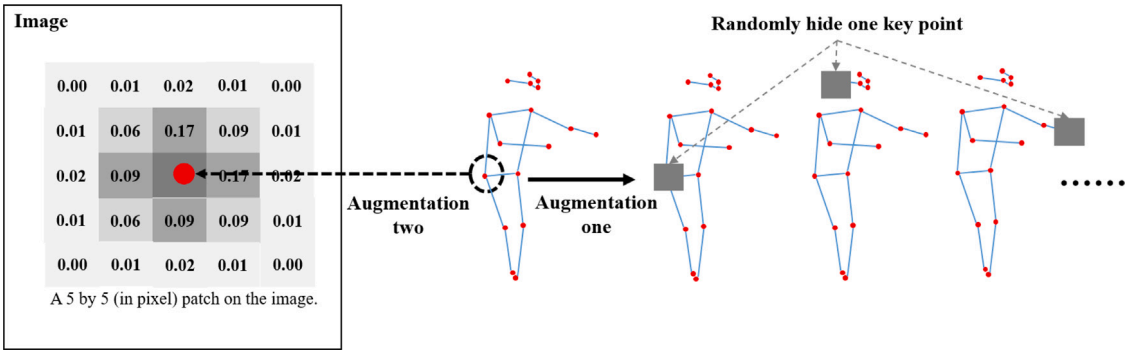


Fig. 8. Two data augmentation methods. Augmentation one randomly hides one key point (set to $(-1,-1)$). Augmentation two adds a random noise to the original key point location. The value in each grid in the left matrix represents the probability that the original key point (red dot) translates to that grid. The two augmentation methods are intended to increase the variability of the original dataset.

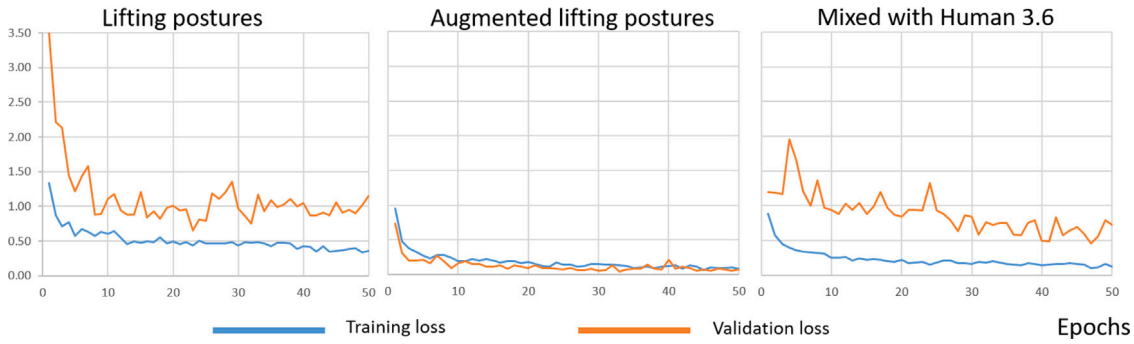


Fig. 9. Training and validation loss. The left graph is from model training with lifting postures exclusively. The middle diagram indicates training with augmented lifting postures. The right graph shows the results for training with augmented lifting postures mixed with Human 3.6 data.

0.65 for loss difference), but not as strong as the first case. In other words, the model in the last case has better generalizability but also has greater classification errors on a specific dataset.

Fig. 10 shows the RULA action level for a full lifting task. Note that only the posture was measured, and the wrist score and workload were assumed to be uniform. So at the beginning and the end of the lifting task, the action level is at the lowest level, as well as when the subject lifted the box at around the elbow height. When the subject lifted the box above shoulder height, the RULA action level was at the highest level.

Table 1 indicates the confusion matrix for the three cases. Table 2 shows the results of applying different evaluation metrics calculated from the confusion matrix. Note that the precision, recall, and F1-score metrics are designed for binary classification. They generally underestimate the sensitivity of an algorithm. For a more comprehensive evaluation, metrics (i.e., micro-averaging and macro-averaging) that are specifically designed for multi-class classification (Sokolova and Lapalme, 2009) were adopted. The formulas for these evaluation metrics can be found in Appendix.

