Applied Ergonomics 80 (2019) 75–88

Contents lists available at ScienceDirect

Applied Ergonomics

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

RGB-D ergonomic assessment system of adopted working postures

Ahmed Abobakr[a,*], Darius Nahavandi[a], Mohammed Hossny[a], Julie Iskander[a], Mohammed Attia[a], Saeid Nahavandi[a], Marty Smets[b]

[a] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, 75 Pigdons Rd, Waurn Ponds, Victoria, 3216, Australia
[b] Ford Motor Company, 29500 Plymouth Rd, Livonia, MI, 48150, USA

ARTICLE INFO

Keywords:
Ergonomics
MSDs
RULA
RGB-D
Deep learning
ConvNet
CNN
Posture analysis

ABSTRACT

Ensuring a healthier working environment is of utmost importance for companies and global health organizations. In manufacturing plants, the ergonomic assessment of adopted working postures is indispensable to avoid risk factors of work-related musculoskeletal disorders. This process receives high research interest and requires extracting plausible postural information as a preliminary step. This paper presents a semi-automated end-to-end ergonomic assessment system of adopted working postures. The proposed system analyzes the human posture holistically, does not rely on any attached markers, uses low cost depth technologies and leverages the state-of-the-art deep learning techniques. In particular, we train a deep convolutional neural network to analyze the articulated posture and predict body joint angles from a single depth image. The proposed method relies on learning from synthetic training images to allow simulating several physical tasks, different body shapes and rendering parameters and obtaining a highly generalizable model. The corresponding ground truth joint angles have been generated using a novel inverse kinematics modeling stage. We validated the proposed system in real environments and achieved a joint angle mean absolute error (MAE) of 3.19 ± 1.57° and a rapid upper limb assessment (RULA) grand score prediction accuracy of 89% with Kappa index of 0.71 which means substantial agreement with reference scores. This work facilities evaluating several ergonomic assessment metrics as it provides direct access to necessary postural information overcoming the need for computationally expensive post-processing operations.

1. Introduction

Musculoskeletal disorders (MSDs) are a common concern across labor intensive industries. A recent statistical study performed by the Bureau of Labor Statistics (BLS) demonstrated that MSD cases account for 31% of all work-related injuries and illness cases (Bureau of Labor Statistic, 2016). These injuries are most commonly in relation to the muscular components of the Neck, Back, Arms and Legs (Luttmann et al. Organizationet al.). In addition to the personal impact these injuries can have on workers, compensation costs and days-away-from-work can greatly effect the productivity of the organization it self (Bureau of Labor Statistic, 2016; Bernard and Putz-Anderson). The manufacturing industries endeavor to constantly provide a safe working environment via the early identification and intervention of problematic procedures. Currently, proactive task planning using digital human models and virtual facilities are helping minimize risk factors of MSDs, however to ensure the maintenance of harm minimization, advancements in injury prevention technology must continue to be implemented. This is due to the complex interactions between force and frequency during automotive assembly tasks. Adopting ergonomically invalid or awkward working postures while performing these manual tasks have the potential to cause long term MSDs (Krüger and Nguyen, 2015; Bernard and Putz-Anderson; Luttmann et al. Organizationet al.). Therefore, ergonomics specialists have been investigating methods and tools to evaluate the adopted working posture and identify potential MSDs risks.

The Rapid Upper Limb Assessment (RULA) (McAtamney and Corlett, 1993) is one of the most popular ergonomic assessment tools in the industry (Plantard et al. Multon; Liebregts et al., 2016). Despite its limitations and low resolution capabilities to problematic working procedures, RULA is simple, easy to compute and does not require prior knowledge in biomechanics or ergonomics. The RULA score quantifies the exposure of the adopted posture to risk factors of MSDs with more focus on the neck, trunk and upper body limbs. It ranges from one to seven representing the level of MSD risk and suggesting an action level that describes whether a method of intervention is required

* Corresponding author.
E-mail address: a.abobakr@deakin.edu.au (A. Abobakr).

https://doi.org/10.1016/j.apergo.2019.05.004
Received 21 April 2018; Received in revised form 10 April 2019; Accepted 14 May 2019
Available online 25 May 2019
0003-6870/

(McAtamney and Corlett, 1993), with one being an acceptable posture and seven requiring immediate intervention. Automating the RULA score evaluation process has gained much attention from ergonomics researchers (Manghisi et al. Monno; Plantard et al., 2017) to overcome the intra- and inter-rater variability problem (Manghisi et al., 2017). However, developing an automated RULA based ergonomic feedback system requires estimating joint angles of the upper body parts.

Recent studies have proposed automating ergonomic assessment methods relying on computer vision and machine learning techniques (Diego-Mas and Alcaide-Marzal, 2014). In particular, the Kinect camera alongside its software development kit (SDK) have been extensively used to analyze the adopted posture and evaluate the RULA score (Plantard et al., 2015; Liebregts et al., 2016; Plantard et al., 2017; Manghisi et al., 2017; Abobakr et al., 2017a). The Kinect SDK tracks the human body and estimates the 3D Cartesian coordinates of 20 joint positions. It uses a random decision forest classifier to segment the body into parts followed by a localization algorithm to infer joint positions (Abobakr et al., 2017a). However, there are several difficulties resulting from using the Kinect SDK. First, it relies on local body part detectors, and hence may produce unrealistic skeletons in cases of occlusions due to cluttered environments (Abobakr et al., 2018; Plantard et al., 2017). Also, the Kinect SDK has a difficulty in tracking self-occluded postures that have arms crossing, trunk bending, trunk lateral flexion and trunk rotation (Manghisi et al., 2017). This requires applying preprocessing operations to correct the resulting kinematic structure as suggested in (Plantard et al., 2015; Plantard et al. Multon). Second, an additional processing stage is required to convert 3D Cartesian coordinates of body joint positions into joint angles. For instance, Plantard et al. (Plantard et al., 2017) corrected Kinect data using the method presented in (Plantard et al., 2017) and estimated missing anatomical landmarks using the approach proposed in (Bonnechere et al., 2014), to make the reconstructed skeleton compatible with the ISB recommendations (Wu et al., 2005) and compute the joint angles. Clark et al. (2012) used the inverse tangent method to convert 3D joint positions into joint angles. Although these approaches have been successful in obtaining joint angles of high quality, they may exhibit large errors from relying on the Kinect skeleton data especially in cases of occluded postures (Plantard et al., 2017; Manghisi et al., 2017). Improving the quality of the Kinect skeleton data for ergonomic studies is an open area of research (Plantard et al., 2017). This work focuses more on addressing limitations of the Kinect V1 sensor, as it uses the structured light technology which has been incorporated in a wide range of depth sensors (Abobakr et al., 2018). This allows better generalization to different depth cameras, for instance ASUS Xtion. The Kinect V2, on the other hand, uses the time of flight imaging technology which helps produce more robust skeleton data, however, it consumes more power and requires cooling (Fankhauser et al., 2015).

In this paper, we propose a skeleton-free holistic posture analysis system that accurately predicts body joint angles from a single depth image without utilizing the temporal information between subsequent images, as shown in Fig. 1. Although incorporating a temporal dynamics modeling stage can help ensure consistency of subsequent frame predictions and achieve higher frame rates, tracking algorithms require regular initialization to avoid leading to drift anomalies (Shotton et al., 2013). The fundamental building block of the proposed method is a cascade of two deep convolutional neural network (ConvNet) models. The depth sensor produces two synchronized video feeds of RGB and depth images. First, we segment the body from the background via passing the RGB image to an object instance segmentation deep ConvNet model. This network computes segmentation masks for a predefined set of objects in a given scene. We apply the obtained person's segmentation mask to the corresponding depth image. Second, depth values of the posture are encoded using a proposed depth encoding algorithm. Third, the encoded image is passed through the second ConvNet model to predict body joint angles. Finally, the estimated joint angles are used to compute the RULA score. Thus, we simplify the overall ergonomic evaluation procedure to be as simple as mapping directly predicted joint angles into a RULA score. Using this score, the MSD risk level is identified and a recommended action is suggested to decrease the risk of work-related injuries as defined in (McAtamney and Corlett, 1993). This is made possible via training our models on a large amount of highly varied synthetic training images with ground truth joint angles that have been biomechanically modeled using a novel inverse kinematics step.

The remainder of this paper is structured as follows. Section 2 describes the proposed method and the used deep ConvNet models. Section 3 presents the experiments and results. Key aspects and limitations of the proposed method are discussed in Section 4. Section 5 highlights the conclusion and future work.

2. Material and methods

We propose a vision based ergonomic posture assessment system composed of two cascaded ConvNets; an object instance segmentation network and a holistic posture analysis network. We utilize both the RGB video and depth feeds of a low cost depth sensor. The input feeds are synchronized which means that each RGB image has an associated depth image. In particular, we employed the segmentation network to detect and segment the person from an input RGB image and reject other background objects. This network produces a segmentation mask for the person in the scene which is then applied to the corresponding depth image. Hence, the proposed system is background independent and can be implemented in any environmental setting. Then, we trained a deep ConvNet model to learn a direct mapping from the segmented depth image of a human posture to body joint angles. The estimated joint angles are used to ergonomically assess the adopted posture by computing the RULA score. The main key for the proposed method is relying on learning from synthetic training images with biomechanically modelled body joint angles. Learning from synthetic depth images has facilitated research and demonstrated effectiveness in several domains (Saleh et al., 2016; Abobakr et al.,2016; Haggag et al. Haggag; Shotton et al., 2013; Abobakr et al., 2017b).

2.1. Data preparation

Training deep learning models is an optimization process with respect to millions of parameters. Therefore, they require large and highly varied training datasets to achieve proper generalization and avoid the risk of overfitting. However, collecting a labelled training dataset of postures for workers of different anthropometric measures is an expensive process. It is also not feasible to cover all rendering scenarios and simulations in real work environments. Also, manually labelling each posture image with the respective joint angles is a hard task that requires expert knowledge and can still remain prone to the inter- and intra-rater variability problem (Manghisi et al., 2017).

Therefore, we build a synthetic data generation pipeline from which we can sample large amounts of training images with plausible reference joint angles. These joint angles are modeled using an inverse kinematic (IK) method that adopts a skeletal model to constraint the movement of body joints in a structured manner. Learning from synthetic training images allows easily simulating a wide range of working postures, anthropometric measurements and rendering parameters for different tasks in model training. The pipeline also adds Kinect noise to the generated data to ensure realistic rendering. Hence, this results into more generalizable and highly invariant learning models at a negligible data preparation cost. These models directly approximate the mapping from a single depth image to a joint angles posture vector from which we easily obtain the RULA score. Thus, the proposed method bypasses the preprocessing steps that were used in the literature to obtain body joint angles.

