


## The accuracy of a 2D video-based lifting monitor

Xuan Wang<sup>a</sup>, Yu Hen Hu<sup>a</sup>, Ming-Lun Lu<sup>b</sup> and Robert G. Radwin<sup>c</sup> 

<sup>a</sup>Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, USA; <sup>b</sup>National Institute for Occupational Safety and Health, Taft Laboratories, Cincinnati, OH, USA; <sup>c</sup>Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA

### ABSTRACT

A widely used risk prediction tool, the revised NIOSH lifting equation (RNLE), provides the recommended weight limit (RWL), but is limited by analyst subjectivity, experience, and resources. This paper describes a robust, non-intrusive, straightforward approach to automatically extract spatial and temporal factors necessary for the RNLE using a single video camera in the sagittal plane. The participant's silhouette is segmented by motion information and the novel use of a ghosting effect provides accurate detection of lifting instances, and hand and feet location prediction. Laboratory tests using 6 participants, each performing 36 lifts, showed that a nominal 640 pixel × 480 pixel 2D video, in comparison to 3D motion capture, provided RWL estimations within 0.2 kg ( $SD = 1.0$  kg). The linear regression between the video and 3D tracking RWL was  $R^2 = 0.96$  (slope = 1.0, intercept = 0.2 kg). Since low definition video was used in order to synchronise with motion capture, better performance is anticipated using high definition video.

**Practitioner's summary:** An algorithm for automatically calculating the revised NIOSH lifting equation using a single video camera was evaluated in comparison to laboratory 3D motion capture. The results indicate that this method has suitable accuracy for practical use and may be, particularly, useful when multiple lifts are evaluated.

**Abbreviations:** 2D: Two-dimensional; 3D: Three-dimensional; ACGIH: American Conference of Governmental Industrial Hygienists; AM: asymmetric multiplier; BOL: beginning of lift; CM: coupling multiplier; DM: distance multiplier; EOL: end of lift; FIRWL: frequency independent recommended weight limit; FM: frequency multiplier; H: horizontal distance; HM: horizontal multiplier; IMU: inertial measurement unit; ISO: International Organization for Standardization; LC: load constant; NIOSH: National Institute for Occupational Safety and Health; RGB: red, green, blue; RGB-D: red, green, blue – depth; RNLE: revised NIOSH lifting equation; RWL: recommended weight limit; SD: standard deviation; TLV: threshold limit value; VM: vertical multiplier; V: vertical distance

### ARTICLE HISTORY

Received 7 January 2019  
Revised 28 March 2019  
Accepted 7 May 2019

### KEYWORDS

Manual lifting; NIOSH lifting equation; video; motion monitoring; lower back pain prevention

## 1. Introduction

Overexertion during manual lifting ranks first among the leading causes of disabling workplace injuries in the US, and in 2017 cost businesses \$13.79 billion in direct costs and accounted for 23% of the overall national burden (Liberty Mutual Research Institute for Safety 2017). In 2016, injuries and illness to the back resulted in higher rates of cases with both days away from work and days of job transfers or restricted work, compared to the head and hands (Bureau of Labor Statistics 2016).

The revised NIOSH lifting equation (RNLE) (Waters et al. 1993) is the most widely used tool to assess the

risk of low back pain associated with lifting and lowering tasks in the workplace (Lowe, Dempsey, and Jones 2018). The recommended weight limit (RWL) for occupational lifting, calculated by the RNLE, is defined for a specific set of task conditions as the weight of the load that nearly all healthy workers could perform over a substantial period without an increased risk of developing lifting-related low back pain. The RWL is implemented in the International Organisation for Standardisation (ISO) standard 11228 Part 1 Manual Lifting and Carrying, and routinely used by practitioners for prevention of lifting-related low back pain. The RWL is calculated as:

$$RWL = LC \times HM \times VM \times DM \times AM \times FM \times CM \quad (1)$$

where  $LC$  is a load constant equal to 23 kg,  $HM$ ,  $VM$  and  $DM$  are a function of the distance between the location of the hands and feet at the origin and destination of the lift,  $AM$  is a function of the angle of torso rotation in relation to the feet,  $FM$  is a function of the frequency of lifting, and  $CM$  is a function of the quality of the hand-to-object coupling.

Today's manual materials handling jobs have evolved from single tasks decades ago to multiple and varying tasks, specifically in the manufacturing and transportation sectors, such as warehousing, distribution centres, package delivery trucks, baggage handling, lean or just-in-time manufacturing, kitting, palletising and shipping. For prevention of work-related low back pain, it has, therefore, become increasingly important to frequently or continuously monitor workers' exposure to physical demands for varying job tasks.

Manually measuring the dimensions needed to calculate the RNLE is challenging, particularly in situations where lifting occurs in numerous locations involving varying body postures throughout the workday (Dempsey 2002). Direct measurement and instrument-based methods have been tested that employ sensors or markers attached to the participant for measuring certain variables (Li and Buckle 1999; Juul-Kristensen et al. 2001; Kim and Nussbaum 2013). However, these methods suffer from the invasiveness of the measurement, the space needed for equipment, and high cost (Patrizi, Pennestrì, and Valentini 2016).

With the advancement of video technologies, cameras have become a prevalent, low-cost, highly efficient, and non-intrusive option for monitoring ergonomic aspects of job tasks. Both 2D (RGB) and depth (RGB-D) cameras have attracted researchers' attention for developing direct-reading technologies for ergonomic risk assessments. In recent years, the development of depth cameras, e.g. Kinect (Microsoft, Redmond, WA), has stimulated a new trend in measuring human body dimensions. These approaches rely on matching the recognised human body parts against a pre-trained skeletal model, which consists of the position of joints and body linkages (Gabel et al. 2012). Previous researchers used conventional cameras for creating a skeletal model from a large image data set (Bogo et al. 2016; Howe, Leventon, and Freeman 2000; Mehrizi et al. 2017, 2018).

Although 2D cameras are more prevalent than depth cameras, 2D cameras propose a more challenging partial body recognition problem. With one more dimension of information, RGB-D camera-based methods have been shown to provide more accuracy than

RGB cameras (Patrizi, Pennestrì, and Valentini 2016; Spector et al. 2014; Delpresto et al. 2013; Plantard et al. 2017). Two studies (Patrizi, Pennestrì, and Valentini 2016; Spector et al. 2014) used the Kinect to measure RNLE parameters. However, these Kinect-based approaches were limited when there was occlusion; namely, when the Kinect did not capture a frontal view of the participant. Since the location of the hands and ankles for calculating the RNLE was dependent on the accumulation of intermediate predictions of each skeletal model linkage position, there were numerous distance errors.

In this paper, we describe a straightforward and practical video-based approach to automatically extract spatial and temporal factors necessary for applying the RNLE using a single 2D camera commonly available in hand-held mobile devices. It leverages motion information to directly detect the hand and ankle locations during lifting. It avoids the challenging partial body recognition problem by not depending on a skeletal model. The proposed method can be used to evaluate a work task with fewer computations and sufficiently high accuracy when compared with 3D camera approaches.

