Video-based Ergonomic Repetitive Hand Motion Analysis

By

Cheng-Hsien Lee

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Electrical Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 8/12/2021

The dissertation is approved by the following members of the Final Oral Committee:

Yu Hen Hu, Professor, Electrical and Computer Engineering
Robert Radwin, Professor, Industrial and Systems Engineering
John A. Gubner, Professor, Electrical and Computer Engineering
Bill Sethares, Professor, Electrical and Computer Engineering

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of numerous individuals.

I would like to express my deepest gratitude to my advisors, Professor Yu Hen Hu and Professor Robert Radwin, for their endless support, inspiration, and guidance throughout my Ph.D. study.

Thanks to my thesis committee members, Professor John A. Gubner and Professor Bill Sethares, for their suggestions and teachings.

Special thanks to the members of the Occupational Ergonomics and Biomechanics Lab. Thanks to Oguz Akkas for cooperating with me in the duty cycle projects. Thanks to Runyu L. Greene and David Azari for always answering my questions patiently.

Great thanks to my labmate, Chia-Hsiung Chen for giving me a lot of advice and directions. I would also like to thank Xuan Wang, Jian Wei Ke, and ZhengYang Lou for providing valuable help.

Last but not least, I would like to express my sincere gratitude to my parents and sister for their support, encouragement, and love.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Hand activity analysis for industrial works that require intensive repetitive hand motions provides important information to facilitate ergonomic analysis of the hand activity levels (HAL). The outcome may lead to a sound work-rest schedule with significant health benefits. The hand activity analysis consists of two important parts: (a) tracking the hand motion trajectory, and (b) determining whether the hand is exerting force or resting. In this research, we develop novel techniques to determine the state of repetitive hand activities (exerting forces or resting) based on kinematic information such as hand motion trajectory.

In a repetitive hand task, the percentage of a cycle that the hand is holding a tool, or a product is called a duty cycle. Force is exerted by the hand during the duty cycle. Accurate estimation of the duty cycle provides the required information for exposure assessment in evaluating repetitive hand work.

A key hypothesis of this research is that the motion trajectory and associated kinematic measurements (speed, acceleration) can provide sufficient information for duty cycle estimation. To validate this hypothesis, we developed a machine-learning algorithm using hand motion kinematic measurements extracted from videos of both simulated laboratory and practical factory repetitive hand work environments.

We further propose a computer vision method to automatically measure wrist flexion and extension from a 2D video for occupational health and safety research. This algorithm tracked skeletal joints of the elbow, wrist, and hand to estimate the wrist flexion/extension angle between the hand and forearm. Based on the estimated angles, wrist posture was classified as flexion (palmar bending), neutral (no bending), or extension (dorsal bending) for each cycle of hand movement. Applying to a set of laboratory videos of a simulated repetitive hand motion task and

selected video frames of hand flexion instances from industrial field video data, we demonstrated the feasibility of using this algorithm for assessing the state of hand activities during manual work.

In summary, we demonstrated two non-intrusive, automated observation methods allowing long-term, large-scale collection of exposure data. They facilitate repetitive hand motion analysis for monitoring occupational health and safety and can help prevent job-related hand injuries.

Chapter 1.

Introduction

Human action analysis has been a focused area of interest in computer vision, human-computer interaction, and related fields [1]–[7]. Among many human action monitoring tasks, vision-based hand activity analysis has received considerable attention recently [8]–[27]. A majority of these works are aimed at hand gesture recognition. Some recent ones [28]–[30], however, focus on hand motion kinematics in the context of ergonomics of industrial jobs requiring intensive upper limb activities [31]–[35].

Human activity can be categorized into four levels: gestures, actions, interactions, and group activities [36]. Gestures are the fundamental movement of the human body part. Actions consist of different gestures temporally. Interactions mainly mean the activity involves two persons or the person and objects. Group activities are the event containing multiple persons generally. There are plenty of researches related to gestures and actions [8], [15], [18], [20], [25], [37]–[41]. However, few studies are about the interaction between humans and objects. Human-object interactions can reveal important contextual information about human action. For instance, if the system can detect a person carrying a suspicious box, then a potential bombing incident may be averted. In the context of ergonomic analysis of human body motion in an industrial work environment, the interaction between the human body, in particular the hands, and objects (tools, materials to be processed) provides useful information such as whether force exertion is applied. Such information will facilitate accurate analysis of hand activity levels [28], [31]–[35], [42].

Duty cycle (DC) is one of the primary measures used in ergonomics for evaluating repetitive exertions, muscle fatigue, and manual materials handling tasks. It is defined as the proportion of

time spent in actual task-related activities and is typically calculated as the exertion time divided by the total time spent doing the task, including rest periods. DC has become an important metric for quantifying work activity and manual exertions in the workplace for optimizing work-rest periods for preventing fatigue and injuries.

DC has played a major role in predicting fatigue and its prevention [43]–[45]. Rohmert [44] described the phenomena of fatigue and recovery to determine work-rest cycles and their influence on workers' strain and stress. A study on continuous and intermittent isometric contractions linked load and work/rest ratio limits to indicators of localized muscle fatigue [46]. Woods et al. [47] considered the work-rest allowance effect on muscle fatigue and predicted an optimum work-rest schedule to minimize this effect. Potvin [48] more recently created an equation based on DC from a meta-analysis of numerous studies in the literature to estimate maximum acceptable force in repetitive manual tasks.

DC has also been used for quantifying exposure to physical stress in repetitive manual work. The strain index, which is used for evaluating repetitive manual tasks, included DC as one of its parameters [49]. A prospective study of biomechanical risk factors for carpal tunnel syndrome found DC for forceful hand exertions was a significant risk factor [50]. Latko et al. [51] introduced a method for quantifying repetitive hand motion, the hand activity level (HAL), which was also related to DC.

HAL applies to mono-cycle tasks and is an observational visual-analog metric that an observer subjectively assesses on a scale from 0 to 10, anchored between the "hand idle most of the time and no regular exertions" and "rapid, steady motions/exertion; difficulty keeping up." The HAL scale is part of the American Conference for Government Industrial Hygienists (ACGIH) threshold limit value (TLV™) for evaluating the risk of work-related distal upper extremity musculoskeletal

disorders [31]. The HAL rating can also be evaluated by directly measuring the exertion frequency and DC against a look-up table. Radwin et al. [52] developed an equation for estimating HAL from these parameters as an alternative to the TLV™ table.

Common ways for measuring DC include time and motion studies, manual video coding, observational methods, and self-reports. Among these, times studies and video coding are the most accurate, whereas observations and self-reports lack consistency and reliability [53]. Bao et al. [54] observed disagreement between observer-rated frequency-DC estimates and detailed time study analyzes. In another study, Garg and Kapellusch [55] discuss the lack of consistency between the methods of evaluating HAL, from observer-rated assessment and table look-up values, and address the need for a consistent method of evaluation. Kapellusch et al. [56] stated a similar need for a robust technique. Moreover, Wells et. al. [57] explains that estimating exposures related to time is difficult, in which self-reports are inconsistent while direct measurements are time and resource-consuming. An objective, automated method for evaluating DC could help resolve these issues.

Yen and Radwin [58] looked at using signal pattern recognition for automatically quantifying cyclical tasks from wrist electrogoniometer signals and concluded that such an approach may be useful, using interactive fine-tuning. Recent advances in computer vision enable HAL to be measured non-invasively without instruments, using automated video processing that employs semi-automatic marker-less tracking of a region of interest (ROI) located near the hand to measure frequency and duty cycle [28]. Such an approach is automatic, repeatable, objective, unobtrusive, and is suitable for a real-time, direct reading assessment.

We previously demonstrated that hand root-mean-square (RMS) speed while exerting force, and DC measures were well suited for automatically estimating HAL and had good agreement

with independent observational ratings of videos for actual industry jobs [59] using DC obtained manually using Multimedia Video Task Analysis™ (MVTA™) frame-by-frame analysis [60]. We hypothesize that a completely automated approach for measuring DC could be achieved using the tracked kinematic record. The current study advances an automatic method to measure DC using marker-less video tracking.

Chen et al. [28] previously calculated DC for a repetitive load transfer task performed in the laboratory. In this method, the local minima of absolute velocity values were first identified. If the acceleration between successive local minima points exceeded a preset threshold, it was determined that the hand was loaded during the period between the pair of local minima points. Such a definition of hand loading was based on observations made for the specific load-transfer task (moving a lead-filled bottle from a tray to a rotating turntable). The DC values obtained using that approach were 1.27 times greater than those measured manually using MVTA ($R^2 = 0.63$).

In this research, we advance the automatic measurement of DC by studying a feature vector training (FVT) algorithm that utilizes kinematic properties (i.e. location, velocity, and acceleration) of a video marker-less tracked ROI to estimate DC. This method is applied to two sets of data. The first is from a simulated repetitive hand-intensive task performed in the laboratory. The second set of tasks are selected from videos taken at real-world industrial work sites [61]. Ground truth time measurements of DC were ascertained utilizing manual frame-by-frame MVTA analysis and compared. The detailed methodology and results will be presented in Chapter 3.

Work-related upper extremity injuries are prevalent in jobs involving repetitive and forceful exertions with awkward postures [69]. A significant portion of those claims are attributed to the hand and wrist [70]. The ability to perform a quick and accurate assessment of workplace injury risk related to joint postures is important for research and injury prevention programs [71].