We cover a wide range of anthropometric measures to ensure that our models generalize well to unseen body shapes. The weight is

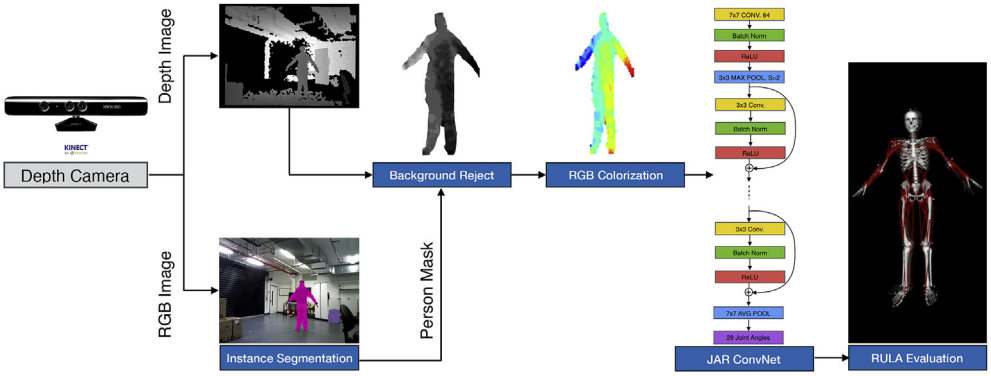


Fig. 1. Ergonomic posture assessment method overview. First, we segment the body from the background via passing the RGB image to an object instance segmentation deep network. This network computes segmentation masks for a predefined set of objects in a given scene. We apply the obtained person's segmentation mask to the depth image. Second, depth values of the posture are encoded using our proposed depth encoding algorithm. Third, the encoded image is passed through our joint angles regression (JAR) deep learning model to predict body joint angles. Finally, the estimated joint angles are used to compute the RULA score and evaluate the exposure to MSDs risk factors. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

reported in kilograms and the remaining measurements are in centimeters.

2.1.1. Synthetic depth generation

Synthetic depth images are generated via rendering animated human models of different shapes and sizes. We used six virtual human models; two males, two females and two neutral bodies generated using MakeHuman software. The anthropometric measures of the used models are detailed in Table 1. Since the depth images are texture invariant, we did not have to apply clothing styles.

The virtual models are animated into realistic postures via retargeting postural information from motion capture (mocap) sequences, as shown in Fig. 2. A mocap sequence is a representation of human movements captured using a marker-based motion capture system. It is composed of the trajectories of 3D Cartesian coordinates of markers attached to a subject. We used the Carnegie Mellon University (CMU) mocap database (raphics Lab Motion C) in generating our training dataset. This database contains a wide range of human activities recorded using a VICON mocap system for real human actors. It features a highly varied set of postures that covers most of upper body articulations involved in manual tasks. In earlier works (Abobakr et al., 2017a, 2017b), we recorded our own mocap dataset for a worker performing 3 different manual tasks. Nevertheless, it does not contain as much varied ranges of postures as the CMU dataset. Therefore, to ensure a more generalized solution, we decided to generate the postures using CMU mocap data. However, the CMU database was recorded at a frame rate of 120 frames per second (FPS), thus it contains many redundant postures. We downsampled the mocap sequences to 1 FPS to reduce this redundancy. Then, we chose the most dissimilar 10 postures from each mocap sequence. This creates a dataset of 3650 sparse postures.

The selected postures are retargeted to articulate the 3D models. We rendered synthetic images for the animated models using 8 virtual Kinect depth sensors with view angles ranging from 0 to 315° with step 45° and different depth distances. The maximum depth distance for rendering was set to 10 m. Fig. 2 depicts the virtual scene and the setup of virtual Kinect depth cameras used in generating our dataset. The scene is created and rendered using the open source software BlenSor (Gschwandtner et al., 2011).

Table 1
Anthropometric measures of 3D virtual human models used in generating training images.

3D Model	Weight	Height	Chest	Waist	Hips
Male_1	107	188	109	89	104
Male_2	84	173	108	98	106
Female_1	78	184	99	78	103
Female_2	61	159	64	87	106
Neutral_1	90	191	103	83	103
Neutral_2	71	166	100	92	105

The rendered depth images are clean with high quality depth measurements. However, deep learning models are generally sensitive to noise patterns augmented on the input data. This is an unsolved problem that is receiving high research interest. We anticipate that to cause an issue with our models, as the real depth sensors exhibit noise due to several environmental effects such as illumination, infrared (IR) interference from ambient light sources and non-IR-reflecting materials (Shotton et al., 2013). Therefore, to ensure resilience with real depth cameras, we utilized the sensor simulation capabilities of BlenSor to obtain Kinect scans that resemble the real sensor output. For each scene scan, BlenSor generates a clean depth image and two noisy images using a realistic and statistically verified noise model (Gschwandtner et al., 2011). This setting creates a synthetic dataset of 350K images that we split into 280K for training and 70K images for validation.

2.1.2. Inverse kinematics

Deriving a deep learning model for a RULA based ergonomic feedback system requires kinematically plausible reference joint angles for training. However, the reference data obtained from the marker-based mocap system represents the 3D Cartesian coordinates of the attached markers. Therefore, an inverse kinematics (IK) step is proposed to transform the 3D marker positions into joint angles. This step employs a skeletal model that is augmented with a set of virtual markers representing bony landmarks. Each virtual marker corresponds to a marker on the mocap data. The skeletal model is animated by minimizing the error between the corresponding marker positions in the skeletal model and in the captured data. The error minimizing process is constrained by different joint angle constraints, i.e. each joint can have a limited range of motion to ensure a natural, realistic human movement. This is done through solving a weighted least-squares problem using a generic quadratic programming solver with a convergence criterion of 10^{-4} and a limit of 1000 iterations, which is implemented in OpenSim platform (Delp et al., 2007; Seth et al., 2011; Reinbolt et al., 2011). The minimization function is

$$\sum_{i \in m} w_i \, ||x_i^{exp} - x_i(q)||^2, \tag{1}$$

where m is the set of markers, ω_i is a weighting factor, q represents the required coordinates, x_i^{exp} and $x_i(q)$ are the i^{th} marker position in the captured marker trajectory and on the model, respectively.

Biomechanical analysis and simulation has been used extensively in assessing workplace activities and activities of daily living (Nimbarte et al., 2013; Vignais et al., 2013; Weston et al., 2017; Hossny et al., 2012; Nahavandi et al., 2016). Therefore, there are various biomechanical models for different parts of the human body such as the upper limbs (Holzbaur et al., 2005; Wu et al., 2016), lower limbs (Delp

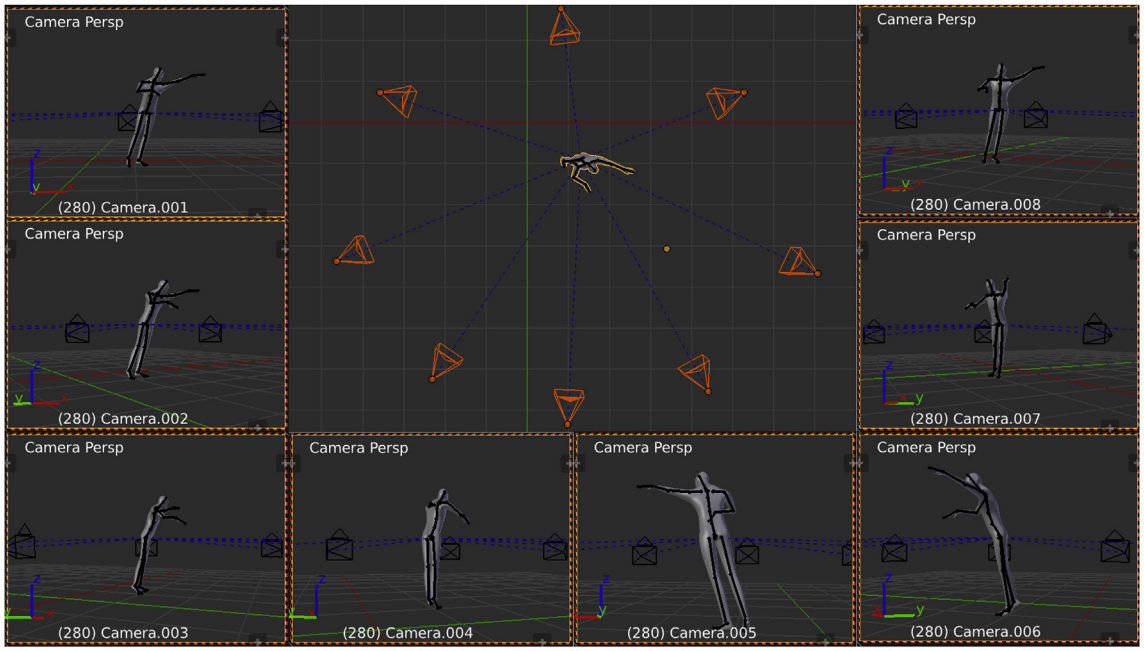


Fig. 2. Synthetic depth generation. In BlenSor, we built a scene containing one 3D model at a time. The model is articulated using retargeted CMU mocap data. We render the scene using eight Kinect sensors with view angles ranging from 0 to 315° with step 45° at different depth distances. Kinect noise is augmented during the rendering process to ensure generalization to data acquired using real depth cameras. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

et al., 1990), and also the eyes (Iskander et al., 2017, 2018a, 2018b). The recorded human activities are simulated and analyzed via the inverse kinematics modelling performed on the skeleton of the Upper and Lower Body Model (ULBmodel) available in OpenSim. This model is an almost full body model, since the neck degree of freedoms are not included. It is a combination of the upper limb model developed by Holzbaur et al. (2005) with the lower limb model developed by Delp et al. (1990).