## 2. Methods

### 2.1. Lifting monitor algorithm

The lifting monitor algorithm takes advantage of motion information to distinguish the moving figure from the static background. Based on the spatial and temporal features of the video scene, a ghost effect is exploited to detect the lifting instance and hand locations. A rectangular bounding box drawn around the human figure is used to locate the feet for each video frame.

The algorithm contains three steps: (1) moving target detection, (2) action feature extraction, and (3) feature recognition, as shown in Figure 1. Each step is illustrated in Figure 2.

The moving target detection step leverages the fact that the figure moves during lifting while most of the surrounding environment is static, and thus a motion-based technique could segment out the moving items. In this algorithm, a mixture-of-Gaussian background subtraction algorithm (Zivkovic 2004; Zivkovic and Van Der Heijden 2006), known as MOG2, is applied. The background subtraction algorithm builds a background model for each pixel using a mixture of Gaussian distributions and updates the weights of the texture to represent the time proportions that the pixel colours remain in the scene. The result is an  $M \times N$  matrix map, where  $M$  and  $N$  are the height and

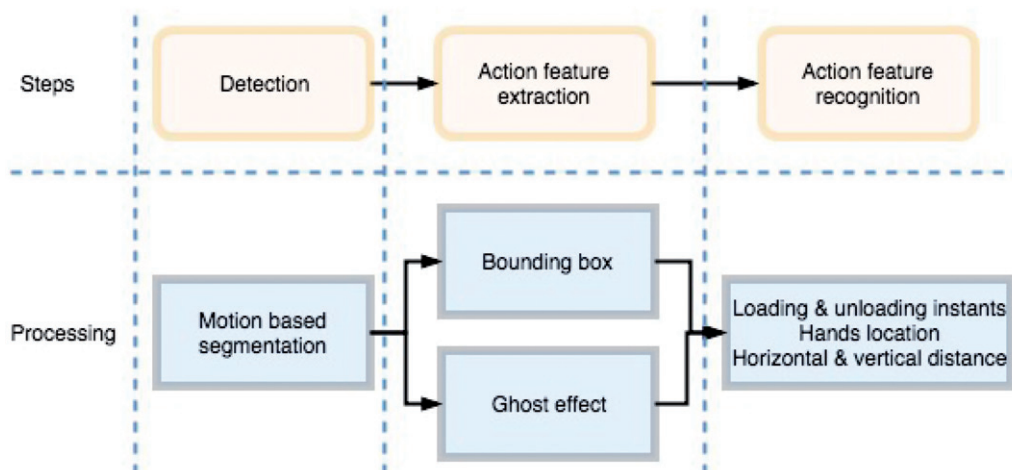


Figure 1. Flowchart of lifting monitoring algorithm.

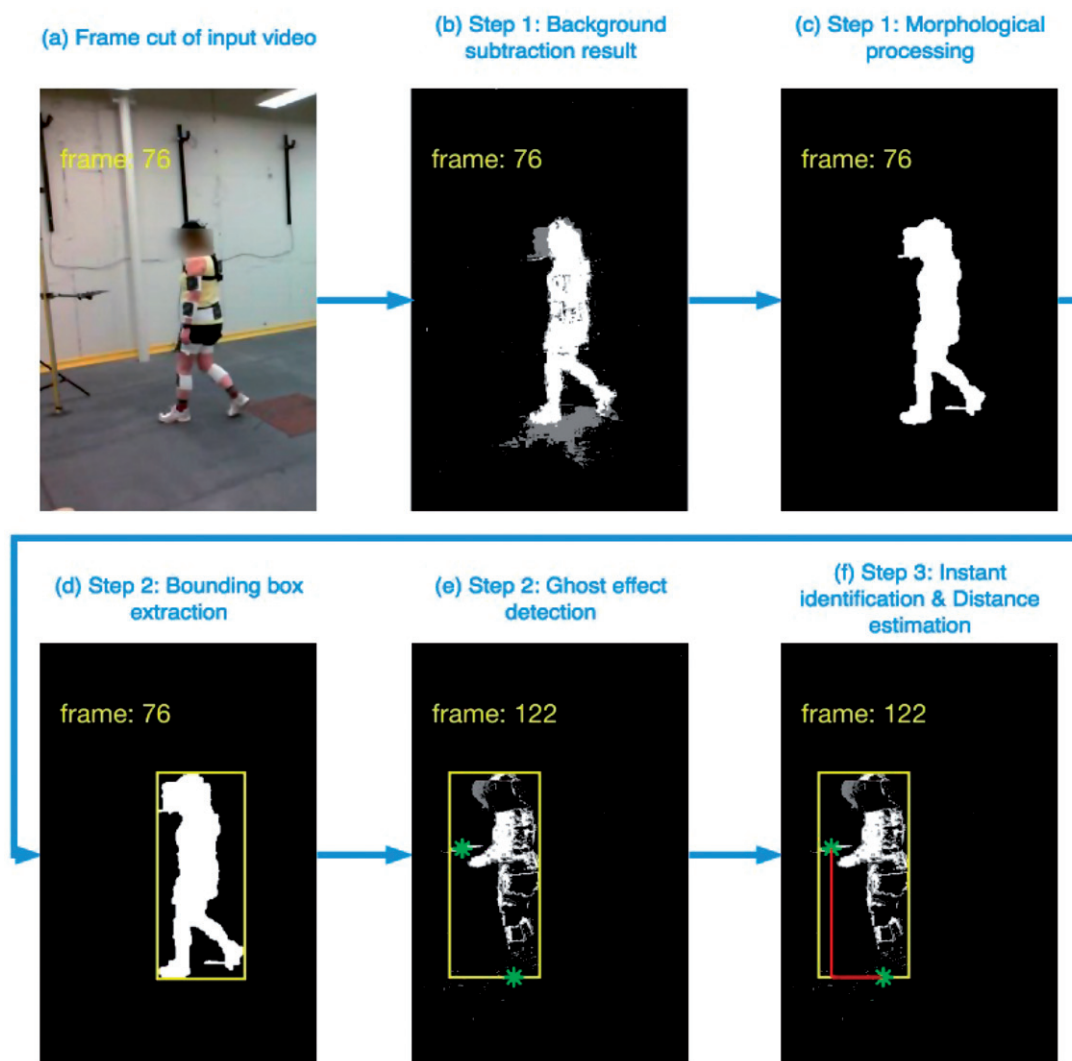


Figure 2. A demonstration of the video processing methodology for each step of the algorithm. Images (a), (b), (c) and (d) show the sequence of procedures for the 76<sup>th</sup> frame of a video where the participant walks to the lift location in the 122<sup>nd</sup> frame. Images (e) and (f) demonstrate hand and ankle detection at the 122<sup>nd</sup> frame of the video where the participant is lifting the object. A rectangular bounding box encloses the worker. The stars show the detected location of the hands and ankles. The lines connecting the stars indicate the horizontal distance from hands to the ankles and the vertical distance from the hands to the ground, respectively.

width of the image in the unit of pixels, labelling each pixel as foreground (moving), background (static), or shadow, with respective colours white, black and grey. Thus, the detected foreground can be identified pixel-wise by the white-colored foreground mask.