The joint angles are a critical factor for assessing work-related musculoskeletal risk [72]. Studies on upper extremity distal wrist disorders have garnered great importance [73], [74], [50]. Strenuous and rapid wrist-bending while gripping is an important risk factor for carpal tunnel syndrome. Wrist posture is one of the six variables, including intensity of exertion, duration of exertion per cycle, efforts per minute, wrist posture, speed of exertion, and duration of task per day, for estimating the Strain Index [49]. Research focusing on methods to assess the physical exposure to work-related musculoskeletal risk have been reported, including wrist posture [75]–[78].

Postural risk assessment methods include both subjective and objective approaches. Subjective methods such as analyst observation, and workers' self-reports have limited reliability and validity due to interrater variability [80]–[82]. Objective approaches including direct measurements, video recording, and computer-aided analysis [79] may offer more accurate assessments [83]–[85] at expense of intrusive measurement methods [86]. A recent study [71] compared observer-estimated wrist angles from video recordings to those measured using electro-goniometers. The agreement of estimated wrist angles between the two methods was 57% for flexion/extension classification using side-view video taken from the sagittal plane.

Recent research has also applied the Kinect range sensor for assessing postures. The influence of sensor view angle relative to the worker on identifying the risk level of recorded postures was studied [87]. Rapid Upper Limb Assessment (RULA) in real work conditions using Kinect skeleton data was also described [88]. In [89] a novel intelligent system for the Rapid Entire Body Assessment (REBA) was applied based on convolutional pose machines. The deviations of the wrist angles were very small in the working postures they chose. Although they did compute wrist angles, the flexion and extension of the wrist were not the focus.

In the present study, a computer vision method is developed to automatically measure the wrist flexion/extension angle from a video recording and classify the wrist flexion/extension state using a 3-bin scale [90]. By automating the process of estimating the wrist flexion/extension angle, the method promises greater accuracy in accessing physical exposure of the upper extremity during work. Tested on 1,464 frames from 61 recorded videos for 16 participants, the algorithm achieved an average performance of 72.40% correct, per-class accuracy. This compares favorably against the 57% consistency rate reported in [71] where the human observer estimated wrist flexion/extension states from video recordings are compared against those measured using electro-goniometers.

The same approach can be extended to automate the measurement of other joint angles for a more accurate assessment of work-related musculoskeletal risks [72]. This automated assessment method also incurs lower costs and can be scaled up to more job types and more observations. The detailed methodology and results will be described in Chapter 4.

The organization of this thesis is as follows:

In Chapter 2, the traditional computer vision algorithms and packages for hand entitativity monitoring are discussed. The ergonomic concerns and Existing video-based approaches introduced.

In Chapter 3, we propose a machine-learning algorithm using hand motion kinematic measurements extracted from videos as the features to predict DC and HAL in video-recorded simulated repetitive motion tasks and real-world factory videos.

In Chapter 4, we propose a computer vision method to automatically measure wrist flexion and extension from a 2D video for occupational health and safety research. We realized this method

by using the technology of 2D skeletal joints estimation. The proposed method was examined on both environment-controlled laboratory videos and real-world factory videos.

Chapter 5 presents a summary of this dissertation and some possible directions for future works.

Chapter 2.

Human Activity Analysis by Videos

2.1 Computer vision algorithms and packages for hand entitativity monitoring

In the past decades, the development of technology in the computer vision field is evident [96]–[98]. The improvement of hand tracking and pose estimation algorithms make many difficult problems solvable. These algorithms might also be used to facilitate evaluating repetitive motion tasks for hand activity levels or wrist postures.

The traditional object tracking methods can be classified into two categories: target representation and filtering. Target representation algorithms use different kinds of methods to track the objects based on their colors, textures, or shapes. One kind of common target representation algorithms is kernel tracking, such as *Mean Shift*. It is an iterative procedure of locating the target based on the maximization of some similarity measure. The other kind of target representation algorithms is silhouette tracking, such as active contour. This kind of method iteratively evolves an initial contour of the target from the previous frame to the location of a new contour in the current frames. Filtering methods consider prior information and observation to deal with the object dynamics under different hypotheses. A famous example is the Kalman filter. It is a two-stage iterative algorithm. In the prediction stage, the filter estimates the current states based on the previous states. And then in the update stage, the filter will update the estimation by a weighted average.

Kernel tracking algorithms have the advantages of high accuracy with low computational complexity and therefore are adopted by most of the existing methods. In this study, we adopt a template tracking algorithm [63], which is also one kind of kernel tracking, to acquire the kinematic trajectories of the hand.

Posture estimation is also an awakening field in computer vision. It is widely used in many applications including human-computer interaction, activity recognition, surveillance, picture understanding, etc.

Felzenszwalb and Huttenlocher [99] applied pictorial structures to match the model of human body parts. The problem of matching a pictorial structure is minimizing an energy function as the following equation:

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right) \tag{2.1}$$

where $L = (l_1, \dots, l_n)$ is the location of body parts. $m_i(l_i)$ is a function measuring the degree of the matching when the part v_i is located at position l_i in the image, and $d_{ij}(l_i, l_j)$ is a function measuring the degree of deformation of the model when the part v_i is located at position l_i and the part v_j is located at position l_j. The method has difficulty dealing with occluded body parts and joint pose estimation of multiple nearby people [100].

Toshev and Szegedy [101] formulated the pose estimation problem as a DNN-based (Deep Neural Network based) regression problem towards body joints. In their settings, each labeled image is denoted by (x, y) where x stands for the image and $y = (\dots, y_i^T, \dots)^T, i \in \{1, \dots, k\}$ is the ground truth pose vector, where y_i contain the 2D coordinates of the ith joint. They used a convolutional neural network with seven layers, each of which is a linear transformation followed by a nonlinear one. The first layer takes an image of predefined size as input, and the last layer outputs the target value of regression, which is the predicted pose vector. To achieve better precision, they trained a cascade of pose regressors. The estimated pose vector will be refined through each subsequent stage.

In this study, we adopt Openpose [91], an open-source real-time system for multi-person 2D pose detection, including body and hand keypoints. It is a two-branch convolutional neural network that takes an image as the input and predicts a set of 2D confidence maps of body part locations and a set of 2D vector fields of part affinities simultaneously. Openpose applies a greedy inference to parse the confidence map and affinity field to output 2D coordinates of 25 key points corresponding to 25 body skeletal joints of the subject. In this study, we only use the locations of the three key points out of the 25 key points: the elbow, wrist, and hand (knuckle) of one hand.

2.2 Ergonomic concerns and Existing video-based approach

The American Conference for Government Industrial Hygienists (ACGIH) hand activity level (HAL) threshold limit value (TLV) is a measure of the risk of work-related upper extremity musculoskeletal disorder and was developed to protect workers who perform repetitive hand exertions for 4 or more hours daily in their tasks.

HAL scale was first introduced by Latko et al. [51]. They proposed a 10-point visual-analog scale for rating HAL which is depicted in Figure 2-1. The repetition rating is low when the worker's hand is idle most of the time or has no regular exertions, and high when the worker's hand is in a rapid steady motion or has difficulty keeping up.

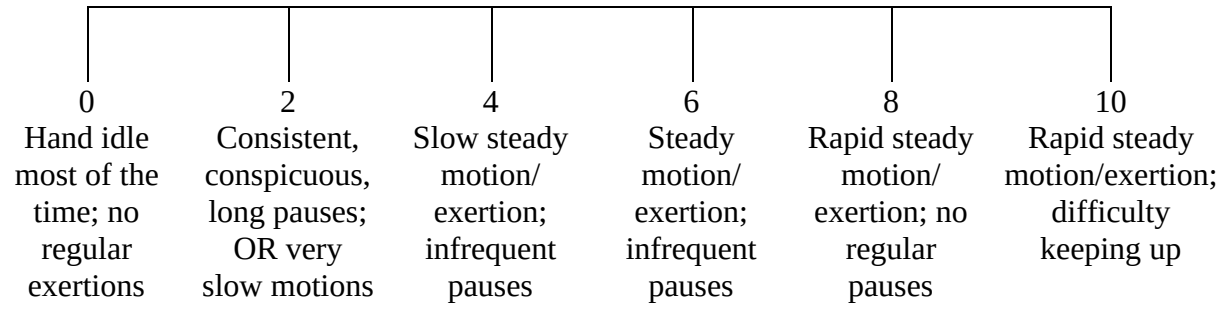


Figure 2-1 Scale for rating hand activity level [51]

The HAL is traditionally rated by trained observers, considering exertion frequency, rest pauses, and speed of motion according to specified guideline descriptions. Radwin et al. [52] proposed an equation for predicting HAL, based on exertion frequency and duty cycle. The duty cycle is defined as the exertion time divided by the total time of the task. Akkas et al. [59] further developed an equation, which depends on the root mean square of the hand speed and duty cycle, for estimating HAL.

Work-related upper extremity musculoskeletal disorder, including injuries of muscles, nerves, and tendons in shoulders, elbows, forearms, wrists, and hands, have been causing hazards to the health of the workers and enormous losses due to lost workdays in the workplace [61], [102], [103].