Table 3 compares the proposed method with other vision-based postural risk assessment algorithms. Because the mixed dataset includes more diverse postures and different viewing angles, the statistical measures are less likely to overfit. Our method achieved comparable overall accuracy and higher efficiency. However, it should be noted that the results should be interpreted with caution because different algorithms were evaluated with different metrics, hardware configurations, and posture dataset.

4. Discussion

The primary aim of this study was to develop a robust and efficient vision-based algorithm to automate RULA assessment. In this study, a CNN-based pose detector was adopted to infer 2-D poses from RGB images, and a second DNN was designed to estimate RULA action levels

Table 1
Confusion matrix for three treatments. The entries on the diagonal in each matrix represent the number of cases that were correctly classified.

Action level	Predicted													
	Lifting postures					Augmented					Mixed with Human 3.6			
		1	2	3	4		1	2	3	4		1	2	3	4
True class	**1**	42	1	0	0	**1**	120	2	0	0	**1**	131	5	6	0
	2	2	13	4	3	**2**	10	119	1	0	**2**	16	168	0	0
	3	0	4	6	1	**3**	0	0	89	0	**3**	8	7	119	0
	4	2	0	1	6	**4**	0	0	0	47	**4**	0	0	0	66

Table 2
Evaluation results. ↑ represents when the larger value is preferred. Avg. denotes the inter-class average. μ denotes micro-averaging. M denotes macro-averaging.

	Lifting postures					Augmented					Mixed with Human 3.6				
	1	2	3	4	Avg.	1	2	3	4	Avg.	1	2	3	4	Avg.
Precision↑	0.91	0.72	0.55	0.60	**0.78**	0.92	0.98	0.99	1.00	**0.97**	0.85	0.93	0.95	1.00	**0.93**
Recall↑	0.98	0.59	0.55	0.67	**0.79**	0.98	0.92	1.00	1.00	**0.97**	0.92	0.91	0.89	1.00	**0.93**
F1-score↑	0.94	0.65	0.55	0.63	**0.79**	0.95	0.95	0.99	1.00	**0.97**	0.88	0.92	0.92	1.00	**0.93**
Precision_{μ}↑	0.79					0.97					0.92				
Recall_{μ}↑	0.79					0.97					0.92				
F1-score_{μ}↑	0.79					0.97					0.92				
Precision_{M}↑	0.70					0.97					0.93				
Recall_{M}↑	0.69					0.97					0.93				
F1-score_{M}↑	0.70					0.97					0.93				

Table 3
Comparison of vision-based postural risk assessment algorithms. The algorithm reported in Manghisi et al. (2017) only introduced the p-value for the Z-test. In Yan et al. (2017b), results from different algorithms were reported (Neural network (NN), K-nearest neighbor (KNN), decision tree (DT), and ensemble classifier (EC)). The overall accuracy takes the average of all the algorithms. '-' represents that information was not mentioned in the paper.

	Assessment	Overall accuracy	Efficiency (FPS)	Hardware
Our method	RULA	0.93	29	CPU Xeon(R) CPU E5-1650 v4 @ 3.60 GHz, 128 GB RAM, GPU Titan V
Kinect v2 Manghisi et al. (2017)	RULA	P<0.001	10	CPU Intel Core i5-4200 @ 2.50 GHz, 4 GB RAM, GT 740 M
Kinect V1 Diego-Mas and Alcaide-Marzal (2014)	OWAS	0.89	25	CPU @ 3.4 GHz processor, 4 GB RAM
NN/KNN/DT/EC Yan et al. (2017b)	OWAS	0.88	–	–