Each of the connected clusters of white pixels is referred to as a 'blob'. In a working environment where only the operator moves, the resulting map would capture the human figure perfectly with a foreground mask identifying the human figure's silhouette. But several moving objects and light reflections could introduce noise (false positives), and the foreground mask of the human figure might be incomplete and distorted (false negatives), as shown in Figure 2(b). These false detections are due to a false match to the background model. False detections are eliminated by morphological operations based on their size relative to the larger size of the subject's silhouette, as shown in Figure 2(c).

The action feature extraction step utilises information from the first step. The height and width of the human figure's silhouette are extracted, represented by a rectangular bounding box tightly covering the human figure's foreground mask (Figure 2(d)). Another feature utilised is a condition of the background subtraction, named the ghost effect, which is a set of connected points detected as in-motion, but not corresponding to any real moving object. This phenomenon appears when an object is moved away or set to a location, changing the appearance of an area (Figure 2(e)). It persists for a short period, depending on how fast the background subtraction restores the original appearance of that area. The ghost effect is detected in accordance with the following four properties:

1. Consistency in time. A blob of the ghost effect area should show up in the same location in subsequent  $N$  frames.
2. Gradual vanishing. The size of the blob should be no larger than the size of the lifted object and should gradually get smaller.
3. Proximity. When the ghost effect shows up, the human figure's silhouette (large-sized blob) should be close to it. In a non-ideal environment where there is more than one moving object, the ghost effect might occur in multiple places. Proximity is therefore used to focus on the ghost effect caused by the human figure.
4. Frame number. The frame number  $N$  is related to the duration that the ghost blob persists when the background model updates the new

appearance of the ghost effect area into the background model.

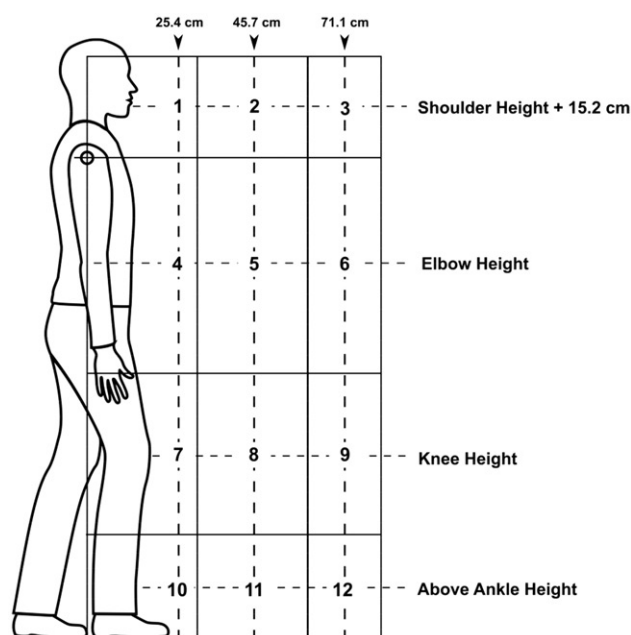
The feature recognition step exploits the fact that the ghost effect occurs when the appearance in a location changes when a stationary object is moved by the participant, and, therefore, the loading instance can be identified from the start of a ghost effect. At these detected lifting instances, the respective location of each ghost effect indicates where the hands were, and the bottom blobs of the silhouette show where the feet were during the initiation of the lift. If the participant is recorded in the view from the sagittal plane, and the hands and feet are moving symmetrically about the sagittal plane, the ghost blob centre can be used as an approximation of hands centre during the initiation of the lift. Similarly, when an object stops moving, such as when setting down an object at the termination of a lift, the ghost effect occurs and persists indicating the instance of release.

The geometric centre of the bottom portion of the silhouette blob (i.e. the lower 10% region of the bounding box while the participant is standing) can be used to approximate the horizontal location of the midpoint of the participant's ankles. Thus, the horizontal distance  $H$  from hands centre to ankles (for the  $HM$  calculation) and the vertical distance  $V$  from hands to the ground (bottom edge of the bounding box for  $VM$  and  $DM$  calculations) can be estimated (Figure 2(f)). The loading and unloading instances are important for  $H$  and  $V$  distance estimation because they indicate when to detect the hand and ankle locations. The algorithm considers the hands as the centre by the ghost blob, which does not move in time. Distances measured in pixels were calibrated against the participant's standing height.

## 2.2. Validation of the algorithm

### 2.2.1. Laboratory data

Data utilised in the current study were from a previous study conducted by researchers at the National Institute for Occupational Safety and Health (NIOSH). The aim of the original study was to record symmetrical lifting tasks using the combination of two lifting task variables (horizontal and vertical distances) defined by the American Conference of Governmental Industrial Hygienists (ACGIH) Threshold Limit Values (TLV) for lifting (ACGIH 2019). The TLV for lifting classifies 12 risk zones using the two task variables for symmetrical lifting in the sagittal plane. The horizontal distance ( $H$ ) is defined as the projected distance on



**Figure 3.** The starting point of 12 different lifting tasks was designed using the 12 risk zones of the ACGIH TLV for Lifting. The intersections of the dotted lines are the origins of the tasks and most were in the centre of the risk zones. The vertical heights for Tasks 1–3 were adjusted to 15.2 cm above the participants' shoulder height. The horizontal distances for tasks 1, 4, 7, and 10 were adjusted to 15.4 cm from the centre of the two ankles. These adjustments were made to create a realistic lifting motion. Adapted from ACGIH\_TLV lifting risk zone system.

the transverse plane from the centre of two ankles to the centre of two hands. The vertical distance ( $V$ ) is defined as the distance from the centre of two hands to the ground. The midpoints of risk zones were used as the positions for starting the lifting tasks, except for Zones 1–3, 4, 7 and 10. The alternative starting locations of the lift tasks for the exceptional zones were chosen for realistic lifting motion within each participant's reach envelope. The origins of the 12 lifting tasks in relation to the participant's neutral body position are shown in Figure 3. Each task was repeated three times for a total of 36 lifting trials for each participant. These trials were assigned to each participant in a random order.