To prevent these occupational injuries, identifying hazardous tasks with exposure assessments becomes crucial. There are three major categories of methods used for exposure assessments: self-report, direct measurements, and observational methods. Self-report assesses the ergonomic risk using a questionnaire with pre-defined questions. It is easy to use but could have a large bias due to perceptional differences of distinct subjects. Direct measurements need to attach instruments to the hands or arms of the workers, and are therefore more accurate. However, it has the disadvantages of time consumption, invasiveness, and high cost [104]. Observational methods estimate the exposure to risk factors by ergonomic experts who watch the work process. Observational methods are also time-consuming but provide more reliable and validated results than self-report.

Many video-based technical tools have been developed to assist those assessment methods. Multimedia Video Task Analysis (MVTA) [60] is a software tool for observational time-based activity and event analysis with synchronized analog data sampling. It can reduce tedious activities in video analysis. The user interface facilitates data management, information entry, and time

analysis. A time-based posture analysis approach developed by SHARP [93] uses a special data processing program for posture estimation by analysts to obtain a continuous angular scale. The software shows two synchronized video frames and asks the analyst to estimate the wrist angle they see by clicking on the posture diagram.

Patrizi et al. [104] compared the marker-less methodology using Microsoft Kinect with BTS SMART system which requires the use of reflective markers to be placed on the subject's body, and concludes that maker-less devices could be successful for ergonomics purposes that do not require high precision.

Chapter 3.

Measuring Duty Cycle Using Hand Motion Kinematics

3.1 Experiment Setup

3.1.1 Laboratory Task Simulation

In order to develop and test the algorithms for automatically measuring DC, we simulated a prototypical repetitive motion task in the laboratory. A subject grasps a ball from one location, moves it to a specified location, releases it, and reaches for another ball. An apparatus (Figure 3-1) was fabricated using an electromechanical linear actuator for indexing the balls that are obtained and deposited for a paced sequence. The device is comprised of an 840-mm travel length linear belt drive actuator (Misumi MSS-625) driven by a bipolar stepper motor (ElectroCraft Model TPP34 with 560 N cm torque) and controlled by a stepper motor controller (IMS MX-CS101–401). The device was capable of moving a 2-kg object every 0.5 s across the actuator length of travel.

The participant stands in front of the apparatus, gets a ball at point A and transfers it to another specified location (Figure 3-2). The balls weighed 59 g and had a 6.5 cm diameter. The distance was calibrated against a measured grid in the image (Figure 3-2).

A fourth-generation i7 quad-core computer recorded video of the task performed. A Logitech C920 fixed focal length web camera was used for imaging the color video stream that was stored as an AVI video file (Xvid compressed) with 640×480 pixels resolution at the frame rate of 30 frames per second.

Estimation of DC requires identifying instances when the hand is loaded and exerts a force against an object in a work cycle. In this simulated task, exertions occur during the time elapsed after the ball is grasped, moved, and until it is released. The sequence of this task can be described

as Reach-Grasp-Move-Release (Figure 3-3). After the subject grasps a ball, the Move element starts and continues until the ball is released. We further define the sequence Move-Release as 'Put', and the sequence Reach-Grasp as 'Get' (Figure 3-3). Thus, Put time represents exertion time, Get time represents rest time when the hand is not interacting with the object, while the total task time cycle time is the sum of the elapsed Put and Get times. The DC for the task is therefore the percentage time: Put/(Put + Get).

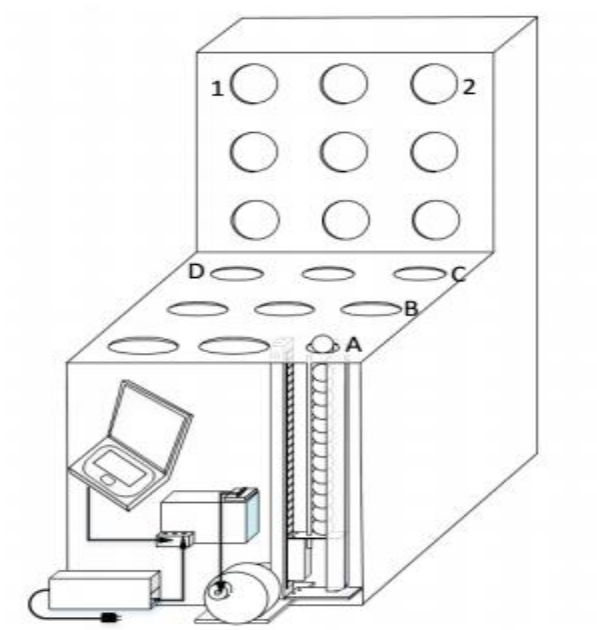


Figure 3-1 Laboratory task simulation apparatus
The participant grasps a ball from location A and moves it to either locations B, D, 1, or 2, depending on the specified task.

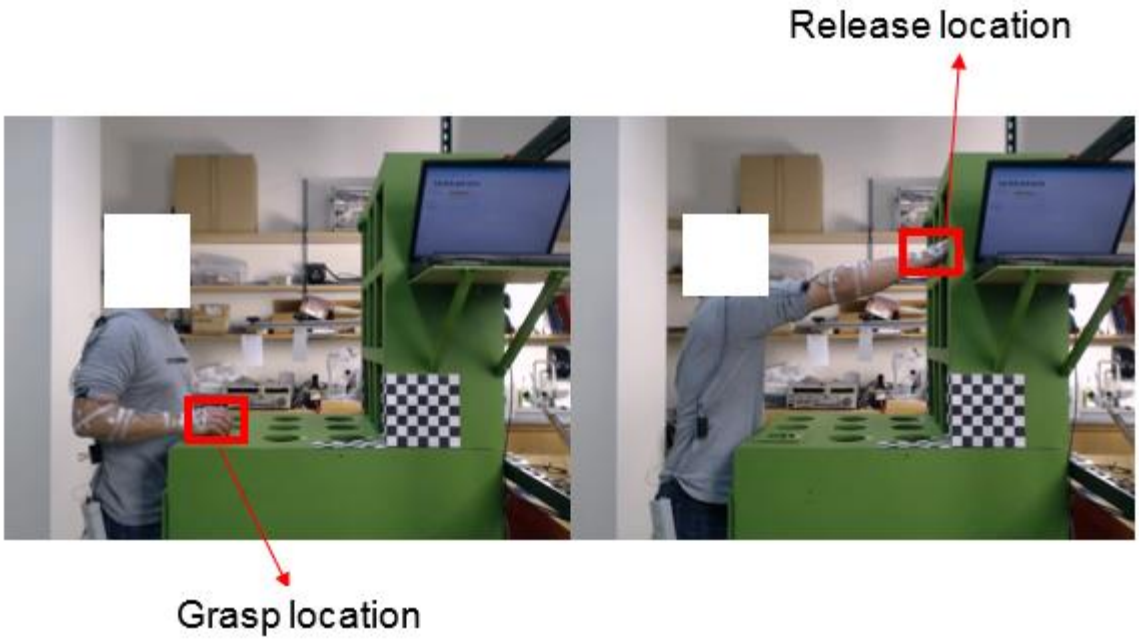


Figure 3-2 Representation of grasp and release location

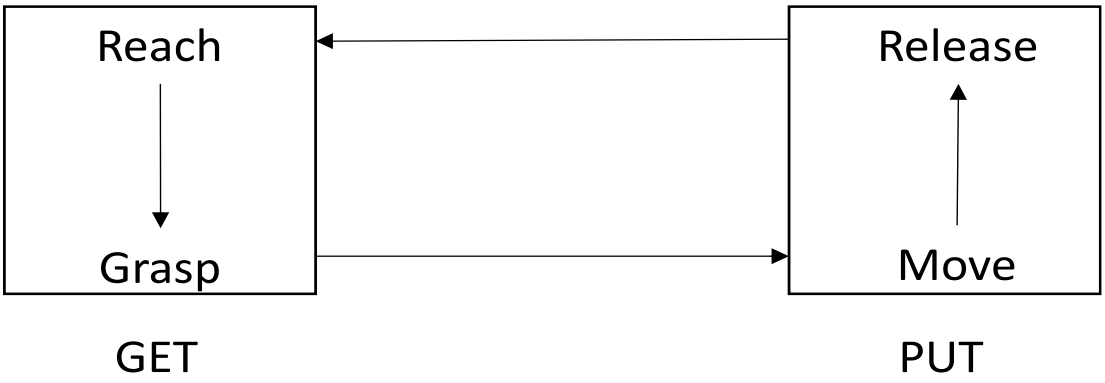


Figure 3-3 Graphical representation of the simulated task

We recruited 19 university student volunteers (6 males and 13 females) with informed consent and IRB approval. They were each video-recorded while performing 15 cycles of the Get-Put task paced at various frequencies; exertion and rest times were controlled using an auditory cue. The paced frequencies and DC for each task are given in Table 3-1. We calculated the paced HAL values for each task based on these frequencies and DC using the equation for HAL in [62], which

are also provided in Table 3-1. The observed HAL for each participant in every task was calculated from frequency and DC measured using MVTA (Table 3-1). These HAL values were used as ground truth measures for testing the computer vision algorithms.

Each subject performed 10-paced tasks in random order. A set of practice tests was provided for each condition. A one-minute rest was provided between each condition to prevent fatigue. Each video clip had a length ranging from 10 to 80 s and consisted of 15 cycles of task execution. Some participants were unable to accomplish every pace and those cases were excluded from the analysis. This was due to a combination of challenging experimental conditions or failures. Failures occurred when the subject missed or dropped the ball. The number of error-free video clips for these experiments are listed in Table 3-1.