We are interested in the joint angles that rotate the trunk, shoulders, elbows, and wrists, shown in Fig. 3. The trunk has three degrees of freedom (DoF), as shown in Fig. 3 (a), flexion, lateral bend and twist. The shoulder being a complex joint, needed two angles to describe the elevation movement, the plane of elevation and the shoulder elevation angle (Holzbaur et al., 2005), as shown in Fig. 3(b). Shoulder rotation was ignored since it is not used in RULA scoring. The lower arm movement is described using the elbow and wrist joints. Fig. 3(c) illustrates the configurations of the elbow flexion angle, and Figures (d)-(f) show the configurations of the wrist twist, deviation and flexion joints (Holzbaur et al., 2005). To ensure smooth natural movement, we applied kinematic constraints on the range of motion for the upper body joints, as discussed in (Rajagopal et al., 2016; Holzbaur et al., 2005). The used dynamic ranges of the upper body joints are listed in Table 2.

Kinematic constrains applied on the range of motion for the upper body joint angles to ensure smoother movements (Rajagopal et al., 2016; Holzbaur et al., 2005). L/R prefixes refer to left and right sides respectively.

The kinematic modeling of mocap sequences allowed obtaining kinematically plausible joint angles. The generated synthetic depth images and corresponding joint angles constituted a large labelled dataset for training the deep ConvNet regression model.

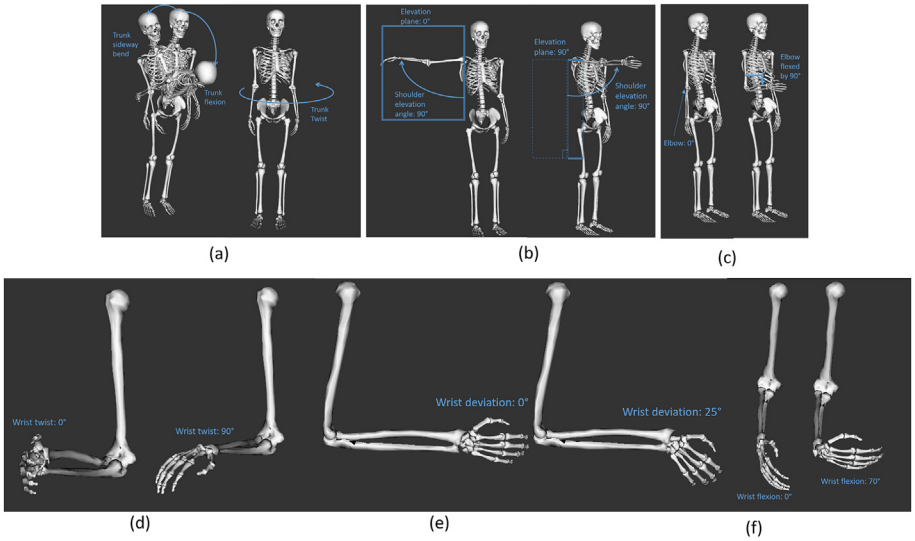


Fig. 3. The different joint angles used. (a) The trunk has 3 DoF, side bending, flexion-extension and twisting. (b) The shoulder joint utilized two angles, the shoulder elevation angle and the elevation plane angle. (c) The elbow rotates in a flexion-extension direction only. The rest of the figures shows the 3 DoF of the wrist. (d)–(f) show wrist twist, deviation and flexion, respectively. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Kinematic range constraints of upper body joints.

Joint name	Dynamic range
Trunk rotation	[−90°, 90°]
Trunk twist	[−90°, 90°]
Trunk bend	[−90°, 90°]
L. Elevation	[−90°, 130°]
R. Elevation	[−90°, 130°]
L. Shoulder	[0°, 180°]
R. Shoulder	[0°, 180°]
L. Elbow	[0°, 130°]
R. Elbow	[0°, 130°]
L. Wrist flexion	[−70°, 70°]
R. Wrist flexion	[−70°, 70°]
L. Wrist deviation	[−10°, 25°]
R. Wrist deviation	[−10°, 25°]
L. Wrist twist	[−90°, 90°]
R. Wrist twist	[−90°, 90°]

Fig. 4. Example results of applying the proposed colorization method on synthetic depth images. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.2. Encoding depth images

The use of depth imaging technologies offers several advantages. First, depth cameras operate independently of scene lighting conditions. Second, they are color and texture invariant (Shotton et al., 2013). Therefore, it is much easier to synthesize realistic depth images. Third, depth sensors facilitate the background subtraction which is an essential preprocessing step for our method, as it is not feasible to generate enough diverse background scenarios for training. Nevertheless, the main limitation of depth images is the weak local gradient information of objects. This issue limits the generalization capabilities of the ConvNet models and biases the network towards detecting object silhouettes (Abobakr et al., 2016).

Therefore, deep learning from depth images has become an active area of research (Abobakr et al., 2016; Couprie et al., 1301; Farabet et al., 2013; Gupta et al., 2014; Eitel et al., 2015). Farabet et al. (2013) suggested an approach to fuse the depth information with RGB images to obtain more expressive learning signal. This method led to good results on the challenging semantic segmentation tasks. On the other hand, it does rely on learning from two modalities. Gupta et al. (2014) proposed extracting three features from depth pixels; horizontal disparity, surface normal and height above the ground. Their approach demonstrated better performance than learning from either raw depth or replicated depth over three channels. However, computing the three features is computationally expensive (Eitel et al., 2015).

Recently, RGB colorization methods (Eitel et al., 2015; Abobakr et al., 2016) achieved better generalization performance and computational efficiency than the aforementioned approaches. In these methods, the depth pixels are shifted to $(0 - 255)$ range and a jet color map is applied to represent each pixel using three RGB channels. This results into a colorized depth image that provides much richer contrast information.

In this work, we propose a normalized encoding method that ensures depth invariant color encoding. The proposed method achieves better and faster generalization performance than the most widely used plane RGB colorization method (Gupta et al., 2014). First, we standardize the depth image by removing the mean and scaling to unit variance. This step ensures robustness to noise that the depth sensor may exhibit and faster convergence. Second, the depth values are transformed to $(0 - 255)$ range and a jet color map is applied. Fig. 4 shows example results of applying the proposed colorization method on synthetic depth images.

2.3. Joint angles estimation using ConvNet

The deep convolutional neural networks (ConvNet) is a class of deep learning models that has a high capacity to approximate an end-to-end mapping function from raw input to target output. This is achieved via a stack of computational layers that learns a hierarchy of features from raw input data (LeCun et al., 2015).

A ConvNet model is a composition of convolution (CONV), subsampling or pooling (POOL) and optionally fully connected (FC) layers at the end. The output of each layer is called feature maps that are passed through a non-linear transformation such as the rectified linear unit (ReLU). An efficient composition of such transformations is capable of approximating complex mapping functions regardless of its complexity (LeCun et al., 2015). Each CONV layer attempts to learn different patterns from its input feature maps, hence building more abstract representation from the raw data. CONV layers are parameterized by kernel size or local receptive field and number of kernels. Each kernel learns a set of weights to extract a certain feature wherever it appears in the input and produces a feature map. The spatial POOL layer reduces the dimensionality of feature maps via merging semantically similar features. The two main POOL operations are average or max pooling. Deep ConvNet models are prone to overfitting due to the high learning capacity they provide. Therefore, dropout (Srivastava et al., 2014) and batch normalization (BN) (Ioffe and Szegedy, 2015) layers have been recently introduced to control the effect of overfitting and ensure faster convergence. Despite their complexity, deep ConvNet models are end-to-end trainable architectures that can be optimized using generic optimization techniques such as the stochastic gradient descent (SGD).

The ergonomic posture assessment task is formulated as a supervised regression problem. The input is a depth image of the posture and the output is the joint angles vector required for computing the RULA score. Therefore, given a dataset $D = \{(x_i, t_i)\}_{i=1}^{N}$ of N samples, where $x_i \in R^d$ is an input depth image and $t_i = (a^{(1)}, a^{(2)}, ...,a^{(k)})$, $t_i \in R^k$ is the reference vector of $k = 15$ joint angles, we approximate a function that maps unseen input images of working postures to joint angles. The resulting posture vector is used to compute the RULA score and identify the MSD risk level and the urgency of intervention to decrease the risk.

2.3.1. Deep residual learning: ResNet

We trained the deep residual network (ResNet) model (He et al., 2016) to map from an input depth image to body joint angles. The ResNet is the state-of-the-art ConvNet architecture for visual perception tasks (He et al., 2016). The design principles of the ResNet follow the residual learning paradigm where layers learn a residual function with reference to layers input instead of learning direct mapping. The ResNet model is easy to optimize, computationally efficient and achieves significant performance gains with increased network depth (He et al., 2016).

The ResNet model is a composition of stacked residual blocks. The

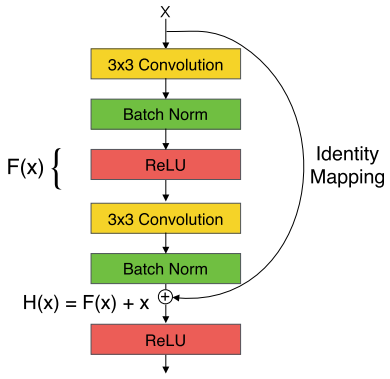


Fig. 5. The architecture of the residual learning block (He et al., 2016). Block layers learn a residual mapping function with reference to the block input. The identity shortcut implements a parameter-free mapping of the input.

most common architecture of the residual block is shown in Fig. 5. This block implements the design principles of the ResNet: residual learning and identity mapping via shortcut connections. Given a mapping function $H(x)$ to be learned, the residual block tries to learn the residual $F(x) = H(x) - x$. Therefore, the function to be learned becomes $F(x) + x$ (He et al., 2016). The identity mapping of the input x is implemented via parameters-free shortcut connections. The depth of the ResNet model is of crucial importance and is defined by the number of stacked residual blocks.

In our experiments, we built two ResNet models of depth 18 and 34 to estimate body joint angles. The differences between both networks are the number of CONV filters and the replication pattern of the residual block. ResNet-18 has 8 stacked residual blocks giving a total of 18 trainable layers. This model is described as: {RGB input - (CONV–BN - ReLU - max POOL) - 8 residual blocks - average POOL - 15 regressors}. On the other hand, the ResNet-34 model is described as: {RGB input - (CONV–BN - ReLU - max POOL) - 16 residual blocks - average POOL - 15 regressors}.