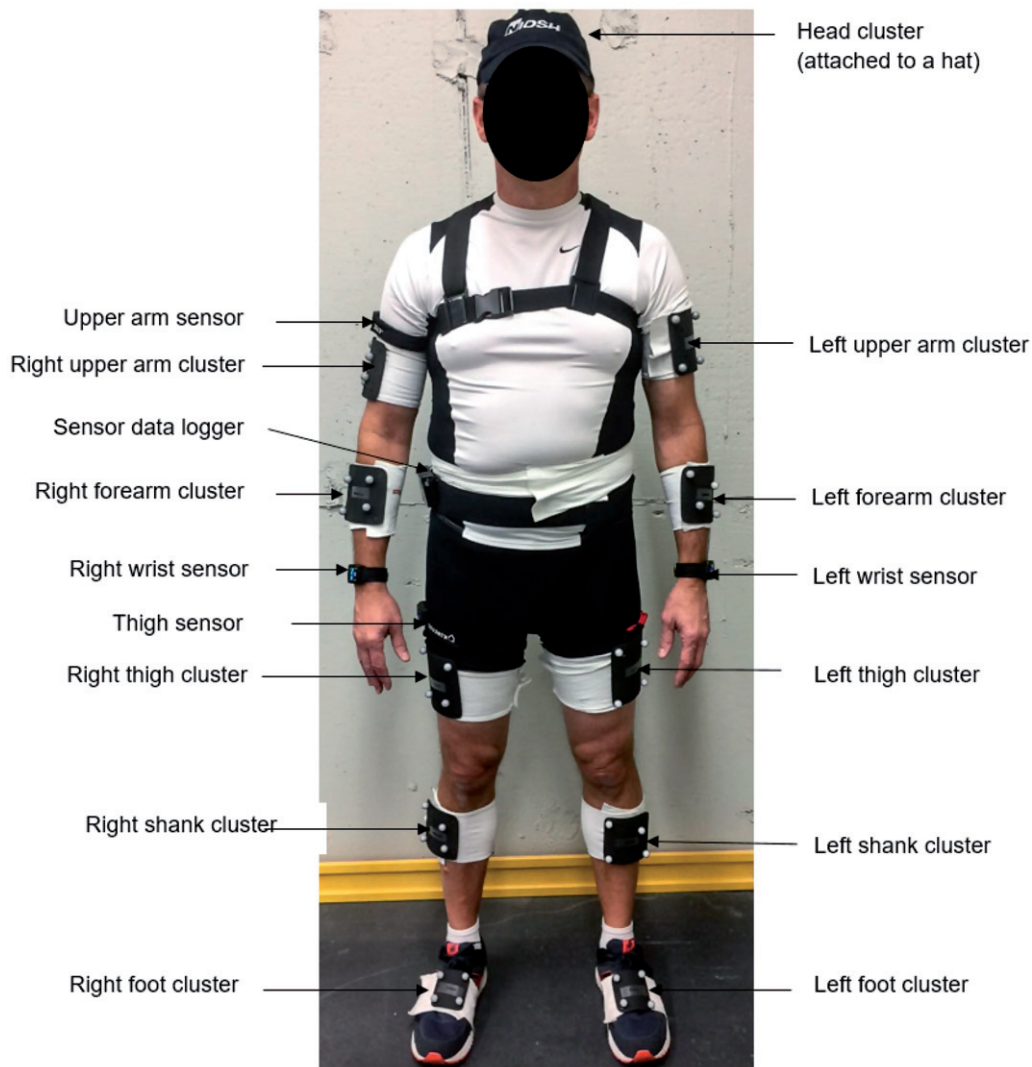
A 0.45 kg wire grid measuring 36 cm  $\times$  12 cm, with two cut-out handles, was used to simulate lifting a tote box during the trials. The grid was designed to help participants create realistic lifting motions while minimising obstructions for motion tracking. The grid was set on a small platform (12 cm  $\times$  12 cm) for setting the initial lifting height. The platform was connected to a movable clamp on a metal pole for adjusting the lifting height. The small platform provided clearance for participants to lift the grid in a

natural motion. Participants aligned their toes against one of three marked lines in the lifting area to create three horizontal distances for three grouped risk zones (Group 1: Risk Zones 1, 4, 7 and 10; Group 2: Risk Zones 2, 5, 8 and 11; Group 3: Risk Zones 3, 6, 9 and 12). These lines were 25.4, 45.7 and 71.1 cm from the centre of the grid (i.e. the centre of two hands) to the centre of their ankles.

Participants were asked to walk from a marked line (i.e. initial position) to the lifting area and line up their toes against one of the marked lines for the lifting task. The vertical height of the grid for the lifting task was set up prior to each trial. The distance between the initial position and the starting point of the lifting task typically required participants to take 3–4 steps prior to lifting the grid. After lifting the grid, participants were asked to turn around and continue to carry the grid and set it down on a shelf in a fixed height of 77.5 cm. After setting down the grid, they were asked to turn around and walk to a marked finish line for completing each trial. The distance between the shelf and the finish line typically required participants to take three to four steps to complete the trial. Participants were instructed to walk and lift/carry the grid with two hands at their own pace and in their preference of direction for the two turnarounds. They were also instructed to carry the grid in front of their body to minimise trunk asymmetry. A trial was completed continuously in about 15 s. Participants practiced a few times until they familiarised themselves with the entire experimental procedure.

The video data were recorded by a web camera (Microsoft 1080p LifeCam) in synchronisation with whole body motion capture system (OptiTrack 12 IR camera system, model Flex 13 with the MotionMonitor data acquisition programme, Innovative Sports, Inc., Chicago, USA). The web camera was set up in a fixed location at the typical eye level (165 cm from the ground) and  $\sim$ 3.92 m in distance from the beginning of each trial. The video camera viewing angle was 88° measured between lines drawn from the camera to the participants' initial standing point and drawn perpendicular to the participant's path and was near perpendicular to the direction of the lift. The camera viewing angle was offset by 31.7° from perpendicular to the sagittal plane at the lifting location. The resolution of the video recordings was set at 640 pixels  $\times$  480 pixels at a 30 fps rate to meet the hardware synchronisation requirements for the motion capture system.

Motion tracking marker clusters were attached to 13 body segments for tracking whole-body motion by the MotionMonitor programme using 12 IR cameras.



**Figure 4.** Demonstration of wearable sensor and marker cluster attachments to the body landmarks for the motion capture system. Each cluster has four small retro reflective Styrofoam spheres geometrically configured to be different from one another for motion measurements. Two clusters (upper and lower back) and one wearable sensor attached to the back of the chest Velcro assist harness are not visible in this picture. White elastic Velcro straps are used for the clusters, and thinner black elastic Velcro straps are used for the upper arm and thigh wearable sensors. Two wrist sensors are attached by adjustable rubber bands.

The body locations for marker clusters and five inertial moment unit (IMU) wearable sensors (Kinetic Inc.) are shown in Figure 4. The IMU sensors were from a previous study and not used. The motion capture system was calibrated according to the standard operating procedure of the OptiTrack company to achieve an average 0.7 mm accuracy across 10 test sessions for motion measurements in three-dimensional space.

Data acquired and derived from the motion capture system were used as ground truth for evaluating the accuracy of the video lifting monitoring algorithm. Four variables were used for validation: the time instance at the beginning of the lift (BOL), the time instance when the plate was set down at the end of the lift (EOL)

as well as  $H$  and  $V$  for the BOL. Two NIOSH researchers reviewed video recordings of the trials in the MotionMonitor programme and manually recorded the video frame numbers to establish the BOL and EOL. The criterion for determining the frame numbers was based on the moment when the grid started to move for the BOL with two hands. The two NIOSH researchers independently reviewed half of the trials, but when in doubt, they discussed questionable frame numbers to reach agreement on the final value. Once the video frames for the BOL were determined, the data were used to identify the  $H$  and  $V$  variables calculated with the motion capture data. A similar procedure was used to determine the EOL when the grid was initially set down.