Table 3-1 Summary of tasks performed and pacing

Location	Frequency	DC	Paced HAL	Measured Ground Truth HAL		N
	(Hz	(%)		Mean	SD	
A to 1	0.25	65	2.9	2.9	0.1	18
A to 1	0.75	60	5.8	5.6	0.2	19
A to 2	0.50	20	3.5	3.9	0.2	16
A to 2	1.25	47	6.4	6.5	0.2	18
A to D	1.25	47	6.4	6.5	0.2	19
A to D	1.50	25	5.6	6.7	0.3	23
A to C	1.00	15	4.2	5.6	0.4	17
A to C	1.50	25	5.6	6.7	0.3	22
A to C	1.75	25	5.8	6.9	0.2	18
A to B	1.75	25	5.8	6.8	0.3	19

3.1.2 Factory Video Clips Selection

We extracted 72 videos clips from the real industrial videos selected from [61]. These videos include different kinds of tasks such as picking mushrooms, picking up blocks, sewing, and brushing, etc.

3.2 Method

3.2.1 Training and Test Data

In developing the algorithms, the entire video data were divided into non-overlapping training and test data sets. A total of 41 video clips of the repetitive laboratory task performed by the first five subjects (nine clips were excluded due to failures) were reserved for the training data and used for training and validating the developed algorithms. There were 87 video clips available for the remaining 14 subjects (53 clips were excluded due to task incompleteness or failures) were reserved as the test data-set.

Actual factory jobs involve different workflows, so in practice, few training video clips may be available for training the algorithm. Therefore, in developing the FVT method, we also experimented with a *first cycle* data partition method. Specifically, we manually labeled the Get and Put elements for the first cycle of the repetitive task in the video and used them for training the FVT algorithm. Then, we tested the algorithm on the remaining cycles in the video as the test data.

3.2.2 Ground Truth Data

Trained analysts extracted ground truth DC measures from the videos of each task using single frame video coding and MVTA software. They marked each frame when they identified a change from Get to Put or from Put to Get. After a frame was marked, all the remaining frames were marked using the same label until the next change occurred. The start of exertion (i.e. start of Put) was identified as the instant when the hands contacted the ball while the end of exertion (i.e. start of Get) was identified as the instant when the ball no longer made contact with the hand.

3.2.3 Hand Location

The hand location on each video frame was tracked using the marker-less video tracking algorithm described in [28], [29], [63]. The analyst initially identifies a rectangular ROI covering the image of the entire hand, which was tracked for each frame by the computer. A cross-correlation template matching tracking algorithm tracks the ROI center trajectory *(x_i, y_i)* over subsequent video frames.

3.2.4 Feature Vector

Based on the trajectory *(x_i, y_i)*, other kinematic features may be derived:

Velocity $v_{x,i} = (x_{i+1} - x_{i-1})/2\Delta$, $v_{y,i} = (y_{i+1} - y_{i-1})/2\Delta$ (3.1)

Speed $|v_i| = \sqrt{(v_{x,i})^2 + (v_{y,i})^2}$ (3.2)

Acceleration $a_{x,i} = (x_{i+1} - 2x_i + x_{i-1})/\Delta^2$, $a_{y,i} = (y_{i+1} - 2y_i + y_{i-1})/\Delta^2$ (3.3)

Acceleration Magnitude $|a_i| = \sqrt{(a_{x,i})^2 + (a_{y,i})^2}$ (3.4)

We also computed the spatiotemporal curvature. The curvature function *k* is sensitive to the change of the direction of a curve relative to its arc length. It is defined as:

$$k = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{3/2}} \quad (3.5)$$

where (x(t), y(t)) is a planar curve. This function is useful for detecting discontinuities in a movement based on velocity, acceleration, and position [64]. The spatiotemporal curvature index peaks when there is a significant change in the action, such as changing direction, stopping, or starting to move. These changes correspond to instances in motions such as grasps or releases [65]. The curvature index K_i is:

$$K_i = \frac{\sqrt{a_{x,i}^2 + a_{y,i}^2 + (v_{x,i}v_{y,i} - a_{x,i}a_{y,i})^2}}{(\sqrt{a_{x,i}^2 + a_{y,i}^2 + 1})^3} \tag{3.6}$$

A representative set of feature vectors for the laboratory task are shown in Figure 3-4, including corresponding location, velocity, acceleration, and spatiotemporal curvature.

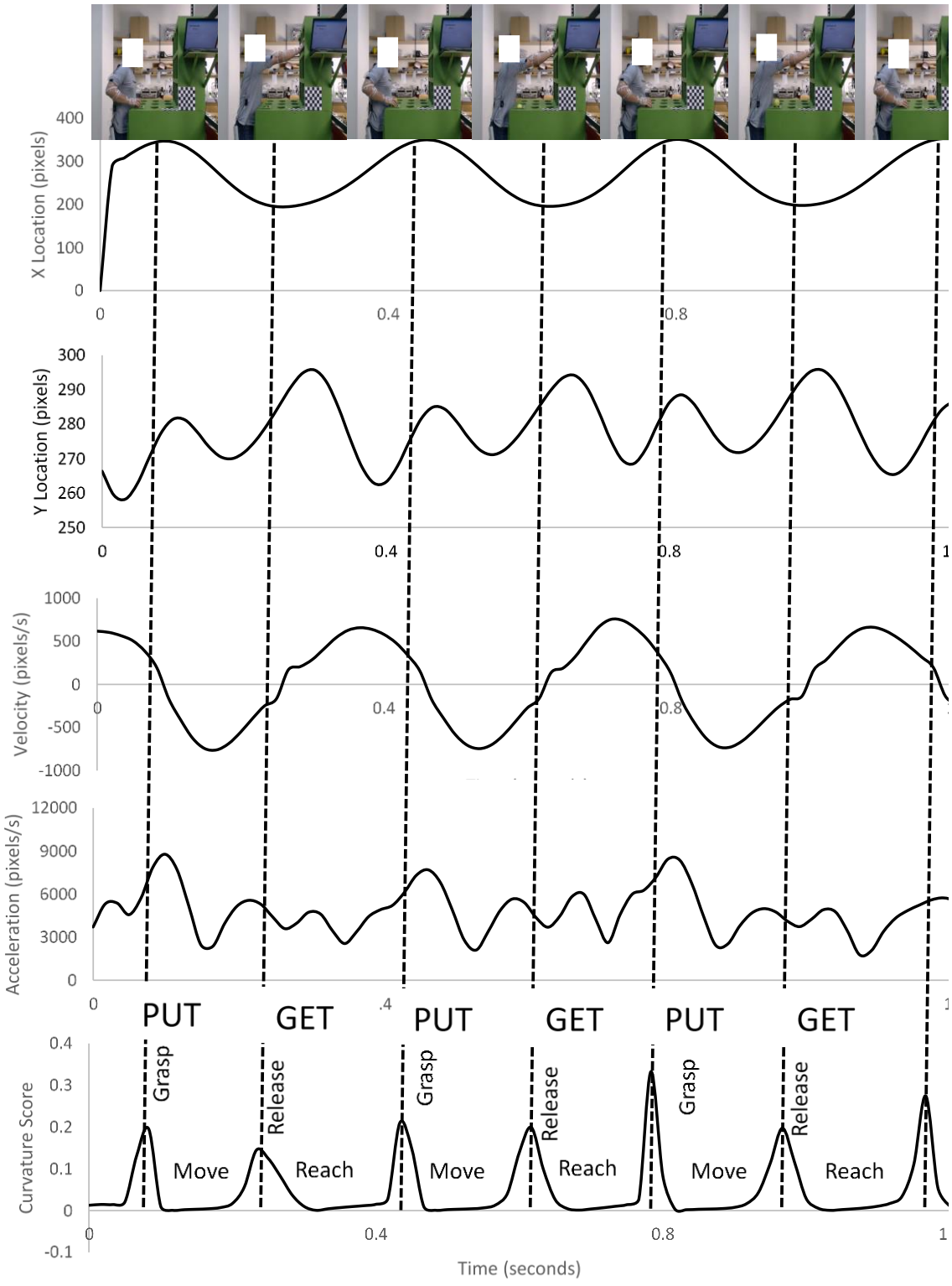


Figure 3-4 Representative hand kinematic feature curves for location, velocity and acceleration aligned with the spatial-temporal curvature score measured using ROI marker-less video tracking

3.2.5 Duty Cycle and HAL Measurement

The DC estimation process was divided into two phases. In the first phase, the hand movement in each video frame was classified as one of two states (Get and Put). The DC for the n^{th} cycle DC_n was estimated as:

$$DC_n = \frac{No.\,frames\ in\ Put\ State\ (\ n^{th}\ cycle)}{No.\,frames\ in\ Put\ State\ (\ n^{th}\ cycle) + No.\,Frames\ in\ Get\ State\ (\ n^{th}\ cycle)} \quad (3.7)$$

Since each video consisted of multiple cycles, multiple estimates of the DC were measured. The overall DC estimates were evaluated using the average of these individually estimated DC.