2.3.2. Models training

We initialized our models with pre-trained feature extractors that were optimized on the ImageNet dataset to discriminate between 1000 natural object categories from RGB images. This practice is known as fine-tuning and it has been proven more effective for many applications than starting the training with random weight initialization (Abobakr et al., 2016). However, it requires relatively large amounts of training data to tune the overall model, the feature extractor and the regressor, to predict joint angles from an input depth image and achieve the desired generalization performance. Therefore, using the data generation pipeline described in Section 2.1, we generated a synthetic dataset of 350K labelled images covering 8 camera view angles from 0 to 315° with step 45° and featuring 6 subjects of different anthropometric measures, detailed in Table 1. The dataset is split into 280K for training and 70K for validation.

The training objective function is minimizing mean square error over a training mini-batch:

$$E_{train} = \frac{1}{N}\sum_{i=1} (H(x_i, W) - t_i)^2 \quad (2)$$

where N is mini-batch size, $H(x_i, W)$ is the predicted joint angles vector for sample x_i given the set of weights W and t_i is the target joint angles vector. We used SGD optimization with an initial learning rate of 0.01, decaying by a factor of 10 every 50 epochs, mini-batch size of 32, weight decay of 0.0001 and momentum of 0.9.

2.4. Background rejection

Generating synthetic training images with different backgrounds is challenging due to the wide range of variations in scene setups and object configurations that may occur in real work environments. Therefore, the posture analysis network is trained on a background-free images. This focuses the analysis to the posture in the input image and requires removing the background during real time deployments. Several approaches in the literature (Abobakr et al., 2016, 2018) made use of the fact that depth images facilitates background subtraction. These approaches require modelling the background as an initial calibration stage. Then, at run time, the modelled background is subtracted and post processing operations follow to remove any remaining residuals. The main limitations for these approaches are; requiring calibration whenever the scene configuration changes which is not practical in manufacturing environments and the run-time complexity of the post processing operations.

Towards obtaining more generic solution, we employ the state-of-the-art fully convolutional instance segmentation network (FCIS) (Li et al., 2017) to segment the person from the background. It has achieved state-of-the-art performance in terms of accuracy and efficiency on the COCO 2016 segmentation challenge administered by Microsoft (Lin et al., 2014). The FCIS network detects and produces segmentation masks of all objects in the scene including the person of interest. The person mask is applied on the depth image and hence obtain a background-free image. This allows the system to operate in dynamic environments without the need for background calibration. Fig. 6 shows sample person segmentation and background rejection results using FCIS with a ResNet-101 (He et al., 1703) backbone network.

2.5. RULA score computation

The predicted body joint angles are the input for the RULA score computation module, which represents the final stage of the proposed system. Fig. 7 shows a snapshot of the proposed system in action. The grand RULA score is computed using the angular thresholds and adjustment parameters defined in the standard RULA worksheet (McAtamney and Corlett, 1993). It is a single page worksheet that divides the body into two sections.

The first section A includes the upper arm, lower arm and wrist segments. The estimated elevation plane and shoulder elevation angles are used to score the upper arm position and detect the shoulder raise and upper arm abduction adjustment parameters. As shown in Fig. 3(b), the upper arm position is scored using the standard RULA thresholds on the predicted shoulder elevation angle. To detect the upper arm abduction, we use the elevation plane angle that determines the plane in which the arm is moving. The upper arm is considered abducted if the elevation plane angle is 0° indicating frontal plane movement, and the shoulder elevation angle is greater than 45° as suggested in (Vignais et al., 2017). The shoulder raise occurs when the arm is raised upward (Vignais et al., 2017). We assume that to happen when the arm is above the horizontal, which refers to a shoulder elevation angle greater than 90°. Shoulder raise in the neutral posture is not considered in this study. The arm support parameter is disabled by default and can be enabled by the operator via the graphical user interface (GUI).

The lower arm is scored based on the predicted elbow flexion angle, as shown in Fig. 3(c). The arms crossing parameter is also disabled by default and can be enabled by the operator. For the wrist position, we apply the thresholds on the predicted wrist flexion angle as shown in Fig. 3(f). The proposed method predicts the wrist radio-ulnar deviation angle, as shown in Fig. 3(e). Therefore, we activate the wrist bending from the midline flag when this angle is below − 5° (radial deviation) or greater than 10° (ulnar deviation) as defined in (Vignais et al., 2017). However, we adjusted the radial deviation threshold to − 5° compared to the − 10° value defined in (Vignais et al., 2017) due to the range of motion for our wrist deviation joint which is [−10°, 25°]. Finally, the proposed approach also estimates the wrist twist, as shown Fig. 3(d). It is considered a mid-range if the absolute value of the predicted twist

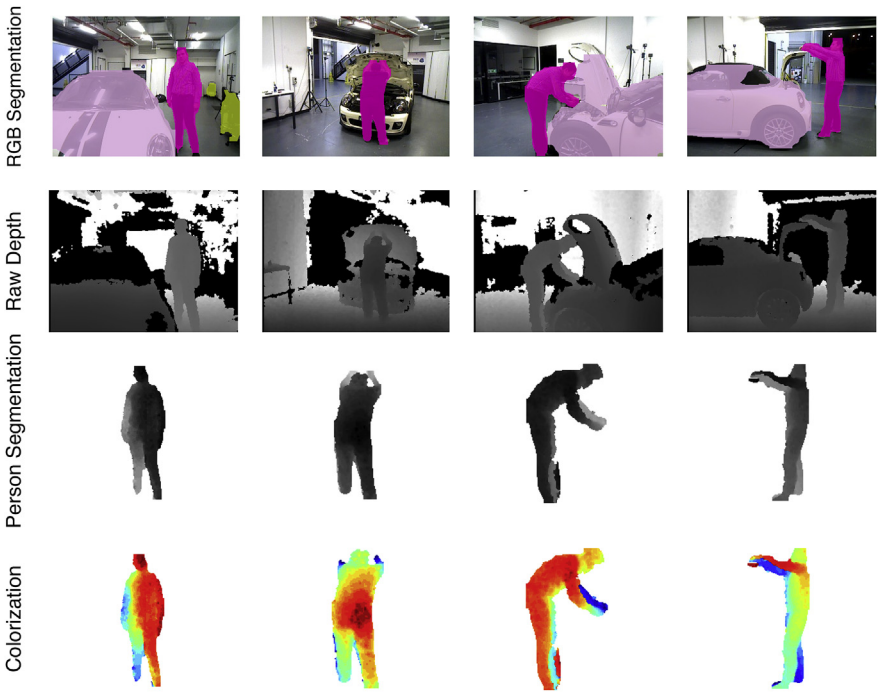


Fig. 6. Example background rejection results on real RGB and depth image pairs. The FCIS network computes segmentation masks for scene objects (Li et al., 2017). We use the person mask to segment the person and remove any other background objects. The resulting human posture is then pre-processed using the proposed colorization method and passed via the trained ConvNet model to estimate body joint angles. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

angle is below 45°, otherwise a near end of range twist is considered.

The second section B covers the neck, trunk and leg segments. The trunk position is scored directly using the estimated flexion, bend and twist angles, as shown in Fig. 3(a). The trunk lateral bend and twist have binary flags that contribute to the RULA score. We activate these flags based on a threshold value of 10° on the predicted bend and twist angles, as suggested in (Vignais et al., 2017). For the neck, due to the aforementioned neck joint limitation, we set its joint angle to be in

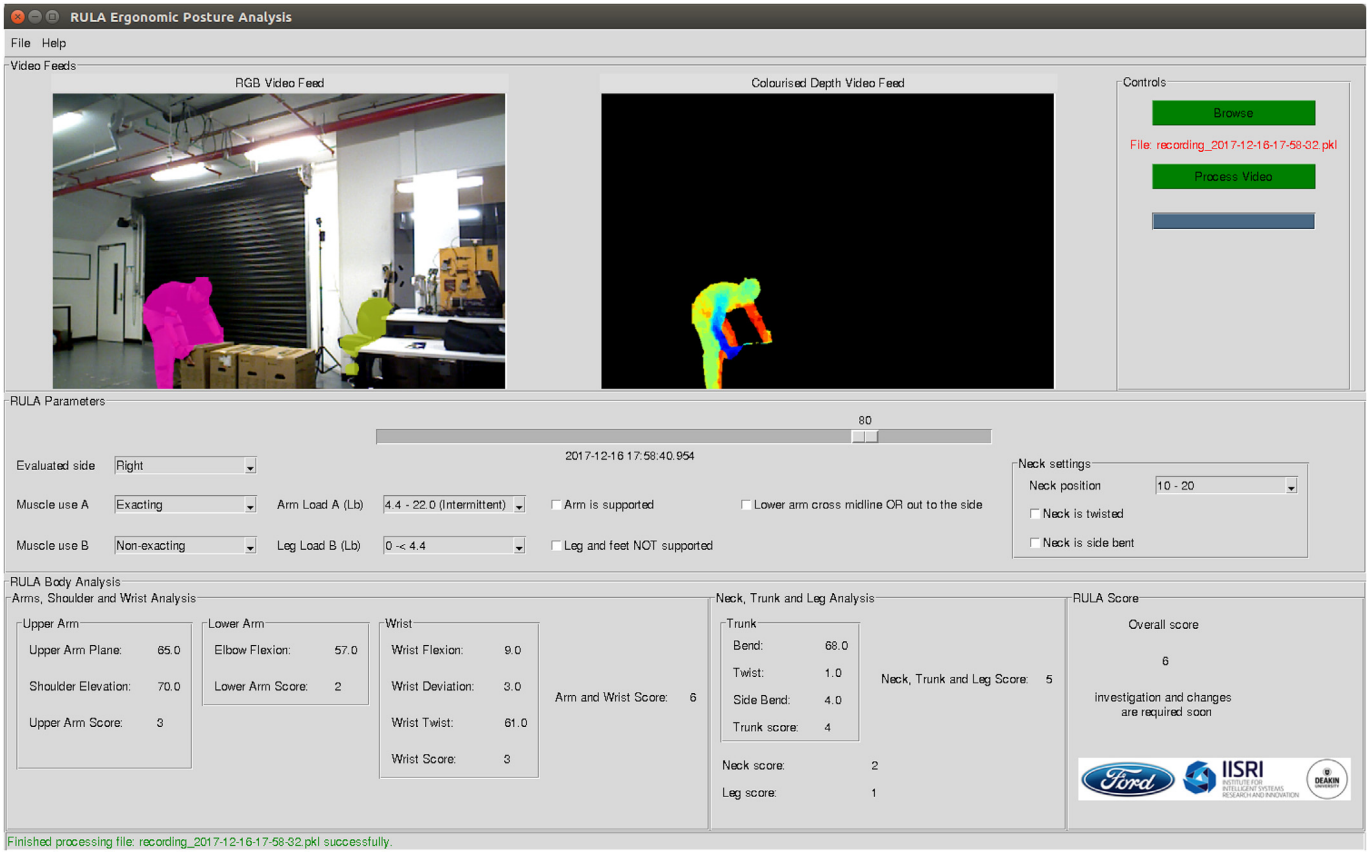


Fig. 7. The proposed system in action. A recorded session is fed to the software for analysis. From a single RGB-D images pair, we estimate the body joint angles of the segmented person and compute the RULA score via the standard angular thresholds and the adjustment parameters. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
RULA action levels and urgency of intervention.