**Table 1.** Demographics and anthropometric properties (Mean  $\pm$  SD) of study participants (Male:  $N=3$ ; Female:  $N=3$ ).

| Gender  | Mass (kg)        | Height (cm)     | Age (years)   | Forearm length (cm) | Upper arm length (cm) | Thigh length (cm) |
|---------|------------------|-----------------|---------------|---------------------|-----------------------|-------------------|
| M       | 101.8 $\pm$ 14.7 | 176.3 $\pm$ 1.4 | 55 $\pm$ 1.9  | 27.7 $\pm$ 1.4      | 32.3 $\pm$ 1.5        | 44.6 $\pm$ 3.5    |
| F       | 69.7 $\pm$ 7.4   | 163.6 $\pm$ 4.6 | 48 $\pm$ 13.6 | 26.4 $\pm$ 1.7      | 30.7 $\pm$ 1.7        | 44.3 $\pm$ 4.5    |
| Average | 82.5             | 169.3           | 51.3          | 27.1                | 31.5                  | 43.3              |
| SD      | 18.9             | 9.0             | 11.5          | 1.9                 | 2.0                   | 3.9               |

### 2.2.2. Participants

Six participants were recruited among employees in the division of the Applied Research and Technology office of NIOSH in Cincinnati, Ohio. Prior to data collection, written consent was obtained from the participants using the NIOSH-approved IRB study protocol. The participant inclusion criteria were individuals that were capable of (1) lifting a 1.4 kg mass in a location combining different horizontal distances from their body to the load within their reach and vertical distances from the shin to shoulder height, (2) lifting and carrying the 1.4 kg mass 3 m, and (3) repeating 12 different tasks (described previously) 3 times for a total of 36 lifting trials. Exclusion criteria were individuals with musculoskeletal disorders or any pain at the time of recruitment or in the past 3 months, individuals under age 18 years, and being pregnant. Demographic and measured anthropometric data relevant to the study are provided in Table 1.

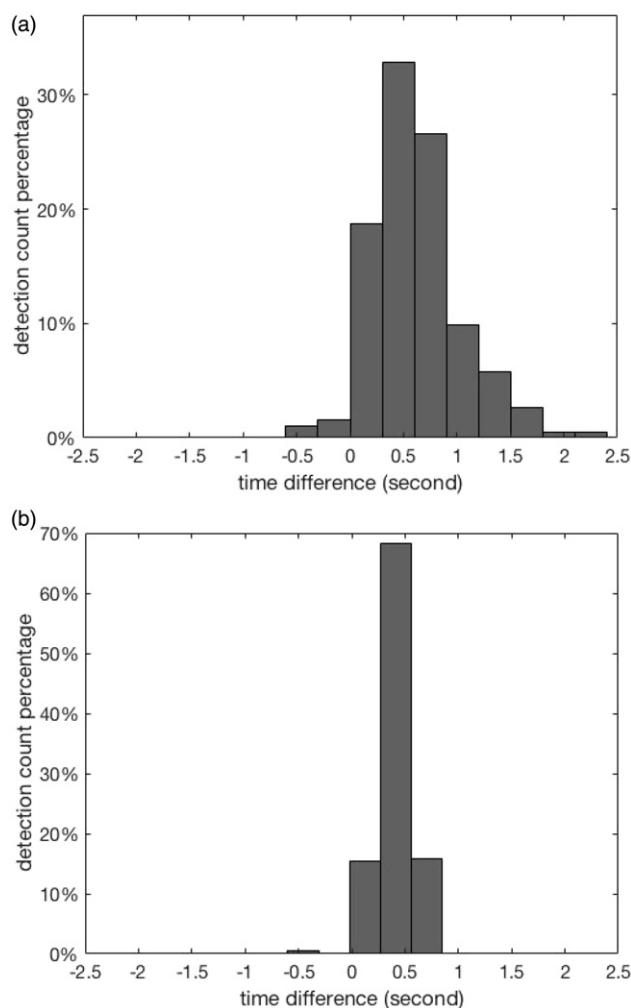
The lifting monitoring algorithm was applied using the validation dataset (6 participants with 36 clips each). Given a video clip as input, the algorithm automatically outputs the bounding box of the human figure for each frame, the detected loading and unloading instance, hands location and feet location at the loading and unloading instance, and the RWL.

The NIOSH study was designed to focus on the origin of the lift in the 12 lift zones. The RWL for the loading instance at the BOL was calculated, including the horizontal distance from the hands centre to the ankles centre ( $H$ ), and the vertical distance from the hands midpoint to the ground ( $V$ ).

## 3. Results

### 3.1. Lift and release instance detection

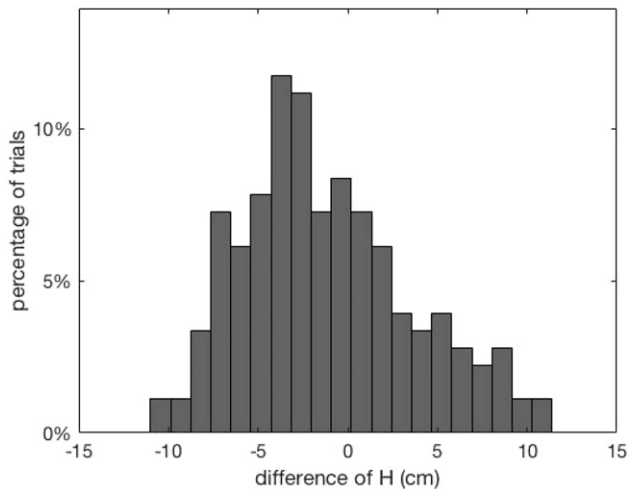
Since the ankles during loading remained steady for a short period, when the predicted lifting instance and the ground truth instance coincided with this period, the predicted lifting instance was considered successful. In this validation experiment, the lifting instance estimation was successful for all clips. The predicted lifting instance timestamp was compared with the manually



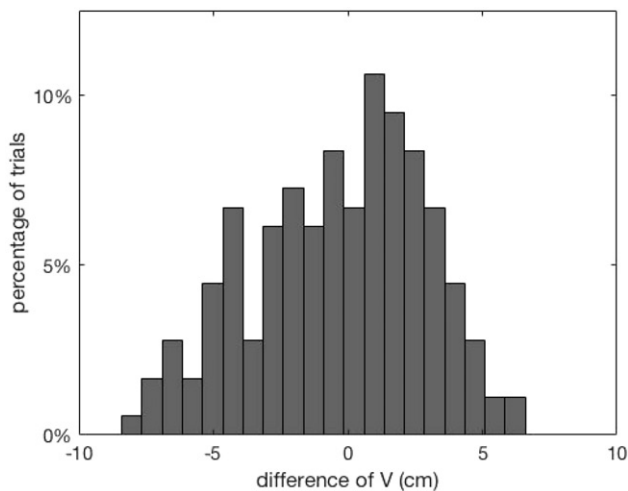
**Figure 5.** Histogram for the time difference between ground truth and (a) the beginning of lift instance detection time, and (b) the end of lift detection time.

observed instance (ground truth). The time difference between prediction and ground truth was calculated.