After the average DC estimate was obtained for a given task and participant, the corresponding *HAL* can be calculated using the equation from [59].

$$HAL = 10 \cdot \left[\frac{e^{-15.87+0.02\,DC+2.25\ln S}}{1+e^{-15.87+0.02\,DC+2.25\ln S}}\right] \quad (3.8)$$

where *s* is RMS speed in mm/s.

3.2.6 Feature Vector Training Algorithm

The FVT algorithm employed machine-learning procedures to systematically develop the DC estimator and hence potentially demands less manual tuning efforts. This is desirable when a large amount of repetitive hand movement videos from different factory work environments are to be analyzed automatically.

3.2.6.1 Applied on Laboratory

We have developed a feature selection method to automatically select a subset of features in order to optimize DC estimation performance. We first used cross-validation and the maximum DC estimate error as the performance criterion to judge whether a specific subset of nine features

(Table 3-2) was most desirable. Maximum DC estimate error was the maximum value of the absolute difference between the estimated DC and ground truth DC over all tasks. Utilizing the 41 training videos, we employed a feature selection procedure to train the algorithm on 40 video clips and then tested the result on the remaining video in the set. We rotated this process for different video clips 41 times and averaged the performance evaluated on the testing video. Since there were $2^9 = 512$ combinations of the 9 features described above, it was too tedious to try every combination exhaustively when the number of features increased. Instead, we opted to use the greedy backward subset selection method described in [66].

We started with evaluating the performance (maximum DC estimate error) using the entire set of nine features. Then, we evaluated the performance of the nine different subsets of eight features each with one feature excluded from the currently selected nine features. And we selected the eight-feature subset that produces the best DC estimate.

We then repeated this process in order to select the best seven-feature subset out of the previously selected best eight-feature subset. This process is repeated until there is only one feature remaining. We compared the accuracies of the estimated DC values from these selected subsets of features and chose the one that yielded the best performance.

Using this process, the nine features are ranked from one to nine, with nine being the most important feature appearing in all nine subsets, and one being the least important feature discarded first when eight features were selected among the nine. The ranking of these nine features is given in Table 3-2. The maximum DC estimate errors of nine different feature sets are plotted in Figure 3-5 where a smaller DC estimation error indicates better performance. Based on these results, we selected the following subset of six features (x, y, v_x, v_y, $|v|$, K) to be used in the FVT algorithm.

We observed that v_x (rank = 9) was the most important feature and a_x (rank = 1) was the least important feature.

Note that x and y are hand coordinates, v_x and v_y are x and y-direction velocities, $|v|=\sqrt{v_x^2+v_y^2}$ is the speed, a_x and a_y are x and y-direction accelerations, $|a|=\sqrt{a_x^2+a_y^2}$ is the magnitude of acceleration, and K is the spatial-temporal curvature.

Table 3-2 Ranking of Selected Features

Feature	x	y	v_x	v_y	$\|v\|$	a_x	a_y	$\|a\|$	K
Rank	8	5	9	7	6	1	2	3	4

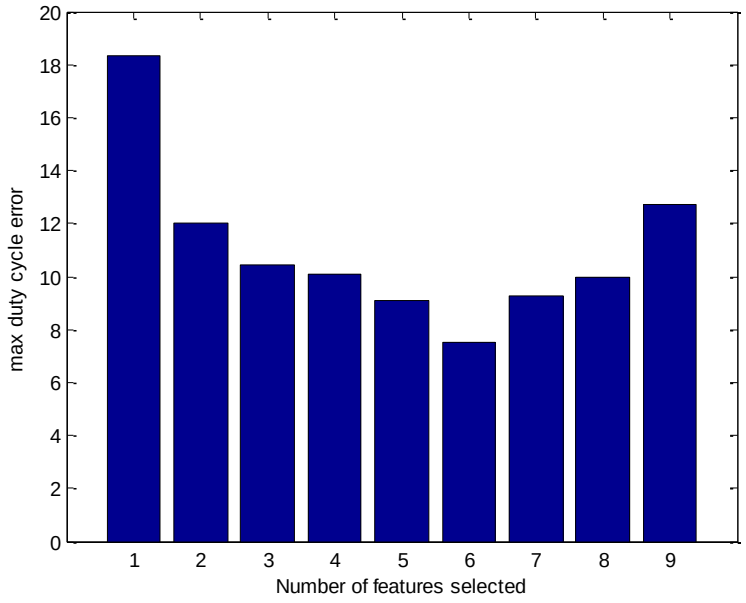


Figure 3-5 The maximum duty cycle errors (ground truth DC – estimated DC) of nine different feature sets

Since these features had quite different ranges, in order to avoid one feature with large magnitudes dominating another feature, a normalization step was applied to ensure that the extracted feature values were approximately zero mean with a sample variance equal to unity.

In this work, we use the *k* nearest neighbor (kNN) classifier as the state estimator. In a kNN classifier, all training data (feature vectors) and corresponding labels (Get, Put) are stored in the classifier. For each test feature vector, a set of *k* (usually chosen as an odd number to facilitate majority voting) training vectors that are most similar to the test vector are identified using a similarity metric. Then, the training vector dominant label (by majority voting of the *k* labels) is assigned to the test vector. In these experiments, we choose $k = 5$ when training with the 41 training videos, and used the Euclidean distance between the training and the test feature vectors as the similarity metric. When training with the first cycle samples, we choose $k = 1$ since in practice there may be few training samples available in the first cycle.

As previously discussed, two different approaches of training and testing were applied to the FVT based algorithm. The first used the 41 video clip training data-set and testing the kNN classifier with the 87 testing dataset (kNN_r). The second approach, called first cycle training, used the feature vectors and corresponding labels of the first cycle of a test video clip as the training data and estimated the DC of the remaining cycles of the same test video clip (kNN_f). The 41 training video clips were not used by the kNN classifier in the first cycle approach. Detailed steps of the training-based DC estimation algorithm are summarized in Algorithm 3-1.

Training Phase

inputs: feature vectors, state labels (Get and Put)

1. Normalize all features into zero mean and unit variance, and store the normalization factors for the testing phase.
2. Train the k-nearest neighborhood classifier by using the chosen feature vectors and the MVTA labels of the training data.

Testing Phase

inputs: frame number, feature vectors

output: states (Get or Put) for each frame

1. Normalize all features with the normalization factors from the training phase.
2. Input the feature vectors into the classifier trained in the training phase frame by frame to classify each frame to be Get or Put.
3. The estimated duty cycle is computed by (# of Put)/((# of Get) + (# of Put))

Algorithm 3-1 Feature Vector Training Algorithm

3.2.6.2 Applied on Factory Videos

There was a wide range of cycle times among the industrial tasks, ranging from less than one second to more than ten seconds. A short first cycle might only be ten frames. Since we only have one cycle as training data, feature selections by cross-validation methods were infeasible. Instead of selecting the best subset of features, we made use of all hand trajectory features which included location, velocity, speed, acceleration, acceleration magnitude, and spatiotemporal curvature.

3.2.7 Camera Motion Compensation

Unlike the stationary camera used in the laboratory, the industrial videos were recorded using hand-held cameras. Therefore, the videos shook, and sometimes the viewing angles changed over time. These factors made the tracking of the hand trajectory irregular and increased the difficulty of estimating the duty cycle by tracked hand trajectory alone.

To compensate for the motion of the camera, first, we tried to find out the geometry transformation between each couple of successive frames. Given two successive frames, we detected the Speeded Up Robust Features (SURF) [67] on both of them, and extracted the matched points using the sum of squared differences (SSD) as our feature matching metric. We calculated the affine transform matrix between the matched points and excluded outliers using the M-

estimator Sample Consensus algorithm (MSAC), which is a variant of the Random Sample Consensus algorithm (RANSAC) [68].

Assuming the transform matrix between the i-th frame and the (i+1)-th frame is Ti, the cumulative transformation of the (i+1)-th frame and the first frame will be the product of T1 to Ti, which is

$$H_{i+1} = \prod_{n=i}^{1} T_i \tag{3.9}$$

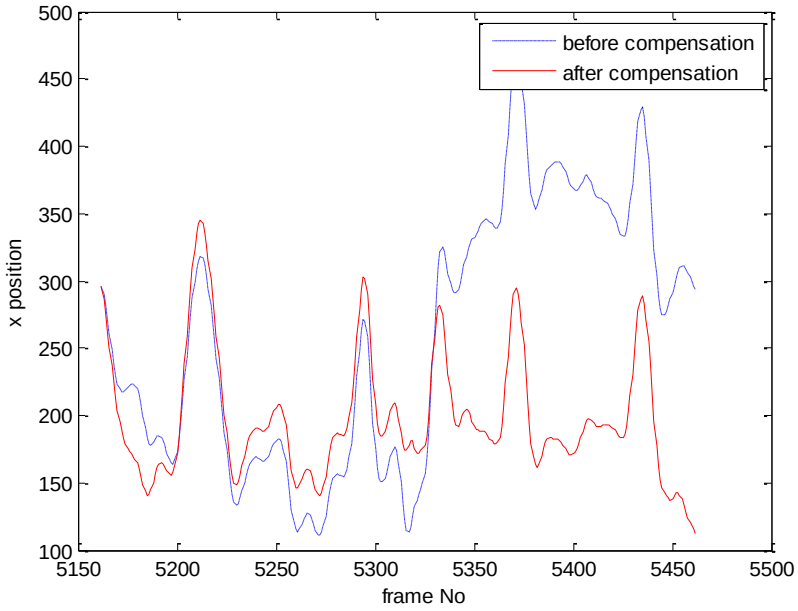
Then, we can compute the compensated hand coordinate of the i-th frame, (x_{ic}, y_{ic}), by