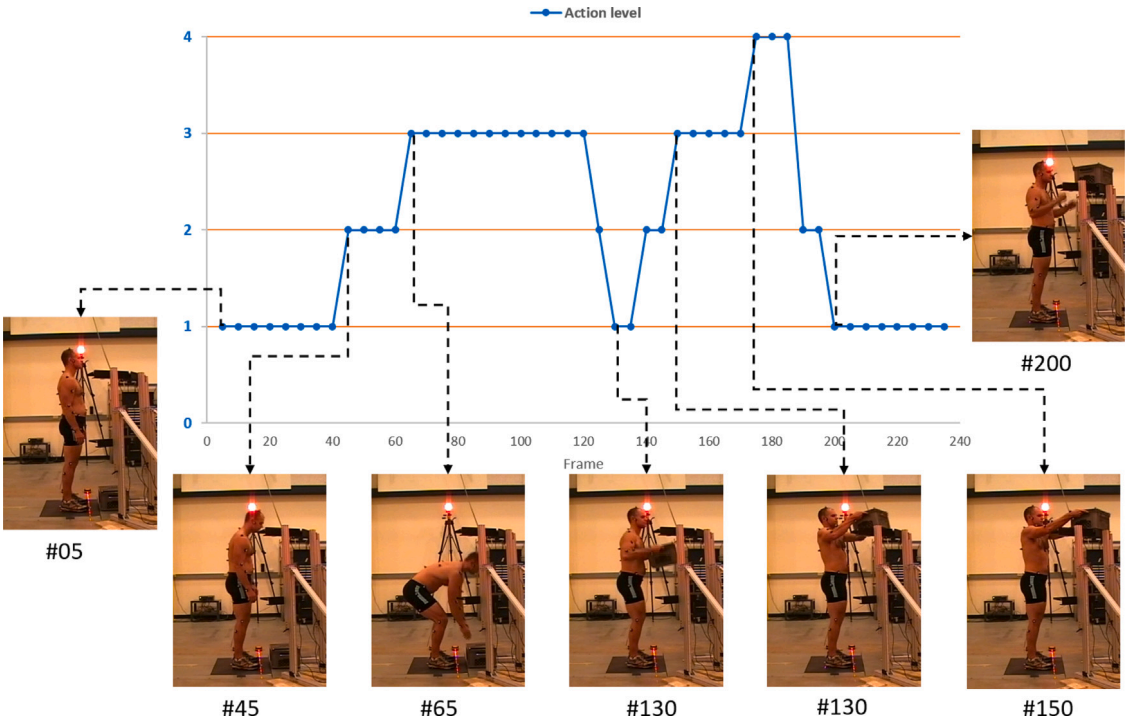


Fig. 10. Predicted action level of a lifting task. Note that the wrist score, muscle use, and workload were assumed uniform. Presented pictures represent several key timestamps across the lifting task. The numbers below the graph represent the frame IDs.

from the detected 2-D pose. The proposed algorithm yielded 93% overall accuracy, and 29 FPS in efficiency. Under multi-class classification metrics, it achieved high accuracy in both micro-averaging (92%) and macro-averaging (93%). Most of the misclassified testing samples are distributed around action levels 2 to 3. In other words, the algorithm is more sensitive to extreme cases, where the posture is either very safe

or very risky. This is because the extreme cases (very safe or risky) have wider margin, while cases in the middle sometimes fall on the decision boundary, which causes the algorithm to fail.

In this study, manual RULA scores by observers were used as true labels. The other way to score postures is by directly calculating the RULA score from kinematics data, which can have higher reliability. To determine segment scores from kinematics data, anatomical planes (e.g., upper body sagittal, coronal, and transverse planes) need to be defined as a reference, which requires precise choice and tracking of anatomical landmarks. However, RULA assessment was initially designed as an observation-based method, and the original paper introducing RULA (McAtamney and Corlett, 1993) only gives general guidelines for determining segment angles. Therefore, the choice of anatomical landmarks could have significant impacts on the consistency of RULA assessment. This is one important reason why we used manual scoring in this study, rather than acquiring RULA score from kinematics data. The trained algorithm learned the latent relations between 2-D pose and RULA scores from manual scoring.

One problem of applying machine learning algorithms in RULA assessment is the lack of training samples. Because manually scoring one posture with RULA takes around 5–7 min, acquiring a large-scale dataset is time-consuming and costly. Because of the data-driven nature of machine learning algorithms, the robustness of algorithms is largely dependent on the size and variability of the dataset. This study proposed two specific data augmentation methods, and the results show that the proposed methods could be used to expand the dataset and augment the performance of the algorithm by (approximately 18% in overall accuracy). This is very useful for on-site applications where we have limited resources to score postures with RULA for algorithm training.