A histogram of the time difference for the BOL is plotted in Figure 5(a). The mean time difference was 0.624 s ( $SD=0.42$  s). Based on this distribution, 99.5% of the measured lifting instance detections were within  $[-0.636, 1.884]$  s of the ground truth measure. A similar time difference for the EOL histogram for the time difference is plotted in Figure 5(b). The mean time difference was 0.144 s ( $SD=1.52$  s). According to this distribution, 99.5% of the measured unloading instance detections were within  $[0.125, 0.833]$  s of the ground truth measure.



**Figure 6.** Histogram of the difference of  $H$  between motion capture and video detection at the loading instance.



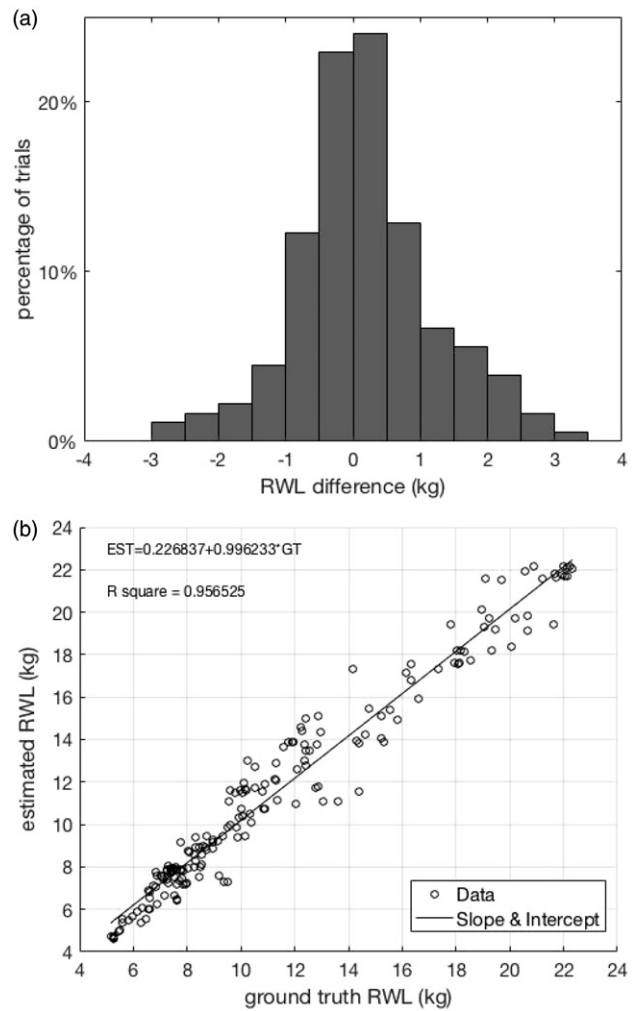
**Figure 7.** Histogram of the difference of  $V$  at the loading instance.

### 3.2. $H$ detection at the loading instance

The horizontal distance from the hands to the ankles ( $H$ ) is a key factor in calculating the RWL. We reference the  $H$  data from the 3D motion capture system as ground truth and calculate the  $H$  difference between the video estimated  $H$  and ground truth  $H$ . A histogram for the  $H$  difference is plotted in Figure 6. The mean of the  $H$  difference was  $-1.16$  cm ( $SD=4.72$  cm). Based on this distribution, 68% of the measured lifting instance detections were within  $\pm 5$  cm of the ground truth measure.

### 3.3. $V$ Detection at the loading instance

The vertical distance from the hands to the floor ( $V$ ) is another key factor in calculating the RWL. We reference the  $V$  data from the 3D motion capture system



**Figure 8.** (a) Histogram of the difference of RWL at loading instance. (b) Linear regression for ground truth motion capture and video estimated RWL.

as ground truth and calculate the  $V$  difference between estimated and ground truth  $V$ . A histogram of the  $V$  difference is plotted in Figure 7. The mean of the  $V$  difference was  $-0.36$  cm ( $SD=3.22$  cm). Based on this distribution, 88% of the measured lifting instance detections were within  $\pm 5$  cm of the ground truth measure.

### 3.4. RWL prediction at the loading instance

Comparing the multipliers between the computer vision estimations and motion tracking ground truth, the 95% confidence interval of the  $VM$  difference was  $0.005 \pm 0.016$ , the  $DM$  difference was  $0.002 \pm 0.011$ , and the  $HM$  difference was  $0.012 \pm 0.111$ . In calculating the RWL, the  $LC$  was 23 kg.  $AM$  was set equal to one because the participants were asked to perform the lifting trials while keeping their torso upright without twisting.  $CM$  was also set equal to one because the

effectiveness of hand-to-object coupling was considered good for all lifting trials. Since each lifting task was performed only once and was not repetitive, we calculated the Frequency Independent Recommended Weight Limit (FIRWL) and set  $FM=1$ . The variables  $V$  and  $H$  were the values at the loading instance.  $D$  was the vertical distance from the origin of the object to the instance before the object is loaded at the destination.  $VM$ ,  $HM$  and  $DM$  were calculated according to the multiplier transformation function required in the RNLE, respectively.

The RWL difference between estimation and ground truth was calculated and shown in Figure 8(a) where the histogram is plotted. The mean RWL difference was 0.18 kg ( $SD=1.03$  kg). Based on this distribution, 91% of the measured lifting instance detections were within  $\pm 2$  kg of the ground truth measure. The linear regression between the video estimated RWL and ground truth RWL ( $R^2 = 0.96$ ) is shown in Figure 8(b). A Bland-Altman analysis for the limits of agreement (mean difference  $\pm 1.96$   $SD$  of differences) for the estimation and the motion capture data were between  $-1.84$  kg and 2.21 kg. This demonstrates that the estimation sufficiently agreed with the ground truth measure.

## 4. Discussion

### 4.1. Accuracy of distances

The calculation of RWL in the validation experiment relies on the variables  $VM$ ,  $HM$ , and  $DM$ , which are functions of  $V$ ,  $H$ , and  $D$ , respectively. The greatest bias of RWL came from the  $HM$  because its error range was on the order of  $10^{-1}$ , while the error range of  $VM$  and  $DM$  was on the order of  $10^{-2}$ . The relatively larger error in  $H$  can be explained by the camera viewing angle in this study.

The algorithm is built on the assumption that the camera viewing direction is perpendicular to the sagittal plane and thus the hands and feet should overlap because of symmetry. Therefore, the detection of the lifted object can be ascertained from either hand. However, in this validation experiment, the camera viewing angle was offset by  $31.7^\circ$  from the perpendicular at the lift location, where ideally there would be no offset. Consequently, the geometric centre of the un-overlapped extremities would likely bias the location measured. Thus, this validation data were for the worst-case scenario. Since the data were from a previous study, it was not possible to change the video camera angles.

The computer vision algorithm measured distance in units of pixels, while the motion capture ground

truth measurements were provided in centimetres. The conversion between these two was therefore dependent on the video camera set-up. The angle between the sagittal plane at the lifting point and the camera viewing plane was  $6.1^\circ$ . After converting the projection of the  $H$  distance from the video camera viewing plane back to the sagittal plane, the distance estimation assumed that all the dimensions of the participant share the same depth from the camera. Without knowing the actual focal length of the camera, we used the participant's standing height as a reference for calibration. However, in actual video recordings, the calibration is distorted at the video frame edge and the variant depths are not consistent with uniform calibration.