$$[x_{ic} \quad y_{ic} \quad 1] = [x_i \quad y_i \quad 1] * H_i \tag{3.10}$$

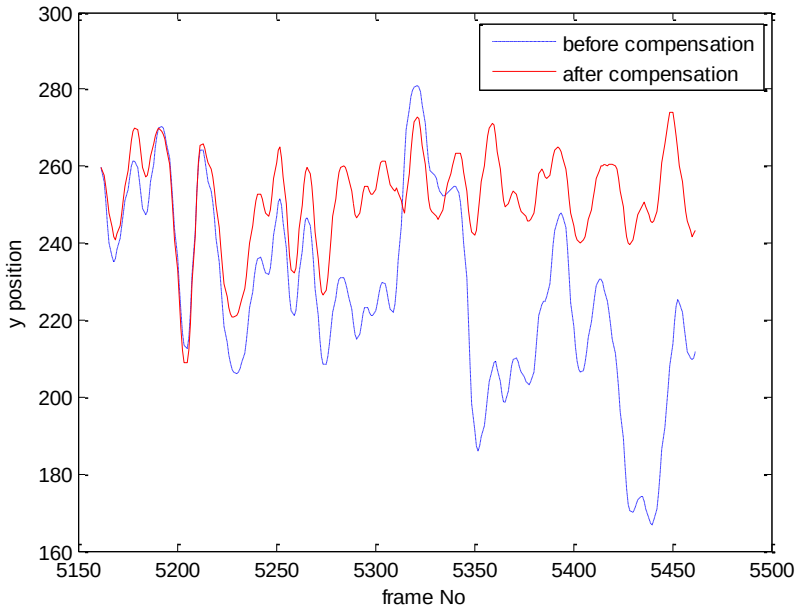
where (x_i, y_i) is the tracking hand coordinate before compensation. Notice that the transform operators are 3 x 3 matrices.

An example of the x and y-direction of hand trajectories before and after camera motion compensation are shown in Figure 3-6 (a) and (b) respectively.

We can see that the hand trajectories tend to concentrate in a smaller range after the camera motion compensation, which is closer to the situation when the camera is stationary.



(a)



(b)

Figure 3-6 Hand trajectories before and after camera motion compensation. (a) The hand trajectory in the x-direction. (b) The hand trajectory in the y-direction.

3.2.8 Sensitivity Analysis

Consider the speed and duty cycle based HAL equation from [59].

$$HAL = 10 \cdot \left[\frac{e^{-15.87+0.02\,DC+2.25\ln S}}{1+e^{-15.87+0.02\,DC+2.25\ln S}} \right] \tag{3.11}$$

where *s* is RMS speed in mm/s.

Let $y = -15.87+0.02 \cdot DC+2.25 \ln S$, then $HAL = 10 \cdot e^y/(1+e^y)$, or

$e^y = (HAL)/(10 - HAL)$, and $1/(1+e^y) = (10-HAL)/10$.

Furthermore, $de^y/d(DC) = e^y \cdot (0.02)$. Also,

Hence

$$\begin{aligned} \frac{d(HAL)}{d(DC)} &= \frac{d(HAL)}{de^y}\frac{de^y}{d(DC)} = \frac{10}{(1+e^y)^2}e^y \cdot (0.02) \\ &= 10 \cdot \left(\frac{10-HAL}{10}\right)^2 \cdot \frac{HAL}{10-HAL} \cdot (0.02) \\ &= 0.002 \cdot (HAL) \cdot (10-HAL) \end{aligned} \tag{3.12}$$

Since $0 \le HAL \le 10$, $(HAL) \cdot (10-HAL) \le 25$ where the maximum occurs when $HAL = 5$. Therefore, the impact of change of *DC* to that of the *HAL* value can be bounded by

$$|\Delta(HAL)| \le 0.05 \cdot |\Delta(DC)| \tag{3.13}$$

Since when $|\Delta(HAL)|$ is greater or equal to 0.5, the rounded *HAL* value will be erroneous, thus, one can see that as long as $|\Delta DC| < 10\%$, the estimated *HAL* value will be correct assuming the speed estimate *S* is correct. Note that *DC* in this equation is a relative error (range from 0 to 100%).

3.3 Result

3.3.1 Simulated Task Videos

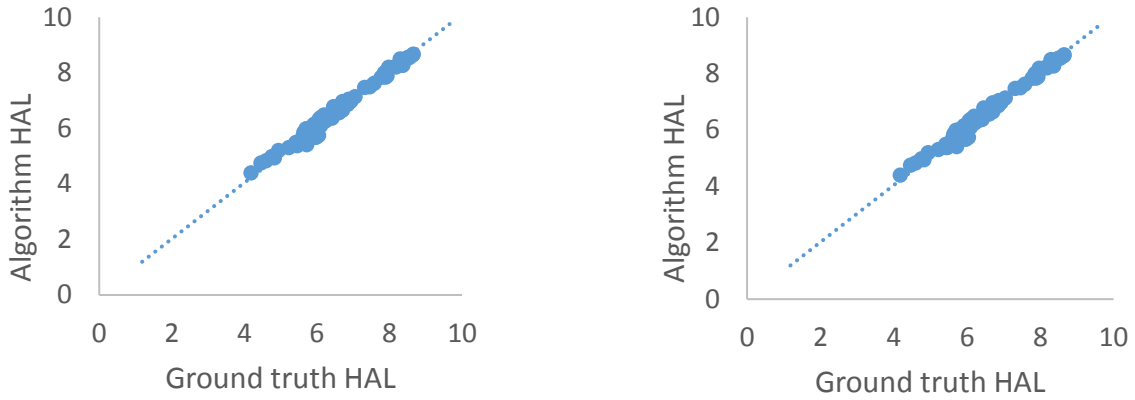
DC estimate error (%), was calculated as the average absolute difference between ground truth DC (%) and predicted DC (%) for the two different algorithms (kNN_r, kNN_f). We provide a summary of the DC estimation errors of these two methods in Table 3-3. The Max and Min give the maximum and minimum absolute values of estimate errors. The Mean and SD give the mean and standard deviation of the estimate errors. We also calculated the HAL estimation errors in Table 3-4. To better understand the HAL estimation outcome, we also provide scatter plots of the HAL estimation outcome by the kNN_r, kNN_f algorithms versus the ground truth HAL values and summarize the results in Figure 3-7.

Table 3-3 Duty Cycle Estimation Error Summary (Unit: percent)

	kNN_r	kNN_f
Max	8.8	10.9
Min	0.1	0.2
Mean	2.8	3.3
SD	2.1	2.5

Table 3-4 HAL Estimation Error Summary (Unit: HAL units)

	kNN_r	kNN_f
Max	0.4	0.4
Min	0.0	0.0
Mean	0.1	0.1
SD	0.1	0.1



(a) kNN_r Algorithm (R^2=.99) (b) kNN_f Algorithm (R^2=.99)

Figure 3-7 Predicted HAL versus ground truth HAL kNN_r and kNN_f algorithms
(a) *k* Nearest Neighbour using 41 video training set (kNN_r), (b)*k* Nearest Neighbour using first cycle training set (kNN_f). The number of the videos in the test set is 87.

3.3.2 Factory Videos

A summary of the statistics of the DC estimate error is provided in Table 3-5. We also calculated the HAL estimate error as the average absolute difference between the ground truth HAL and predicted HAL in Table 3-6. The HAL was computed by the HAL equation, given the corresponding root mean square speed and duty cycle. Within the 72 videos, 19 videos contain tasks in which the subject is holding the object almost all the time. In these cases, the state in every frame of the first cycle is always a Put. As a result, the prediction for every frame will also be a Put, which makes a 100% correct prediction. For that reason, we separated the results of the overall videos and the result of only non-holding videos in different columns. The results before and after camera motion compensation are all listed for comparison. We show the scatter plot of predicted HAL vs. ground truth HAL in Figure 3-8 and the histogram of the HAL estimate error in Figure 3-9 to understand the result better.