Score	Level	Urgency of intervention
1–2	1	the posture is acceptable if it is not maintained or repeated for long periods
3–4	2	further investigation is needed and changes may be required
5–6	3	investigation and changes are required soon
7	4	investigation and changes are required immediately

range [0°, 10°] with neither twisting nor bending, and allowed the operator to override these settings through the GUI. Finally, the legs and feet are considered supported and can be changed by the operator through the GUI.

Each of these sections has a corresponding table that is used to obtain an intermediate score based on the scores for its segments. This score is further adjusted according to additional parameters such as muscle use, force load and frequency of operation. These parameters contribute effectively to the grand RULA score. However, they are difficult to be automated (Manghisi et al., 2017; Plantard et al., 2017). Therefore, we have set a default setting for these parameters and allowed the operator to override them in the GUI, as shown in Fig. 7. The third and final table returns the grand RULA score associated with the previous two scores. This score is mapped to an action level that indicates the urgency of the intervention required to decrease the likelihood of MSD injuries, as defined in (McAtamney and Corlett, 1993) and listed in Table 3.

The RULA grand score and action levels indicate the risk of MSD injuries and the urgency of intervention to reduce this risk (McAtamney and Corlett, 1993; Manghisi et al., 2017).

3. Results

We have trained a deep ConvNet model on a synthetic training dataset to predict 15 body joint angles from a single depth image. The estimated joint angles are then used to compute the RULA score for the adopted posture. In this section, we evaluate the performance of the proposed method and explore the generalization capabilities on a real test dataset. Table 4 provides a detailed description of the synthetic and real datasets used in training and evaluating the proposed system. We also examine the effect of several aspects on the generalization performance of our method. We report the mean absolute (MAE) and root mean square (RMSE) error rates.

Details of the datasets used in training and evaluating the proposed system. We report the number of subjects, number of samples, image modalities, the mocap system used in collecting the motion sequences and the depth sensor used for generating or recording the images.

3.1. Comparison between ResNet-18 and 34

Fig. 8 shows a per-joint MAE of both models on the synthetic validation set. The reported errors are scaled by the biomechanical range of motion of the joints, listed in Table 2. These results demonstrate that deeper ResNet-34 achieves better generalization performance than ResNet-18 for all body joints.

There are several reasons for ResNet-34 performing better than ResNet-18. First, deeper models have a higher learning capacity and can learn more powerful representations. Second, the employed deep residual learning framework makes efficient use of network depth while ensuring easy optimization and fast convergence (He et al., 2016). However, after a certain extent, the models saturate and the extra runtime computational cost resulting from going deeper becomes a challenge. Also, in some cases this may lead to performance degradation due to overfitting.

The improvements that ResNet-34 provides are slightly significant compared to the added runtime computational complexity, as shown in the benchmark in Table 5. That makes it challenging to deploy the overall system on embedded devices. Further, RULA scores for body limbs are evaluated based on angular thresholds, which means that we can compensate these little improvements without affecting the final outcome of the RULA metric. Therefore, we chose to use ResNet-18 for body joint angles regression. The remainder of this section investigates the effect of depth preprocessing and the generalization capabilities of ResNet-18 to real depth images.

Benchmarking ResNet-18 and ResNet-34 models for joint angles prediction. Errors are evaluated on the 70K validation images and none of these images is included in models training. As shown, ResNet-34 has a slightly better generalization performance than ResNet-18 with extra runtime computational cost. JTX2 refers to the embedded Nvidia Jetson TX2 GPU device.

3.2. The effect of depth encoding

We compared the effect of the proposed depth encoding method with the standard RGB colorization method (Gupta et al., 2014; Abobakr et al., 2016). We trained the ResNet-18 model using each encoding method on the same training set. Table 6 reports the prediction errors achieved using each encoding method on the validation set. The proposed colorization method achieves better generalization performance and faster convergence than the RGB colorization approach.

We compared the proposed depth encoding method with the state-of-the-art RGB colorization method (Gupta et al., 2014). The reported average joint angle prediction MAE and RMSE errors are evaluated using the ResNet-18 model trained on the synthetic dataset.

3.3. Generalization to real data

To validate the performance of the proposed method on real depth images, we recorded a real dataset of 24K postures for 6 subjects of different body shapes while doing a set of manual tasks. The anthropometric characteristics of the real test subjects are detailed in Table 7. We used an XSENS mocap system and an ASUS Xtion depth camera for recording the images. The three feeds are synchronized in such a way that each posture frame in the mocap sequence has a corresponding RGB and depth images pair from the camera. Joint angles for the recorded motion sequences are generated using the biomechanical model described in Section 2.1.2.

We collected the real test dataset from 6 male subjects with different body shapes and sizes. The reported measurements are in centimeters and the weight is in kilogram. The Sh. Prefix refers to the shoulder body part.

We fine-tune the trained ResNet-18 model on the recorded real

Table 4
Description of datasets used in this study.

Dataset	Subjects	Samples	Modalities	MoCap system	Depth camera
Synthetic training	6	280,000	Depth	VICON (raphics Lab Motion C)	Virtual Kinect
Synthetic validation	6	70,000	Depth	VICON (raphics Lab Motion C)	Virtual Kinect
Real fine-tuning	6	19,000	RGB, Depth	XSENS	ASUS Xtion
Real validation	6	5000	RGB, Depth	XSENS	ASUS Xtion

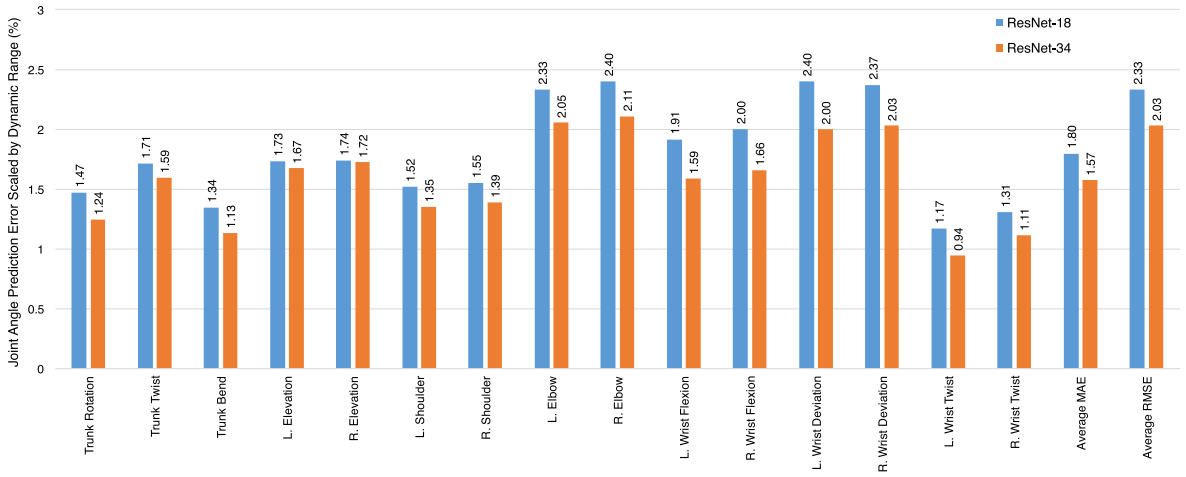


Fig. 8. Comparison of the generalization performance of the two deep residual models on the validation set of 70K images. We report the MAE error of each joint scaled by its range of motion, defined in Table 2. None of the validation images is included in training our models.

Table 5
Benchmarking deep residual models for holistic ergonomic posture analysis.

Method	Capacity (Milions)		Prediction Errors		FPS		
	Parameters	Neurons	MAE	RMSE	Core-i7 CPU	Titan-X GPU	JTX2
ResNet-18	12	12	2.60 ± 1.44	3.45 ± 2.07	20	250	50
ResNet-34	22	18	2.31 ± 1.31	3.07 ± 1.86	10	125	30

Table 6
The effect of depth encoding on prediction errors of ResNet-18.

Encoding method	MAE	RMSE
RGB Colorization (Gupta et al., 2014)	3.00 ± 1.61	3.98 ± 2.32
Proposed	2.60 ± 1.44	3.45 ± 2.07

dataset. The dataset is split into 19K postures for training and 5K for validation, as detailed in Table 4. The motivation is to capture the data distribution of the real depth sensor and to learn the shape difference that the segmentation mask may cause. Hence, this ensures proper generalization to real data. It is worth noting that this is not a calibration process, so it is not required during the real time deployment. We record a final average MAE error of 3.19 ± 1.57° and an average RMSE of 4.27 ± 2.32° in real test environments. Table 8 details a per-joint breakdown of these results. Since the joint angles have different ranges of motion, we report the error of each joint scaled by its dynamic range, defined in Table 2. The results demonstrate low prediction error rates for most of the joints and difficulties in estimating elbow and wrist joint configurations. However, as the RULA score is evaluated based on angular thresholds, it is not susceptible to small error variations. Further, these errors are highly unlikely to change the final RULA score. As a result, the proposed system achieves a RULA grand score prediction accuracy of 89% with a substantial Kappa index of 0.71. This accuracy represents the average percentage of correct RULA score predictions over both right and left body sides in comparison with scores computed using reference mocap based joint angles. The RULA scores are computed using the standard angular thresholds and the default parameter settings described in Section 2.5.

Table 8
Prediction errors on real data.