As detection of centre of the hands and feet directly comes from the centre of the corresponding blobs, the precision of the algorithm relies on the precision of the blob. If the blob is incomplete because of dilution or enlarged by noise, the centre of the blob will be biased, and thus the location estimation will be biased. If the resolution of the video was greater so that the calibration ratio was less, then the error caused by non-uniform shaped blobs would be reduced.

### 4.2. False negative detections

If the algorithm cannot identify a ghost blob, or if the location of the detected ghost blob does not comply with human body structure, the algorithm would fail. Using this criterion, there were 17.13% false negative detections in this dataset.

In order to be detected by the algorithm, the ghost blob should be complete and not diluted for a consecutive sequence of  $N$  frames. To avoid noisy blobs, an arbitrary  $N$  was set as 20 frames. If the isolated ghost blob lasted less than  $N$  frames after detaching from the participant's silhouette, the algorithm would not identify the ghost blob. A long-duration overlap of the ghost blob and the participant's silhouette mostly depends on the participant's moving speed. In this validation dataset of six participants, one participant moved significantly slower than the others. This made the isolated blob persist for a shorter time resulting in 10 false negative detections for this participant, while for other participants, there were no more than three false negative detections. Because the algorithm depends on motion, a very slow-motion lift could potentially impair accuracy. This effect will be studied in future investigations.

A weak appearance of an object can lead to a weaker blob. In this validation experiment, the lifted object was a thin grid, and thus its ghost blob was often weak and seen as a shadow or was diluted. Additionally, this study used a relatively low definition video (640 pixels  $\times$  480 pixels) due to the limitation of the 3D motion capture system. Consequently, the plate appeared as a line segment when viewing in the plane of the plate surface. The background subtraction for this plate was represented as a 2 pixels  $\times$  20 pixels blob. This caused the ghost blob to rapidly diminish and was considered as noise with high probability. This coarse representation undoubtedly provided less information for the algorithm, while a higher definition video should significantly improve the resolution of the blobs and thus the algorithm accuracy. A solid object would certainly form a much stronger blob.

The current study was conducted in the laboratory under controlled conditions where the background was stationary and there were no extraneous motions. In an industrial setting, it is anticipated that the background might contain other workers, vehicles, or machinery that can introduce multiple ghosting images from their movements. Proximity was used to identify the ghost that was caused by the participant. This methodology will be refined in future work in order to reject noise introduced by extraneous motions.

Since the algorithm relies on motion information to detect a moving target if the appearance (colour and tone) of the foreground and the background were close, the motion would be hard to differentiate, and the algorithm would be challenged, introducing error. However, the algorithm does not detect targets according to colour features, so the markers worn by participants for the motion capture system did not have any noticeable effect on the outcome.

Compared with previous research aimed at estimating body position using 2D or 3D videos (Bogo et al. 2016; Howe, Leventon, and Freeman 2000; Mehrizi et al. 2017, 2018), the current algorithm directly detects the hand and feet locations while avoiding the previous intermediate steps of recognising specific body segments or fitting skeletal models. This approach not only reduces computational complexity but is not sensitive to possible errors caused by miscalculation of any single body segment location, which can offset the entire measurement. Future research will study how this feature will help prevent errors potentially introduced in the industrial setting where obstacles, poor illumination or occlusion may interfere with obtaining accurate measurements.

Spector et al. (2014) collected a large dataset comprising six participants performing lifting tasks in a laboratory setting. The RNLE parameters were provided using a Kinect, while the ground truth was provided by a 3D optical motion capture system. A manual inspection was performed to exclude the obvious outliers before data processing. However, in the current study, no outliers were excluded, and all data were used regardless of lifting style or anomalies encountered. Comparing with the modelled data, our  $H$  estimation error was more concentrated around 0 with its 25<sup>th</sup> to 75<sup>th</sup> percentiles ranging within  $[-0.04, 0.02]$  m and median being  $-0.02$  m. For the modelled data in Spector et al. (2014), the 25<sup>th</sup> to 75<sup>th</sup> percentiles ranged within  $[0.06, 0.13]$  m, and the median was 0.09 m. The  $V$  estimation error in the current study had 25<sup>th</sup> to 75<sup>th</sup> percentiles ranging within  $[-0.03, 0.02]$  m and a median of 0 m, while the 25<sup>th</sup> to 75<sup>th</sup> percentiles in Spector et al. (2014), ranged within  $[-0.02, 0.04]$  m and median was  $-0.02$  m. This shows that the  $V$  estimation errors were similar for both studies. The cited values from Spector are the average of approximated values observed from the boxplot figures published in Spector et al. (2014). The current study estimation of  $H$  and  $V$  had no outliers removed, compared with the many outliers in Spector et al. (2014).

The accuracy of location prediction for the current algorithm is dependent on the detection and accuracy of blobs. Shadows, shading and highlights caused by illumination, blur the boundary between the foreground and the background, and are important factors affecting the performance of segmentation accuracy. Chondagar et al. (2015) acknowledge that although multiple solutions for these problems have been proposed, they have not yet been solved. As discussed previously, there were 17.13% false negative detections in this dataset (i.e. 82.9% correct detections). The state-of-art semantic segmentation approach by Long, Shelhamer, and Darrell (2015), which uses a fully convolutional network, presents pixel segmentation accuracy only as high as 90.3%. Consequently, good illumination is necessary for accuracy using the current approach. The current approach combines the ghost appearance feature and motion information to improve the accuracy of blob detection. Considering its computation complexity, the current approach has achieved good performance in blob segmentation for a laboratory setting.

### 4.3. Recommendations for improvements

An effective improvement in error could be achieved by providing a quality assurance measurement. Some

factors influencing the algorithm accuracy include, but are not limited to, how close the participant's hands and shoes are similar to the background in colour as well as whether there was proper illumination to make the significant objects clear without shadows.

A higher definition video could significantly help to detect the ghost blobs. For example, the thin plate which was represented by a 2 pixels  $\times$  20 pixels blob in the current 640 pixels  $\times$  480 pixel resolution video, would appear as a 9 pixels  $\times$  120 pixels blob in a 4K (3840 pixels  $\times$  2160 pixels) resolution video. Thus, it is less probable it would be eliminated as a noise blob. The videos in the current study were limited to 640 pixels  $\times$  480 pixels resolution in order to synchronise the video with the 3D motion capture data, but we anticipate that a more dense pixel resolution will provide more distinct blobs with greater contrast and details. Also, an object with more visual density than the current transparent thin plate would certainly improve detection. Furthermore, increased pixel resolution would provide greater precision in the distance measurements.

This validation dataset was conducted under a worst-case viewing angle (31.7°) and would certainly achieve a better distance estimation if the camera were set perpendicular to the sagittal plane. Future studies will examine the amount of error that was contributed by a camera viewing angle.

The camera technology of today's devices has better resolution, the algorithms offer suitable computational simplicity, and the method has suitable stability for a hand-held camera. Additionally, an anti-shake algorithm will be explored. In the natural work setting, the most challenging problem for this algorithm is multiple moving objects. As the final design of the video lifting monitor is implemented in a hand-held device, most of the noisy moving objects could be avoided in the image by manually choosing the recording angle.