This work modularized a pose detector and RULA estimator. The two neural networks were trained separately. The advantage of training separately is that it preserves the flexibility to extend to other applications. With modifications to the RULA estimator, it is possible to easily apply this real-time framework for other observational methods, such as REBA, LUBA, and OWAS, because they all have postures as the primary input.

Tensorflow was used as the backend in this study. The advent of Tensorflow Lite allows the proposed algorithm to be transferred onto Andriod and IOS platforms (Manning et al., 2018). The images used in this study were extracted from streamed videos, and the proposed algorithm can process 29 streamed images per second (29 FPS), which exceeds most webcam frame rates (15 FPS) and is very close to the frame rate of a regular camcorder (29.97 FPS). Therefore, the proposed algorithm can be run on a hand-held cell phone to perform a real-time end-to-end RULA assessment onsite, and the user can easily evaluate an ongoing task. Through analysis of the predicted scores over time (Fig. 10), the user has convenient access to the average, percentile, maximum, and variation of RULA scores of a task. It should be noted that one can use a wide-angle camera to improve the field of view, and thus cover a greater area while the worker moves around. However, a high-resolution camera may compromise the processing efficiency of the algorithm because the image input may be larger dimensions. The trade-off between efficiency and field of view is usually application-dependent.

Additionally, due to the adoption of the end-to-end pipeline, the availability of intermediate outputs (e.g., segment angles) used for RULA calculation was sacrificed for efficiency. In the case that the intermediate segment angles are needed for work redesign, safety practitioners can review the detected images that have a high RULA score and find which body segment resulted in the high RULA score. For example, for postures with action level 4 (high-risk postures), safety practitioners should check the associated images immediately; and for postures within action level 2–3 (low and medium risk postures), safety practitioners could check review at a later time if time is limited. This will substantially reduce the workload of safety practitioners as they will not need to observe the task of interest for the entire period and can prioritize the potentially riskier postures. These selected postures could be further investigated through video-based tools (Hanse and Forsman, 2001; Reiman et al., 2014) for a more detailed evaluation.

A major limitation of this study is that wrist score, muscle use, and workload were assumed uniform. Also, the proposed algorithm only examines static postures. Body movement frequency and the level of muscle use are not considered. Therefore, the application of the current algorithm is limited to light-duty tasks with moderately repetitive body motions. To date, there have been few pioneering studies attempting to estimate object weight from an image (Standley et al., 2017). A very recent study (Liu, 2019) used a deep learning-based method to estimate lifting frequency, duration, and muscle load from videos. However, results showed that the model is highly task-specific, and the accuracy of the estimated object weight needs to be further improved to infer muscle use. Moreover, the wrist postures are indistinguishable or even invisible in a full-body image because of image resolution and self-occlusion. In the current study, the wrist score was assumed to be uniform among all data. An additional telephoto camera dedicated to capturing wrist posture may be necessary for inferring the RULA score of the wrists (e.g., Leap Motion (Marin et al., 2014)).

Although the algorithm demonstrated sufficient performance on scoring postures in different laboratory environments, its performance in on-site applications may not be guaranteed. For example, visual noise could challenge the algorithm. The unpredictable nature of on-site environments, factors such as lighting conditions, and dust could result in noise on the images collected. Current CNN structures with a fully-supervised learning framework can only learn features present in the training data, so overfitting would be inevitable when visual noise is introduced. The accuracy achieved in a laboratory environment should be interpreted as the best scenario for testing in an on-site environment. Therefore, images of a variety of postures collected from actual workplaces need to be further coded to improve the generalizability and robustness of this approach.

In the future, other pose detectors and more structures of RULA estimator will be explored further to enhance the robustness and efficiency of the algorithm. In addition, our ongoing study is focusing on collecting a more comprehensive set of working postures covering a wider range of on-site postures. The collected posture data, as well as the synchronized videos will be made publically available. We expect that it will further contribute to the application of vision-based RULA assessment.

5. Conclusion

In this study, a novel vision-based real-time RULA assessment method was explored. It allows for efficient RULA assessment of a single image taken by a normal RGB camera. The method was trained and validated with postures during lifting tasks and daily activities. This method achieved an overall accuracy of 93% in RULA action level inference, and can be implemented with 29 FPS, which guarantees real-time applications. The results also indicate that using data augmentation techniques can augment the performance for action level inference and demonstrate the potential for real-time RULA assessment with a single camera.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This manuscript is based upon work supported by the National Science Foundation under Grant No 1822477.