Table 3-5 Duty Cycle Estimate Error Summary (Unit: percent)

	All Videos	All Videos after Compensation	Non-holding Videos	Non-holding Videos after Compensation
Max	52.52	33.02	52.52	33.02
Min	0.00	0.00	0.25	0.00
Mean	8.35	7.83	10.95	10.25
SD	9.76	9.29	10.00	9.56

Table 3-6 HAL Estimate Error Summary (Unit: HAL units)

	All Videos	All Videos after Compensation	Non-holding Videos	Non-holding Videos after Compensation
Max	1.90	1.55	1.90	1.55
Min	0.00	0.00	0.01	0.00
Mean	0.37	0.34	0.48	0.44
SD	0.39	0.39	0.39	0.40

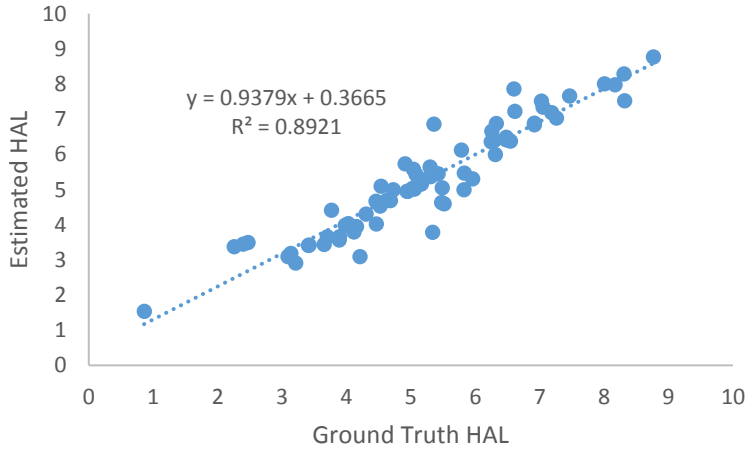


Figure 3-8 Predicted HAL versus ground truth HAL (The kNN_f algorithm)

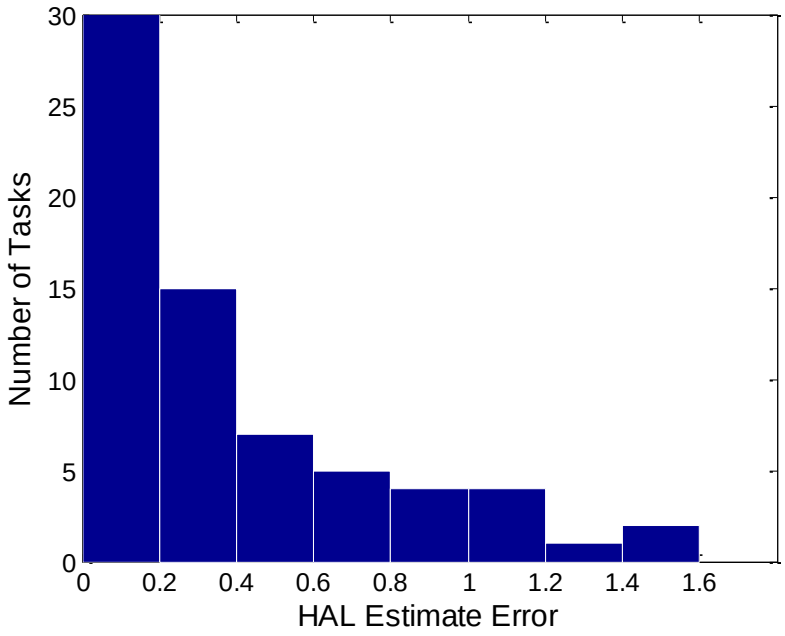


Figure 3-9 Histogram of the HAL estimate error (The First Cycle Training algorithm)

3.4 Discussion

In our first cycle training algorithm, one of the assumptions is that the tasks are involved with cyclic hand motions. The larger DC estimate errors for some videos might come from the larger deviation of the motions in different cycles of the same factory videos. Choosing a typical cycle that can represent the task or choosing more than one cycle for training data might help to alleviate the errors.

In Section 3.2.7, our sensitivity analysis has shown that the HAL estimate error should be less than 0.05 times the DC estimate error. In Figure 3-10, the HAL estimate error vs. the DC estimate errors for all videos are plotted. The points are all located below the line of y=0.05x, which verifies our conclusion.

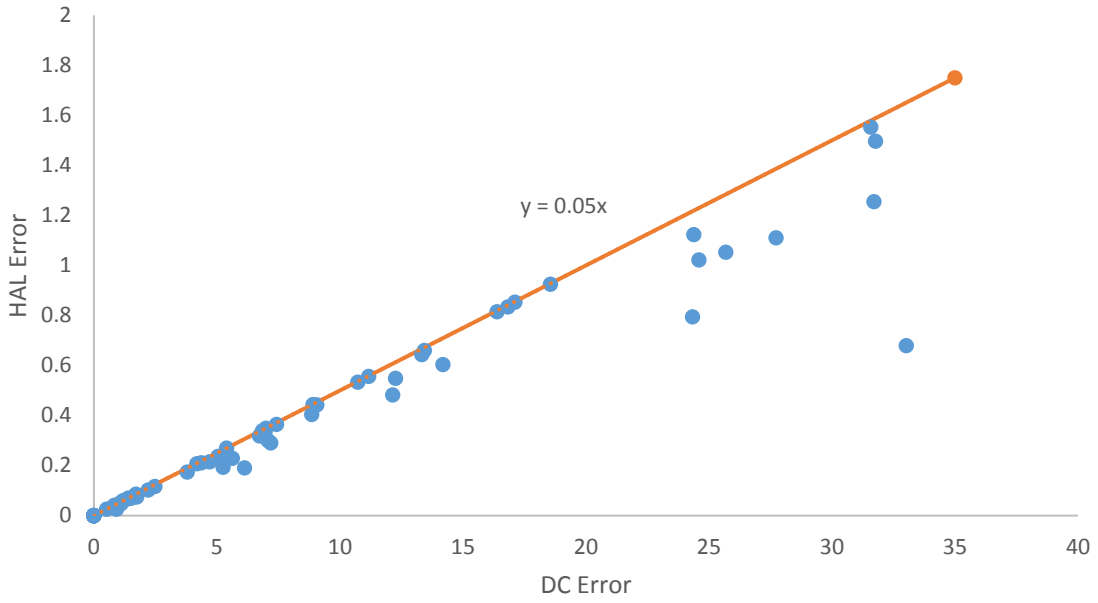


Figure 3-10 HAL estimate error versus DC estimate error
Each point represents a video. It verifies that $|\Delta(HAL)| \leq 0.05 \cdot |\Delta(DC)|$ since all points are below the y=0.05x line.

We used the absolute DC error, that is the difference between the ground truth DC and the predicted DC, as the criterion to estimate our performance. It might not be able to precisely represent the ability of the algorithms since false alarms and missed detections would cancel each other out. That means the algorithm having the minimum DC estimate error might not have the highest classification rate.

In real cases, the ground truth states are usually a bunch of consecutive GET states followed by a bunch of consecutive PUT states in turn. Our feature vector training algorithm estimated whether the current state was GET or PUT frame by frame and did not utilize the temporal information. The accuracy of the DC estimation would be improved if we can exploit the sequential property of our data.

Chapter 4.

Video-based Automatic Wrist Flexion and Extension Classification

4.1 Method

The method assumes that each video frame is viewed from a direction perpendicular to the sagittal plane of the hand. Hence the 2D coordinates of three upper limb joints, the elbow, wrist, and the back of the hand (knuckle) were measured using an open-source human body key point estimation package, OpenPose [91]. From these three key points, the wrist flexion angle was computed. These estimated angles were classified into three hand-posture classes: flexion, neutral, or extension. From a laboratory experiment, we collected 1464 video frames from 61 videos of controlled repetitive hand tasks performed by 16 participants. Each video frame yielded a wrist flexion angle estimate. Motion capture (MoCap) IR markers of the Optotrak motion capture system [92] were attached to the upper limb of the participants to provide ground truth measurements of the three key points. The same wrist flexion angle was estimated from these ground truth keypoint locations and similarly classified into one of the three classes as ground truth. Finally, the computer-vision hand state classification results were compared against the ground truth label. In addition to the laboratory experiment data, we also applied this algorithm to a set of videos acquired from a field study (the SHARP study [93]).