Joint name	MAE (deg.)	Scaled MAE (%)	RMSE (deg.)	Scaled RMSE (%)
Trunk rotation	3.23 ± 3.34	1.79 ± 1.86	4.64	2.58
Trunk twist	3.13 ± 3.05	1.74 ± 1.70	4.37	2.43
Trunk bend	2.30 ± 2.09	1.16 ± 1.28	3.10	1.72
L. Elevation	4.13 ± 4.70	1.88 ± 2.14	6.26	2.84
R. Elevation	4.04 ± 4.34	1.83 ± 1.98	5.93	2.70
L. Shoulder	4.19 ± 4.32	2.33 ± 2.40	6.02	3.34
R. Shoulder	4.27 ± 4.65	2.40 ± 2.59	6.31	3.51
L. Elbow	4.14 ± 4.54	3.18 ± 3.49	6.14	4.72
R. Elbow	4.19 ± 4.97	3.22 ± 3.82	6.50	5.00
L. Wrist flexion	2.59 ± 2.46	1.85 ± 1.76	3.58	2.56
R. Wrist flexion	2.76 ± 2.72	1.97 ± 1.94	3.87	2.77
L. Wrist deviation	1.06 ± 1.06	3.03 ± 3.04	1.50	4.29
R. Wrist deviation	1.15 ± 1.22	3.29 ± 3.49	1.68	4.79
L. Wrist twist	3.05 ± 2.75	1.70 ± 1.53	4.11	2.28
R. Wrist twist	3.58 ± 3.21	1.99 ± 1.79	4.81	2.67
Average	3.19 ± 1.57	2.23 ± 1.12	4.27 ± 2.32	2.94 ± 1.64

Table 7
Anthropometric measures of the real test subjects.

Subject	Height	Weight	Arm Span	Hip Height	Hip Width	Sh. Height	Sh. Width
1	168	75	166	94	34	135	47
2	182	88	182	90	27	155	48
3	167	97	168	91	28	135	37
4	184	95	180	102	30	157	39
5	175	78	178	97	27	144	38
6	179	85	184	102	31	153	48

Table 9
The effect joint angle errors on RULA postural scores.

RULA Score	RMSE	Accuracy P_o	kappa (k)
Upper arm Right	0.29	0.92	0.88
Upper arm Left	0.32	0.90	0.86
Lower arm Right	0.22	0.95	0.82
Lower arm Left	0.20	0.96	0.84
Wrist score Right	0.50	0.78	0.67
Wrist score Left	0.50	0.78	0.67
Score A (arm and wrist) Right	0.39	0.86	0.78
Score A (arm and wrist) Left	0.41	0.84	0.76
Score B (neck, trunk and legs)	0.64	0.82	0.63
RULA Grand Score Right	0.49	0.86	0.66
RULA Grand Score Left	0.51	0.85	0.67

Per-joint MAE prediction errors of the proposed system on real depth images. The scaled errors, by ranges of motion in Table 2, show that elbow and wrist joints are the most challenging for prediction. L/R prefixes refer to left and right sides respectively.

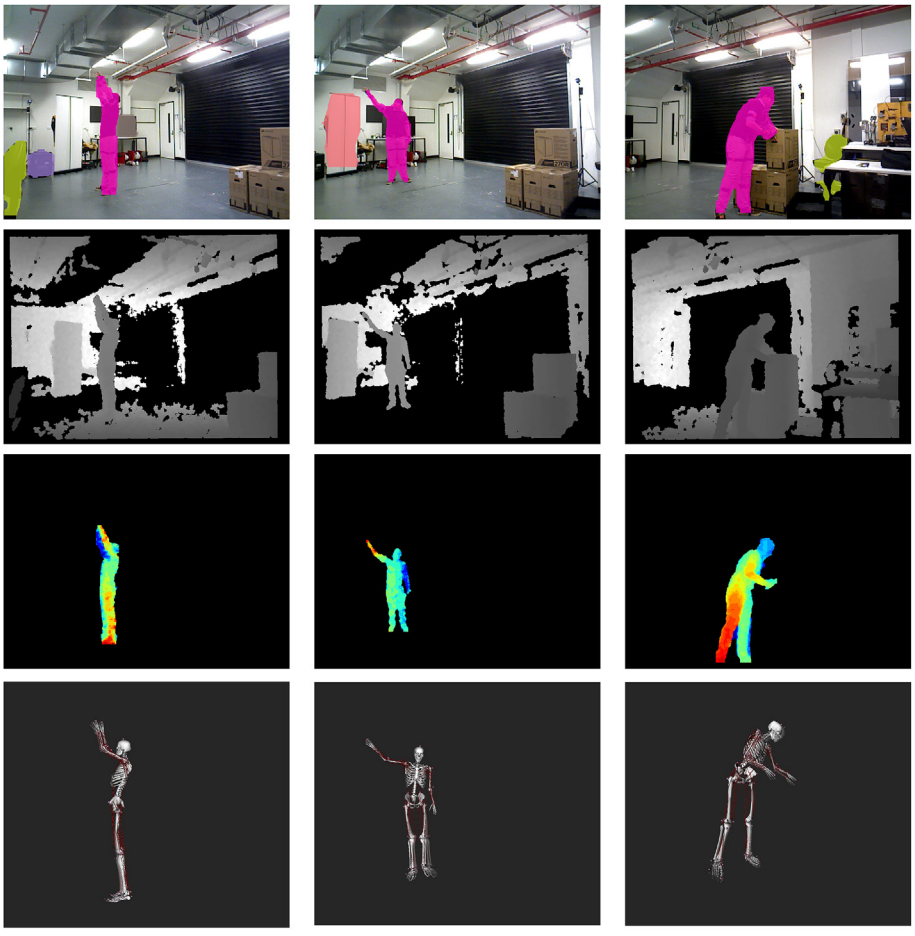
We report RMSE of RULA scores, P_o and Cohen's kappa index, between RULA scores computed using predicted joint angles and reference joint angles generated from recorded mocap data in real environment.

3.4. RULA score analysis

Further, we study the effect of the reported joint angle errors on the final RULA scores in real conditions. Table 9 reports the RMSE, accuracy or agreement values p_o and level of agreement (Cohen's kappa) (Cohen, 1960), between RULA scores computed from estimated joint angles using the proposed method and scores computed using reference joint angles. The reference joint angles are very accurate as they are generated from recorded mocap sequences in real conditions. Thus, the mocap system represents the expert observations. We achieve high grand score accuracy and a substantial strength of agreement according to the scale of (Landis and Koch, 1977).

3.5. Qualitative results

Fig. 9 shows example inferences of the proposed ergonomic posture analysis system on a set of real frames, where the first two rows display input RGB and depth image pairs captured using an ASUS Xtion sensor. The third row displays the encoded posture after rejecting the background. The colorized images are then passed to the ResNet-18 model for estimating the body joint angles. Predicted joint angles are applied to the biomechanical model in OpenSim for comparison with the input posture in the forth row. The proposed method does not rely on either joint positions prediction or posture calibration. Moreover, using the holistic posture analysis approach allows the system to be robust in



(a) RULA Grand Score Right: 4
RULA Grand Score Left: 4
Score A Right (arm and wrist): 6
Score A Left (arm and wrist): 5
Score B (neck, trunk and legs): 2
Action: 2, further investigation is needed and changes maybe required

(b) RULA Grand Score Right: 4
RULA Grand Score Left: 2
Score A Right (arm and wrist): 6
Score A Left (arm and wrist): 2
Score B (neck, trunk and legs): 2
Action: 2, further investigation is needed and changes maybe required

(c) RULA Grand Score Right: 4
RULA Grand Score Left: 5
Score A Right (arm and wrist): 2
Score A Left (arm and wrist): 4
Score B (neck, trunk and legs): 5
Action: 3, investigation and changes are required soon

Fig. 9. Example ergonomic posture analysis on real test images. The first two rows display input RGB and depth image pairs captured using an ASUS Xtion sensor. The results of person segmentation followed by depth encoding preprocessing are shown in the third row. The forth row shows predicted joint angles applied to the biomechanical model in OpenSim. The transparent skeleton represents the reference joint angles, and the displacement between skeletons is the model prediction error. The RULA scores are computed using the default parameter settings described in Section 2.5. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

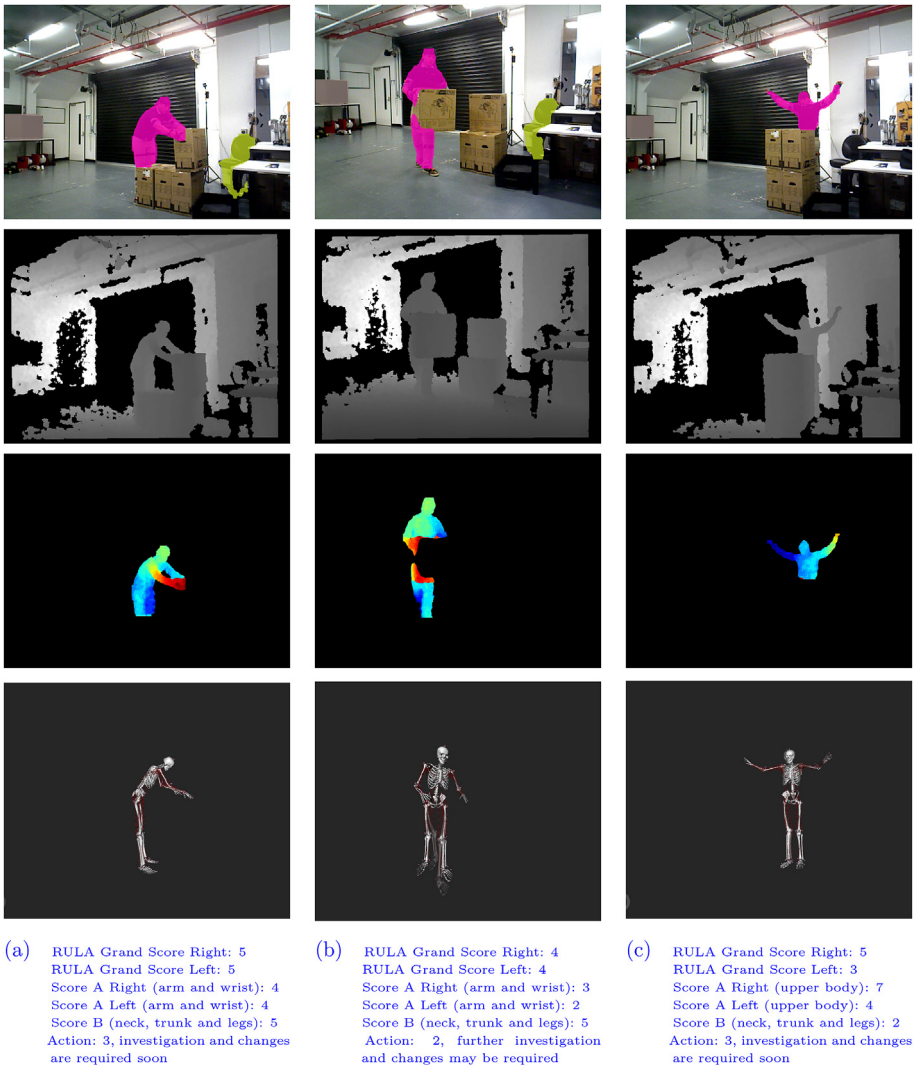


Fig. 10. Example system inferences in case of occlusions. The proposed method shows robustness to invisible body parts due to self-occlusions or clutters. The forth row shows predicted joint angles applied to the biomechanical model in OpenSim. The RULA scores are computed using the default parameters setting described in Section 2.5. The transparent skeleton represents the reference joint angles, and the displacement between skeletons is the model prediction error. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

cases of invisible body parts due to self-occlusions or cluttered environments, as shown in Fig. 10.