Appendix. Metrics used in evaluation

A.1. Precision, recall and F1-score

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

$$F_1 score = \frac{PrecisionRecall}{Precision + Recall} \tag{5}$$

Where tp represents true positives; tn denotes true negatives; fn is the number of false negatives and fp is the false positives.

A.2. Micro-averaging

$$Precision_\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fp_i)} \tag{6}$$

$$Recall_\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)} \tag{7}$$

$$F1score_\mu = \frac{2 \times Precision_\mu Recall_\mu}{Precision_\mu + Recall_\mu} \tag{8}$$

Where, tp_i represents true positives for the ith class; tn_i denotes true negatives; fn_i is the number of false negatives; and fp_i is the number of false positives. l equals total number of classes. $Precision_\mu$ and $Recall_\mu$ evaluate the effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions, and the $Fscore_\mu$ reveals the relationships between the data's positive labels and those from classifier based on sums of per-text decisions.

A.3. Macro-averaging:

$$Precision_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{9}$$

$$Recall_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \tag{10}$$

$$F1score_M = \frac{2 \times Precision_M Recall_M}{Precision_M + Recall_M} \tag{11}$$

The macro-averaging measures the agreement of the average per-class of a classifier to identify class labels. These systematic measures will give a more objective and reliable evaluation for the algorithm.

References

Abobakr, Ahmed, Nahavandi, Darius, Iskander, Julie, Hossny, Mohammed, Nahavandi, Saeid, Smets, Marty, 2017. A kinect-based workplace postural analysis system using deep residual networks. In: 2017 IEEE International Systems Engineering Symposium (ISSE). IEEE, pp. 1–6.

Balaguier, Romain, Madeleine, Pascal, Rose-Dulcina, Kevin, Vuillerme, Nicolas, 2017. Trunk kinematics and low back pain during pruning among vineyard workers—A field study at the chateau larose-trintaudon. PLoS One 12 (4).

Bhatia, Vibha, Kalra, Parveen, Randhawa, Jagjit Singh, 2019. Upper body postural analysis in sitting workplace environment using microsoft kinect v2 sensor. In: Research Into Design for a Connected World. Springer, pp. 575–586.

Bureau of Labor Statistics, 2017a. Employer-Reported workplace injuries and illnesses. https://stats.bls.gov/iif/oshcdnew.htm.

Bureau of Labor Statistics, 2017b. Occupational injuries/illnesses and fatal injuries profiles. https://data.bls.gov/gqt/RequestData.

Cao, Zhe, Simon, Tomas, Wei, Shih-En, Sheikh, Yaser, 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299.

Cao, Wenming, Zhong, Jianqi, Cao, Guitao, He, Zhiquan, 2019. Physiological function assessment based on kinect v2. IEEE Access 7, 105638–105651.

Diego-Mas, Jose Antonio, Alcaide-Marzal, Jorge, 2014. Using Kinect™ sensor in observational methods for assessing postures at work. Appl. Ergon. 45 (4), 976–985.

Ding, Zewei, Li, Wanqing, Ogunbona, Philip, Qin, Ling, 2019. A real-time webcam-based method for assessing upper-body postures. Mach. Vis. Appl. 1–18.

Dockrell, Sara, O'Grady, Eleanor, Bennett, Kathleen, Mullarkey, Clare, Mc Connell, Rachel, Ruddy, Rachel, Twomey, Seamus, Flannery, Colleen, 2012. An investigation of the reliability of rapid upper limb assessment (RULA) as a method of assessment of children's computing posture. Appl. Ergon. 43 (3), 632–636.

Gonzalez-Jorge, H, Rodríguez-González, P, Martínez-Sánchez, J, González-Aguilera, D, Arias, P, Gesto, M, Díaz-Vilariño, L, 2015. Metrological comparison between kinect I and kinect II sensors. Measurement 70, 21–26.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Guo, Feng, Qian, Gang, 2007. Human pose inference from stereo cameras. In: 2007 IEEE Workshop on Applications of Computer Vision (WACV'07). IEEE, p. 37.