This investigation was a laboratory study, involving controlled conditions, limited to symmetrical lifting and included relatively few participants (six). Future field studies will further test the algorithm performance, including actual lifting conditions and a larger number of participants to truly validate the method; the current investigation, however, was important because it enabled comparison of conventional video against ground truth motion capture data, something very difficult to control in the field. Future demonstration studies will be conducted in industrial facilities involving actual manual materials handling tasks in production, warehousing and supply chain operations.

It is acknowledged that generalizability of this study is limited, given the number of participants, controlled lifting parameters, and restrictions imposed by laboratory conditions, however, the results demonstrate that this novel approach has promise for future field applications.

## 5. Conclusion

In this paper, a 2D-video based lifting analysis was presented. The system provides a robust, non-intrusive method to automatically extract the spatial and temporal factors necessary for applying the RNLE using a single video camera in the view from the sagittal plane. Compared with 3D motion capture, our approach provides sufficiently high accuracy of the RWL predictions. It also provided automatic detection of lifting instances. Efficiency in computation load and non-intrusive properties will make it possible to be implemented on a hand-held device and accessible for a wide variety of applications.

## Acknowledgements

The findings and conclusions in this paper are those of the author(s) and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centres for Disease Control and Prevention.

## Funding

Funding for this publication was made possible (in part) by grants from the National Institute for Occupational Safety and Health (NIOSH/CDC), [R01OH011024] (Radwin).

## ORCID

Robert G. Radwin  <http://orcid.org/0000-0002-7973-0641>

## References

- ACGIH. 2019. "ACGIH TLVs and BEIs: Threshold limit values for chemical substances and physical agents biological exposure indices." Cincinnati, OH: American Conference of Governmental Industrial Hygienists.
- Bogo, F., A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. 2016. "Keep it SMPL: Automatic Estimation of 3d Human Pose and Shape from a Single Image." In *Proceedings of European Conference on Computer Vision*, 561–578. Cham, Switzerland: Springer.
- Bureau of Labor Statistics. 2016. "Non fatal occupational injuries and illnesses requiring days away from work, 2015." *Bureau of Labor Statistics*, November 10. <https://www.bls.gov/news.release/osh2.nr0.htm>
- Chondagar, V., H. Pandya, M. Panchal, R. Patel, D. Sevak, and K. Jani. 2015. "A review: Shadow detection and removal."

- International Journal of Computer Science and Information Technologies* 6 (6): 5536–5541.
- Delpresto, J., C. Duan, L. M. Layiktez, E. G. Moju-Igbene, M. B. Wood, and P. A. Beling. 2013. "Safe Lifting: An Adaptive Training System for Factory Workers Using the Microsoft Kinect." In *IEEE Systems and Information Engineering Design Symposium, SIEDS*, 64–69. Piscataway, New Jersey, United States: IEEE.
- Dempsey, P. G. 2002. "Usability of the Revised NIOSH Lifting Equation." *Ergonomics* 45(12): 817–828. doi:10.1080/00140130210159977.
- Gabel, M., R. Gilad-Bachrach, E. Renshaw, and A. Schuster. 2012. "Full Body Gait Analysis with Kinect." In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 1964–1967*. Piscataway, New Jersey, United States: IEEE.
- Howe, N. R., M. E. Leventon, and W. T. Freeman. 2000. "Bayesian Reconstruction of 3d Human Motion from Single-Camera Video." *Advances in Neural Information Processing Systems* 12: 820–826.
- Juul-Kristensen, B., G. Å. Hansson, N. Fallentin, J. H. Andersen, and C. Ekdahl. 2001. "Assessment of Work Postures and Movements Using a Video-Based Observation Method and Direct Technical Measurements." *Applied Ergonomics* 32(5): 517–524. doi:10.1016/S0003-6870(01)00017-5.
- Kim, S., and M. A. Nussbaum. 2013. "Performance Evaluation of a Wearable Inertial Motion Capture System for Capturing Physical Exposures during Manual Material Handling Tasks." *Ergonomics* 56(2): 314–326. doi:10.1080/00140139.2012.742932.
- Li, G., and P. Buckle. 1999. "Current Techniques for Assessing Physical Exposure to Work-Related Musculoskeletal Risks, with Emphasis on Posture-Based Methods." *Ergonomics* 42 (5): 674–695. doi:10.1080/001401399185388.
- Liberty Mutual Research Institute for Safety. 2017. Liberty Mutual Workplace Safety Index. Boston, MA: Liberty Mutual.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. Boston, MA: IEEE.
- Lowe, B., P. Dempsey, E. Jones, and National Institute for Occupational Safety and Health (NIOSH). 2018. "Assessment Methods Used by Certified Ergonomics Professionals." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 838–842. Los Angeles, CA: SAGE Publications. doi:10.1177/1541931218621191.
- Patrizi, A., E. Pennestrì, and P. P. Valentini. 2016. "Comparison between Low-Cost Marker-Less and High-End Marker-Based Motion Capture Systems for the Computer-Aided Assessment of Working Ergonomics." *Ergonomics* 59(1): 155–162. doi:10.1080/00140139.2015.1057238.
- Plantard, P., H. P. Shum, A. S. Le Pierres, and F. Multon. 2017. "Validation of an Ergonomic Assessment Method Using Kinect Data in Real Workplace Conditions." *Applied Ergonomics* 65: 562–569. doi:10.1016/j.apergo.2016.10.015.
- Mehrizi, R., X. Xu, S. Zhang, V. Pavlovic, D. Metaxas, and K. Li. 2017. "Using a Marker-Less Method for Estimating L5/S1 Moments during Symmetrical Lifting." *Applied Ergonomics* 65: 541–550. doi:10.1016/j.apergo.2017.01.007.
- Mehrizi, R., X. Peng, X. Xu, S. Zhang, D. Metaxas, and K. Li. 2018. "A Computer Vision Based Method for 3D Posture Estimation of Symmetrical Lifting." *Journal of Biomechanics* 69: 40–46. doi:10.1016/j.jbiomech.2018.01.012.
- Spector, J. T., M. Lieblich, S. Bao, K. McQuade, and M. Hughes. 2014. "Automation of Workplace Lifting Hazard Assessment for Musculoskeletal Injury Prevention." *Annals of Occupational and Environmental Medicine* 26(1): 15.
- Waters, T. R., V. Putz-Anderson, A. Garg, and L. J. Fine. 1993. "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks." *Ergonomics* 36(7): 749–776. doi:10.1080/00140139308967940.
- Zivkovic, Z. 2004. "Improved adaptive Gaussian Mixture Model for Background Subtraction." In *Proceedings of the 17th International Conference on Pattern Recognition, 2004 ICPR 2004*, 28–31. Cambridge, UK: IEEE.
- Zivkovic, Z., and F. Van Der Heijden. 2006. "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction." *Pattern Recognition Letters* 27 (7): 773–780. doi:10.1016/j.patrec.2005.11.005.