3.6. Comparison with the Kinect SDK in occluded cases

The Kinect depth sensor and its SDK have demonstrated effectiveness and acceptable accuracy when used in computerizing observational ergonomic assessment metrics (Plantard et al., 2017). However, the Kinect SDK has a limitation in estimating joint positions of occluded postures due to clutters or self-occlusions (Plantard et al., 2017). Therefore, we followed the holistic reasoning approach to learn kinematically valid skeletal structures via exploiting the full posture context. Hence, this ensures robustness to invisible body parts due to occlusions or cluttered environments. Fig. 11 shows example inferences using the Kinect SDK and the proposed method respectively, in challenging conditions. This figure confirms the Kinect difficulties discussed in this study and surveyed in the literature (Plantard et al., 2017; Manghisi et al., 2017). It also demonstrates challenges of Kinect SDK to estimate posture information from a side view of the human body.

3.7. Frame rate

The trained ergonomic posture analysis model ResNet-18 achieves up to 30 FPS on a MacBook Pro with Core-i7 CPU, up to 250 FPS on a Nvidia Titan X GPU and 50 FPS on the Jetson TX-2 embeded GPU device. The reported frame rate includes the preprocessing depth encoding step. However, the main bottleneck is the FCIS segmentation network due to relying on ResNet-101 as a backbone model. The overall system runs at a frame rate of up to 5 FPS on the TITAN-X GPU. Therefore, the proposed system supports an offline analysis mode for recorded sessions.

4. Discussion

This paper proposed a semi-automated ergonomic assessment system of adopted working postures. The proposed method analyzes the posture holistically and estimates body joint angles directly from a single depth and RGB image pair. Hence, we do not exploit temporal dependencies or skeleton data from the Kinect SDK. The estimated joint angles are used to compute the RULA score, the MSD risk level and subsequently the urgency of intervention required to reduce the risk of injury. The RULA score is computed based on the standard angular

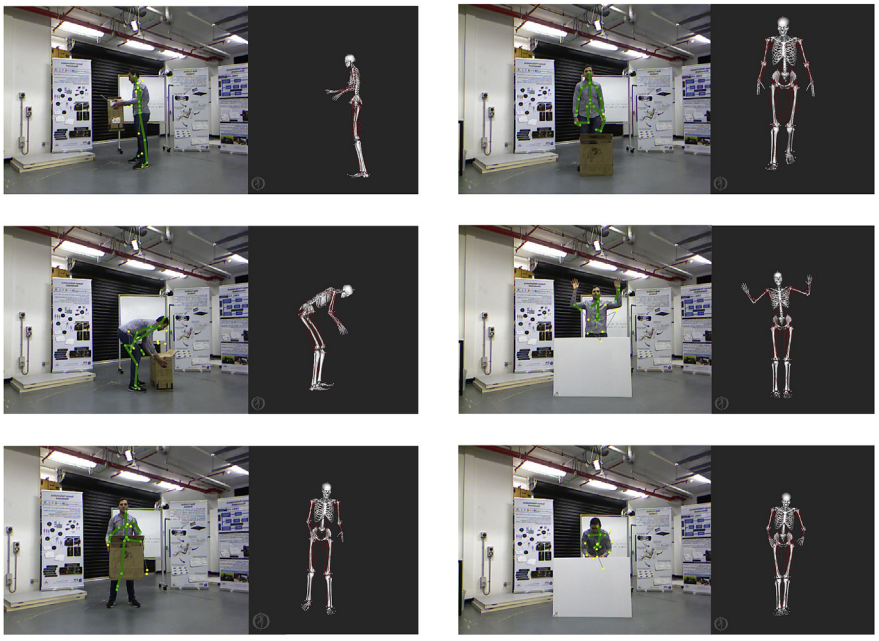


Fig. 11. The proposed method in comparison with the Kinect SDK in existence of occlusions and different view angles. The green skeleton in the RGB image is obtained from the Kinect SDK, and the skeletal model represents our predictions. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

thresholds and a default and easily adjustable parameters setting.

The proposed mapping, from input images to joint angles, has been approximated via training deep machine learning models on a highly varied set of synthetic depth images with biomechanically modelled reference joint angles. The biomechanical modelling stage ensures learning valid skeletal structures. To the best of our knowledge, this is the first work that proposes a method to obtain body joint angles directly from an input image of a posture without using skeleton data from the Kinect SDK. Therefore, the proposed approach could help accelerate the development of vision based ergonomic assessment systems via providing necessary information such as joint angles. This allows training machine learning models directly on high quality and kinematically modelled joint angles and provide opportunities to incorporate and evaluate different sensors and different image modalities such as RGB color images. The current implementation supports Microsoft Kinect and ASUS Xtion depth cameras. It can also be extended to accommodate any depth sensor as described in (Saleh et al., 2017).

We have validated the proposed system using an XSENS mocap system and an ASUS Xtion depth camera. The validation procedure has been done via recording mocap data synchronized with RGB and depth image pairs for 6 subjects of different anthropometric measures. The recorded mocap data are used to generate the reference joint angles via an inverse kinematics process in OpenSim software. The reference RULA scores are then computed based the resulting joint angles and the default parameters setting discussed in Section 2.5. Thus, in this setting, the mocap system represents the expert assessment. We achieved a joint angle MAE error of 3.19 ± 1. 5° and RMSE error of 4.27 ± 2.32° and an average RULA grand score prediction agreement of 89% over both right and left body sides, with a substantial Kappa index level of 0.71. Further, the proposed method demonstrated robustness to self-occlusions and missing body parts due to cluttered environments as shown in Fig. 10. We have also qualitatively compared the inferences of the proposed method with the predictions of the Kinect SDK in challenging conditions as depicted in Fig. 11. The holistic reasoning approach allowed the proposed method to be more robust to occlusions and clutters than the Kinect SDK.

4.1. Limitations

There are three main limitations for the current implementation of the proposed system. First, the used biomechanical model does not support the neck joint which contributes in evaluating the grand RULA score. In the used model, the torso, neck and skull are all acting as one body part with no degrees of freedom between them. This is done for simplification and faster biomechanics simulation times. The more degrees of freedom, the more complex is the model which means longer processing time. Furthermore, due to this complexity, the most recent full body model published in (Rajagopal et al., 2016) did not have the neck degrees of freedom. As per our knowledge, there are no full body models that include degrees of freedom for the head, neck and spine together available in OpenSim software. However, there are studies that computed the degrees of freedom at the head, neck (Mortensen et al., 2018) and spine (Raabe and Chaudhari, 2016) levels separately. Therefore, we assumed the neck to be in range [0°, 10°] with neither twisting nor bending, and allowed the operator to adjust these settings through the GUI.

Second, our RULA score computation module assumes a default setting for parameters that are difficult to be automated such as the force load and muscle use, as well as the aforementioned neck parameters. We have also had to adjust the wrist radial deviation threshold defined in the literature to be compatible with our modeled ranges. Further, we also assumed that a shoulder raise occurs when the arms are above the horizontal which is detected when the shoulder elevation angle exceeds 90°. Thus, the proposed system is semi-automated and requires manual adjustments if necessary. Future advancements will focus on researching the development and validation of a more articulated skeletal model to overcome the neck limitation.

Third, the frame rate of the system is up to 5 FPS on a NVIDIA TITAN-X GPU. This high computational cost is mostly attributed to the person segmentation stage. The need for this stage was due to generating synthetic depth images without background, due to the difficulty of modeling diverse backgrounds. Hence, our trained models require removing the background as a preprocessing step. Several approaches have used calibration methods to build a model of the background which is then subtracted from the input image (Abobakr et al., 2018). However, recalibration is required on every change in scene configuration which makes this approach not practical for real work conditions with frequently changing scene settings. The instance segmentation approach we followed in this work overcomes this limitation and offers background independent person segmentation method with the aforementioned high computational cost. The proposed system will be enhanced via a multi-task model to perform the

segmentation and joint angles prediction tasks simultaneously in a single forward pass. Further, we will also enlarge the collected dataset of mocap sequences converted into joint angles and synchronized with RGB color images to allow training deep learning models for ergonomics evaluation from color cameras.

5. Conclusions

This paper proposed a semi-automated holistic ergonomic posture assessment system. It is composed of an instance segmentation model that detects and segments the person in the scene and a deep convolutional neural network that we trained to estimate body joint angles directly from a single depth image. The joint angles prediction model is trained on synthetic depth images. This allows simulating a wide range of manual tasks performed by workers of different body shapes and sizes from several view angles. The corresponding reference joint angles are generated using a biomechanical model. The proposed system does not require calibration or specific sensor placement, is marker-free, supports different depth sensors such as the Kinect and ASUS Xtion and does not rely on skeleton data. Moreover, it predicts body joint angles directly from the input image and achieves an average RULA grand score prediction accuracy of 89%, over both left and right body sides, with a substantial agreement of 0.71 Kappa index with reference scores. Thus, this system overcomes challenges inherited from using the Kinect skeleton data and accelerates developing vision based ergonomic assessment methods using different sensors and different image modalities via providing an approach to obtain necessary postural information.

Acknowledgement

This research was supported by the Institute for Intelligent Systems Research and Innovation (IISRI) at Deakin University, Australia. The project was funded via Ford's university research program (URP 2014-4055R), Ford Motor Co., USA. The data used in this project was obtained from the CMU graphics lab motion capture database, mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217, USA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apergo.2019.05.004.