Hanse, Jan Johansson, Forsman, Mikael, 2001. Identification and analysis of unsatisfactory psychosocial work situations: A participatory approach employing video–computer interaction. Applied Ergon. 32 (1), 23–29.

Hignett, Sue, McAtamney, Lynn, 2000. Rapid entire body assessment (REBA). Appl. Ergon. 31 (2), 201–205.

Hsu, Chi-Fang, Lin, Ta-Te, 2019. Development of an ergonomic evaluation system based on inertial measurement unit and its application for exoskeleton load reduction. In: 2019 ASABE Annual International Meeting. American Society of Agricultural and Biological Engineers, p. 1.

Ioffe, Sergey, Szegedy, Christian, 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Ionescu, Catalin, Li, Fuxin, Sminchisescu, Cristian, 2011. Latent structured models for human pose estimation. In: 2011 International Conference on Computer Vision. IEEE, pp. 2220–2227.

Ionescu, Catalin, Papava, Dragos, Olaru, Vlad, Sminchisescu, Cristian, 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36 (7), 1325–1339.

Kang, Dongmug, Kim, Young-Ki, Kim, Eun-A, Kim, Dae Hwan, Kim, Inah, Kim, Hyoung-Ryoul, Min, Kyoung-Bok, Jung-Choi, Kyunghee, Oh, Sung-Soo, Koh, Sang-Baek, 2014. Prevention of Work-Related Musculoskeletal Disorders. BioMed Central.

Kee, Dohyung, Karwowski, Waldemar, 2001. LUBA: An assessment technique for postural loading on the upper body based on joint motion discomfort and maximum holding time. Applied Ergon. 32 (4), 357–366.

Kingma, Diederik P., Ba, Jimmy, 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, Li, Xu, Xu, 2019. A deep learning-based RULA method for working posture assessment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63. (1), SAGE Publications Sage CA: Los Angeles, CA, pp. 1090–1094.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, Zitnick, C Lawrence, 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, Meiyin, 2019. Video-Based Human Motion Capture and Force Estimation for Comprehensive On-Site Ergonomic Risk Assessment (Ph.D. thesis).

Liu, Meiyin, Han, SangUk, Lee, SangHyun, 2016. Tracking-based 3D human skeleton extraction from stereo video camera toward an on-site safety and ergonomic analysis. Constr. Innov.

López-Quintero, Manuel I, Marín-Jiménez, Manuel J, Muñoz-Salinas, Rafael, Madrid-Cuevas, Francisco J, Medina-Carnicer, Rafael, 2016. Stereo pictorial structure for 2D articulated human pose estimation. Mach. Vis. Appl. 27 (2), 157–174.

Lowe, David G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Manghisi, Vito Modesto, Uva, Antonio Emmanuele, Fiorentino, Michele, Bevilacqua, Vitoantonio, Trotta, Gianpaolo Francesco, Monno, Giuseppe, 2017. Real time RULA assessment using Kinect v2 sensor. Appl. Ergon. 65, 481–491.

Manning, Jacob, Langerman, David, Ramesh, Barath, Gretok, Evan, Wilson, Christopher, George, Alan, MacKinnon, James, Crum, Gary, 2018. Machine-learning space applications on smallsat platforms with tensorflow. In: Proceedings of the 32nd Annual AIAA/USU Conference on Small Satellites, Logan, UT, USA. pp. 4–9.

Marin, Giulio, Dominio, Fabio, Zanuttigh, Pietro, 2014. Hand gesture recognition with leap motion and kinect devices. In: 2014 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1565–1569.

McAtamney, Lynn, Corlett, E. Nigel, 1993. RULA: A survey method for the investigation of work-related upper limb disorders. Appl. Ergon. 24 (2), 91–99.

Nair, Vinod, Hinton, Geoffrey E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 807–814.

Namwongsa, Suwalee, Puntumetakul, Rungthip, Neubert, Manida Swangnetr, Chaiklieng, Sunisa, Boucaut, Rose, 2018. Ergonomic risk assessment of smartphone users using the Rapid Upper Limb Assessment (RULA) tool. PLoS One 13 (8), e0203394.

Newell, Alejandro, Yang, Kaiyu, Deng, Jia, 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, pp. 483–499.

Occhipinti, Enrico, 1998. OCRA: A concise index for the assessment of exposure to repetitive movements of the upper limbs. Ergonomics 41 (9), 1290–1311.

Öztürk, Nilüfer, Esin, Melek Nihal, 2011. Investigation of musculoskeletal symptoms and ergonomic risk factors among female sewing machine operators in Turkey. Int. J. Ind. Ergon. 41 (6), 585–591.

Parsa, Behnoosh, Samani, Ekta U, Hendrix, Rose, Devine, Cameron, Singh, Shashi M, Devasia, Santosh, Banerjee, Ashis G, 2019. Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks. IEEE Robot. Autom. Lett. 4 (4), 3153–3160.

Peppoloni, Lorenzo, Filippeschi, Alessandro, Ruffaldi, Emanuele, 2014. Assessment of task ergonomics with an upper limb wearable device. In: 22nd Mediterranean Conference on Control and Automation. IEEE, pp. 340–345.

Punnett, Laura, Wegman, David H., 2004. Work-related musculoskeletal disorders: The epidemiologic evidence and the debate. J. Electromyography Kinesiol. 14 (1), 13–23.

Reiman, Arto, Pekkala, Janne, Väyrynen, Seppo, Putkonen, Ari, Forsman, Mikael, 2014. Participatory video-assisted evaluation of truck drivers' work outside cab: Deliveries in two types of transport. Int. J. Occup. Saf. Ergon. 20 (3), 477–489.

Ribeiro, Nuno Ferrete, Santos, Cristina P., 2017. Inertial measurement units: A brief state of the art on gait analysis. In: 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG). IEEE, pp. 1–4.

Sezgin, D., Esin, M.N., 2015. Predisposing factors for musculoskeletal symptoms in intensive care unit nurses. Int. Nurs. Rev. 62 (1), 92–101.

Shakhnarovich, Gregory, Viola, Paul, Darrell, Trevor, 2003. Fast pose estimation with parameter sensitive hashing.

Sokolova, Marina, Lapalme, Guy, 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45 (4), 427–437.

Standley, Trevor, Sener, Ozan, Chen, Dawn, Savarese, Silvio, 2017. image2mass: Estimating the mass of an object from its image, in: Conference on Robot Learning, pp. 324–333.

Toshev, Alexander, Szegedy, Christian, 2014. Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660.

Umar, Radin Zaid Radin, Ling, Chai Fong, Ahmad, Nadiah, Halim, Isa, Lee, Fatin Ayuni Mohd Azli, Abdullasim, Nazreen, Pembuatan, Fakulti Kejuruteraan, 2018. Initial validation of RULA-kinect system–comparing assessment results between system and human assessors. In: Proceedings of Mechanical Engineering Research Day 2018, vol. 2018. Centre for Advanced Research on Energy, pp. 67–68.

Vignais, Nicolas, Miezal, Markus, Bleser, Gabriele, Mura, Katharina, Gorecky, Dominic, Marin, Frédéric, 2013. Innovative system for real-time ergonomic feedback in industrial manufacturing. Appl. Ergon. 44 (4), 566–574.

Waters, Thomas R, Putz-Anderson, Vern, Garg, Arun, Fine, Lawrence J, 1993. Revised NIOSH equation for the design and evaluation of manual lifting tasks. Ergonomics 36 (7), 749–776.

Wei, Shih-En, Ramakrishna, Varun, Kanade, Takeo, Sheikh, Yaser, 2016. Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732.

Xu, Xu, Chang, Chien-chi, Faber, Gert S, Kingma, Idsart, Dennerlein, Jack T, 2011. The validity and interrater reliability of video-based posture observation during asymmetric lifting tasks. Hum. Factors 53 (4), 371–382.

Yan, Xuzhong, Li, Heng, Li, Angus R., Zhang, Hong, 2017a. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. Autom. Constr. 74, 2–11.

Yan, Xuzhong, Li, Heng, Wang, Chen, Seo, JoonOh, Zhang, Hong, Wang, Hongwei, 2017b. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. Adv. Eng. Inform. 34, 152–163.