References

Abobakr, A., Hossny, M., Nahavandi, S., 2016. Body joints regression using deep convolutional neural networks. In: Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on. IEEE, pp. 003281–003287.

Abobakr, A., Nahavandi, D., Iskander, J., Hossny, M., Nahavandi, S., Smets, M., 2017a. A kinect-based workplace postural analysis system using deep residual networks. In: Systems Engineering Symposium (ISSE), 2017 IEEE International. IEEE, pp. 1–6.

Abobakr, A., Nahavandi, D., Iskander, J., Hossny, M., Nahavandi, S., Smets, M., 2017b. Rgb-d human posture analysis for ergonomie studies using deep convolutional neural network. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2885–2890.

Abobakr, A., Hossny, M., Nahavandi, S., 2018. A skeleton-free fall detection system from depth images using random decision forest. IEEE Syst. J. (99), 1–12. https://doi.org/10.1109/JSYST.2017.2780260.

Bernard, B.P., Putz-Anderson, V., 1997Bernard and Putz-Anderson. Musculoskeletal Disorders and Workplace Factors; a Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back.

Bonnechere, B., Jansen, B., Salvia, P., Bouzahouene, H., Omelina, L., Moiseev, F., Sholukha, V., Cornelis, J., Rooze, M., Jan, S.V.S., 2014. Validity and reliability of the kinect within functional assessment activities: comparison with standard stereophotogrammetry. Gait Posture 39 (1), 593–598.

Bureau of Labor Statistics, US Department of Labor, 2016. Nonfatal Occupational Injuries and Illnesses Resulting in Days Away from Work in 2015. https://www.bls.gov/news.release/pdf/osh2.pdf, Accessed date: 22 February 2018 Online; accessed.

Clark, R.A., Pua, Y.-H., Fortin, K., Ritchie, C., Webster, K.E., Denehy, L., Bryant, A.L., 2012. Validity of the microsoft kinect for assessment of postural control. Gait Posture 36 (3), 372–377.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor Semantic Segmentation Using Depth Information, arXiv preprint arXiv:1301.3572..

Delp, S.L., Loan, J.P., Hoy, M.G., Zajac, F.E., Topp, E.L., Rosen, J.M., 1990. An interactive graphics-based model of the lower extremity to study orthopaedic surgical procedures. IEEE Trans. Biomed. Eng. 37 (8), 757–767.

Delp, S.L., Anderson, F.C., Arnold, A.S., Loan, P., Habib, A., John, C.T., Guendelman, E., Thelen, D.G., 2007. Opensim: open-source software to create and analyze dynamic simulations of movement. IEEE Trans. Biomed. Eng. 54 (11), 1940–1950.

Diego-Mas, J.A., Alcaide-Marzal, J., 2014. Using kinect sensor in observational methods for assessing postures at work. Appl. Ergon. 45 (4), 976–985.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., 2015. Multimodal deep learning for robust RGB-D object recognition, Intelligent Robots and Systems (IROS). In: IEEE/RSJ International Conference on, pp. 681–687.

Fankhauser, P., Bloesch, M., Rodriguez, D., Kaestner, R., Hutter, M., Siegwart, R.Y., 2015. Kinect v2 for mobile robot navigation: evaluation and modeling. In: 2015 International Conference on Advanced Robotics (ICAR). IEEE, pp. 388–394.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling, pattern analysis and machine intelligence. IEEE Transac. 35 (8), 1915–1929.

Gschwandtner, M., Kwitt, R., Uhl, A., Pree, W., 2011. Blensor: blender sensor simulation toolbox. Adv. Vis. Comput. 199–208.

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation. Comput. Vis. 345–360.

H. Haggag, M. Hossny, S. Nahavandi, O. Haggag, An adaptable system for RGB-D based human body detection and pose estimation: incorporating attached props, IEEE Conference on Systems, Man, and Cybernetics (SMC).

K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-Cnn, arXiv preprint arXiv:1703.06870..

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Holzbaur, K.R., Murray, W.M., Delp, S.L., 2005. A model of the upper extremity for simulating musculoskeletal surgery and analyzing neuromuscular control. Ann. Biomed. Eng. 33 (6), 829–840.

Hossny, M., Filippidis, D., Abdelrahman, W., Zhou, H., Fielding, M., Mullins, J., Wei, L., Creighton, D., Puri, V., Nahavandi, S., 2012. Low cost multimodal facial recognition via Kinect sensors. In: Proceedings of the Land Warfare Conference (LWC): Potent Land Force for a Joint Maritime Strategy. Commonwealth of Australia, pp. 77–86.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456.

Iskander, J., Hossny, M., Nahavandi, S., 2017. Simulating eye-head coordination during smooth pursuit using an ocular biomechanic model. In: Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on. IEEE, pp. 3356–3361.

Iskander, J., Hossny, M., Nahavandi, S., Del Porto, L., 2018a. An ocular biomechanic model for dynamic simulation of different eye movements. J. Biomech. 71, 208–216.

Iskander, J., Hossny, M., Nahavandi, S., 2018b. A review on ocular biomechanic models for assessing visual fatigue in virtual reality. IEEE Access 6, 19345–19361. https://doi.org/10.1109/ACCESS.2018.2815663.

Krüger, J., Nguyen, T.D., 2015. Automated vision-based live ergonomics analysis in assembly operations. CIRP Ann. - Manuf. Technol. 64 (1), 9–12.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. biometrics, pp. 159–174.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep Learning, Nature 521 (7553), 436–444.

Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y., 2017. Fully convolutional instance-aware semantic segmentation. In: IEEE Conf. On Computer Vision and Pattern Recognition. CVPR), pp. 2359–2367.

Liebregts, J., Sonne, M., Potvin, J., 2016. Photograph-based ergonomic evaluations using the rapid office strain assessment (rosa). Appl. Ergon. 52, 317–324.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

A. Luttmann, M. Jager, B. Griefahn, G. Caffier, F. Liebers, W. H. Organization, et al., Preventing Musculoskeletal Disorders in the Workplace..

Manghisi, V.M., Uva, A.E., Fiorentino, M., Bevilacqua, V., Trotta, G.F., Monno, G., 2017. Real time rula assessment using kinect v2 sensor. Appl. Ergon. 65, 481–491.

McAtamney, L., Corlett, E.N., 1993. Rula: a survey method for the investigation of work-related upper limb disorders. Appl. Ergon. 24 (2), 91–99.

Mortensen, J.D., Vasavada, A.N., Merryweather, A.S., 2018. The inclusion of hyoid muscles improve moment generating capacity and dynamic simulations in musculoskeletal models of the head and neck. PLoS One 13 (6) e0199912.

Nahavandi, D., Iskander, J., Hossny, M., Haydari, V., Harding, S., 2016. Ergonomic effects of using lift augmentation devices in mining activities. In: Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on. IEEE, pp. 002012–002019.

Nimbarte, A.D., Sun, Y., Jaridi, M., Hsiao, H., 2013. Biomechanical loading of the shoulder complex and lumbosacral joints during dynamic cart pushing task. Appl. Ergon. 44 (5), 841–849.

Plantard, P., Auvinet, E., Pierres, A.-S.L., Multon, F., 2015. Pose estimation with a kinect for ergonomic studies: evaluation of the accuracy using a virtual mannequin. Sensors 15 (1), 1785–1803.

Plantard, P., Shum, H.P., Multon, F., 2017. Filtered pose graph for efficient kinect pose reconstruction. Multimed. Tool. Appl. 76 (3), 4291–4312.

Plantard, P., Shum, H.P., Le Pierres, A.-S., Multon, F., 2017. Validation of an ergonomic assessment method using kinect data in real workplace conditions. Appl. Ergon. 65, 562–569.
Raabe, M.E., Chaudhari, A.M., 2016. An investigation of jogging biomechanics using the full-body lumbar spine model: model development and validation. J. Biomech. 49 (7), 1238–1243.
Rajagopal, A., Dembia, C.L., DeMers, M.S., Delp, D.D., Hicks, J.L., Delp, S.L., 2016. Full-body musculoskeletal model for muscle-driven simulation of human gait. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (10), 2068–2079.
CMU Graphics Lab Motion Capture Database, http://mocap.cs.cmu.edu.
Reinbolt, J.A., Seth, A., Delp, S.L., 2011. Simulation of human movement: applications using opensim. Procedia IUTAM 2, 186–198.
Saleh, K., Hossny, M., Hossny, A.H., Nahavandi, S., 2017. Cyclist detection in lidar scans using faster r-cnn and synthetic depth images. In: Intelligent Transportation Systems Conference (ITSC), 2017 IEEE International Conference on. IEEE.
Saleh, K., Hossny, M., Nahavandi, S., 2016. Kangaroo vehicle collision detection using deep semantic segmentation convolutional neural network. In: IEEE Conference on Digital Image Computing: Techniques and Applications (DICTA).
Seth, A., Sherman, M., Reinbolt, J.A., Delp, S.L., 2011. Opensim: a musculoskeletal modeling and simulation framework for in silico investigations and exchange. Procedia Iutam 2, 212–232.
Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al., 2013. Efficient human pose estimation from single depth images, Pattern Analysis and Machine Intelligence. IEEE Transac. 35 (12), 2821–2840.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Vignais, N., Miezal, M., Bleser, G., Mura, K., Gorecky, D., Marin, F., 2013. Innovative system for real-time ergonomic feedback in industrial manufacturing. Appl. Ergon. 44 (4), 566–574.
Vignais, N., Bernard, F., Touvenot, G., Sagot, J.-C., 2017. Physical risk factors identification based on body sensor network combined to videotaping. Appl. Ergon. 65, 410–417.
Weston, E., Le, P., Marras, W.S., 2017. A biomechanical and physiological study of office seat and tablet device interaction. Appl. Ergon. 62, 83–93.
Wu, G., Van der Helm, F.C., Veeger, H.D., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A.R., McQuade, K., Wang, X., et al., 2005. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion-part ii: shoulder, elbow, wrist and hand. J. Biomech. 38 (5), 981–992.
Wu, W., Lee, P.V., Bryant, A.L., Galea, M., Ackland, D.C., 2016. Subject-specific musculoskeletal modeling in the evaluation of shoulder muscle and joint function. J. Biomech. 49 (15), 3626–3634.