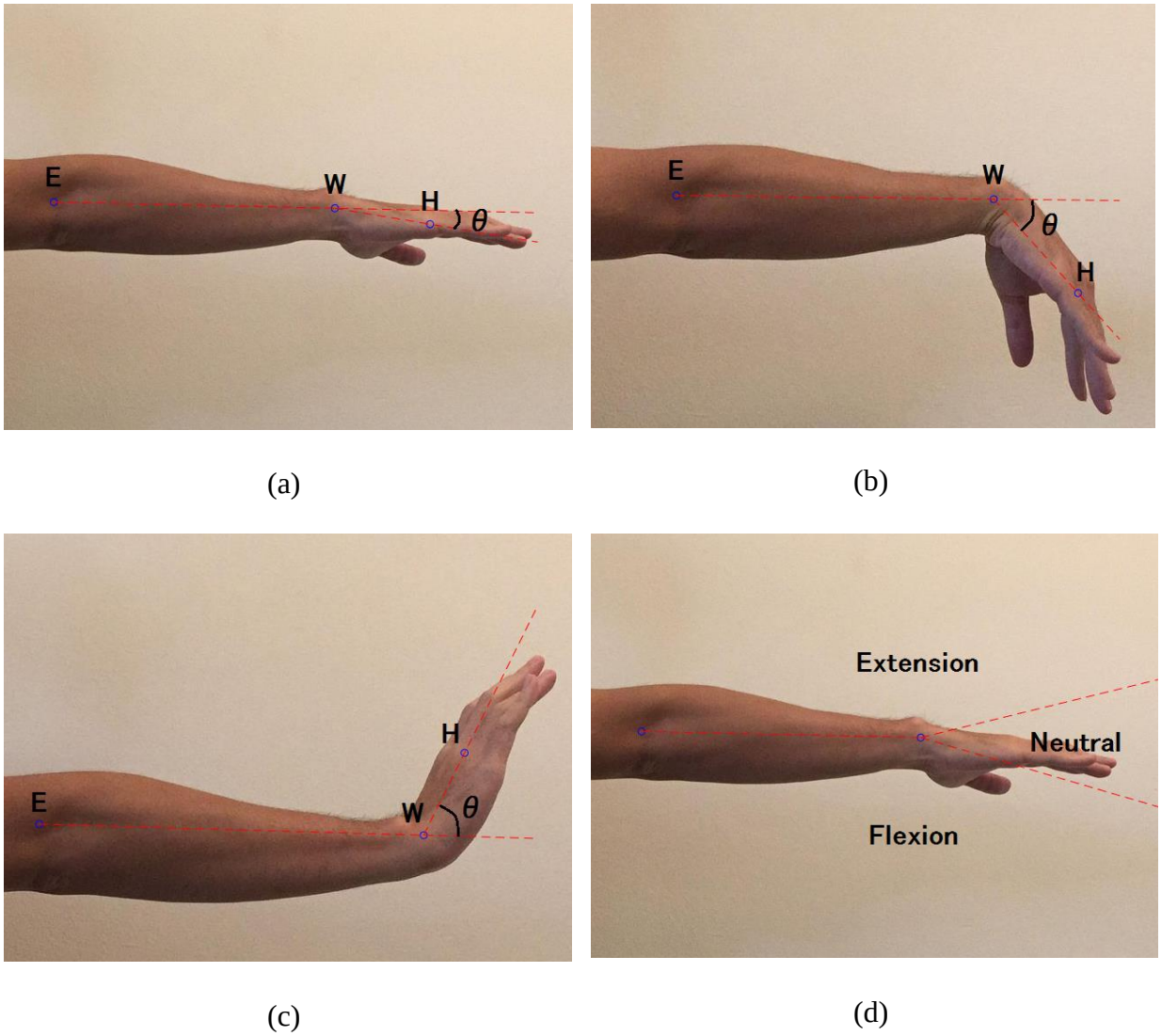


Figure 4-1 Types of wrist postures (a) neutral (b) flexion (c) extension (d) Ranges of the three postures.

4.1.1 Wrist Flexion/Extension Angles

Referring to Figure 4-1, when the palm faces downward, the angle between the elbow (E), wrist (W), and the knuckle of the middle finger on the back of the hand (H) may be used to define three states of hand postures: flexion, neutral, and extension.

These classifications are determined by the wrist flexion/extension angle θ. Assume the vector $\overrightarrow{EW}$ points to the positive direction of the reference axis. θ is the angle between the vector $\overrightarrow{WH}$

and the reference vector $\overrightarrow{EW}$. If the hand bends toward the back of the hand, θ is positive. If the hand bends toward the palm, θ is negative. The magnitude of θ should be smaller than 90°. For example, the angle θ in Figure 4-1 (a) is approximately zero, (b) is negative, and (c) is positive (or extension).

Based on widely used posture observational methods (e.g. [81]), we define the wrist posture as neutral if θ is between -15 to 15 degree. If $\theta > 15^{\circ}$, the hand posture state is extension. If $\theta < -15^{\circ}$, the hand posture state is flexion.

4.1.2 Wrist Flexion Angle Estimation Algorithm

Denote the image coordinates of three manually selected landmark points H, W, and F as (x_h, y_h), (x_w, y_w) and (x_E, y_E) respectively. The angle θ is the angle between line segments $\overrightarrow{WH}$ and $\overrightarrow{EW}$. An example of the three landmark points is shown in Figure 4-2.

Denote x-y image coordinates of vectors $\overrightarrow{WH}$ and $\overrightarrow{EW}$ as:

$$\overrightarrow{WH} = (x_h - x_w, y_h - y_w) \tag{4.1}$$

and

$$\overrightarrow{EW} = (x_w - x_e, y_w - y_e) \tag{4.2}$$

Using trigonometry, one has

$$|\theta| = \cos^{-1}\left(\frac{\overrightarrow{EW} \cdot \overrightarrow{WH}}{\left\|\overrightarrow{EW}\right\|\left\|\overrightarrow{WH}\right\|}\right) \tag{4.3}$$

θ is negative if the rotation is toward the palm direction (flexion) and positive otherwise (extension).

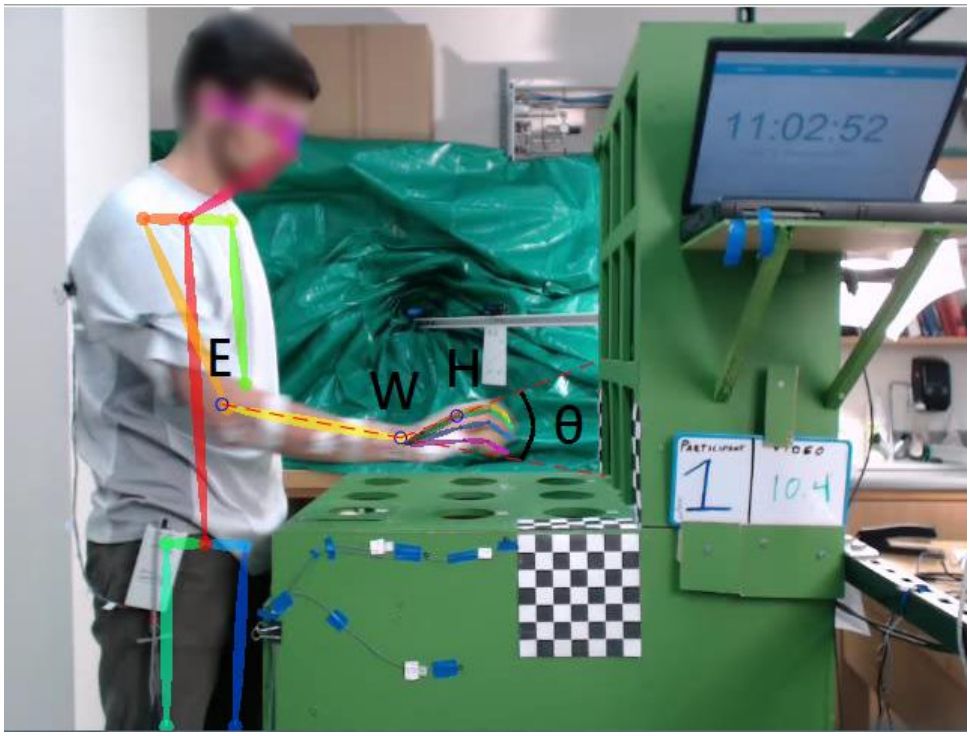


Figure 4-2 Three tracked points and the hand posture angle θ.

4.1.3 Skeletal Joint Position Estimation

To estimate skeletal joint locations in the frame of the elbow (E), the wrist (W), and the knuckle on the back of the hand (H), an open-source 2D human pose estimation software package Openpose [91] was applied. Openpose takes a video frame as the input to a two-branch Convolutional Neural Network and predicts a set of 2D confidence maps of body part locations and a set of 2D vector fields of part affinities simultaneously. A greedy inference is then applied to parse the confidence map and affinity field to output 2D coordinates of 25 key points corresponding to 25 body skeletal joints of the subject. In this work, we use the key point locations of the elbow, wrist, and hand (knuckle) of one hand. The OpenPose algorithm can automatically start when presented with a video clip and will estimate key point locations for each video frame. The accuracy of the joint position estimation is compromised if part of the limb is occluded. In

this laboratory experiment, the camera is placed facing the sagittal plane to minimize occlusion. However, for videos taken from factories, manual screening of the videos is required to select segments of frames where the arm and hand movements are always visible. A diagram of the proposed wrist flexion angle estimation algorithm is summarized in Figure 4-3.

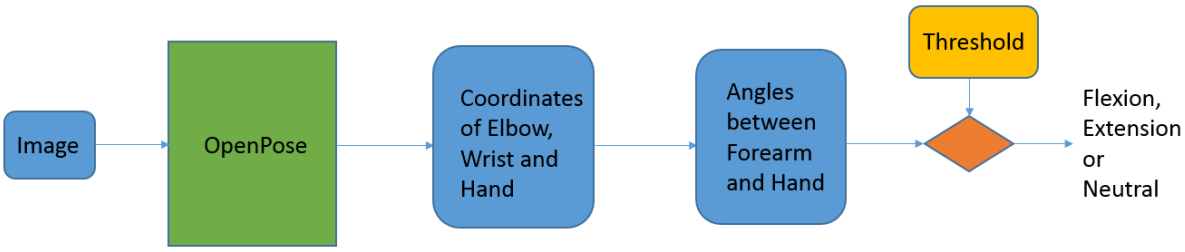


Figure 4-3 Diagram of our method.

4.2 Experiment

4.2.1 Data Collection of Laboratory Videos

The data set consists of 61 videos from 16 participants (3 males and 13 females). These videos were selected among videos recorded in the laboratory of a simulated repetitive motion task [94]. The subjects were recruited with informed consent and IRB approval. In each video, the participant stands in front of an apparatus and moves tennis balls from one location to another for 15 cycles, as demonstrated in Figure 4-2 and Figure 4-4.

The apparatus (Figure 4-4) is comprised of an 840 mm travel length linear belt drive actuator (Misumi MSS-625) driven by a bipolar stepper motor (ElectroCraft Model TPP34 with 560 N cm torque) and controlled by a stepper motor controller (IMS MX-CS101-401). The device can move a 2 kg object every 0.5 s across the actuator length of travel.

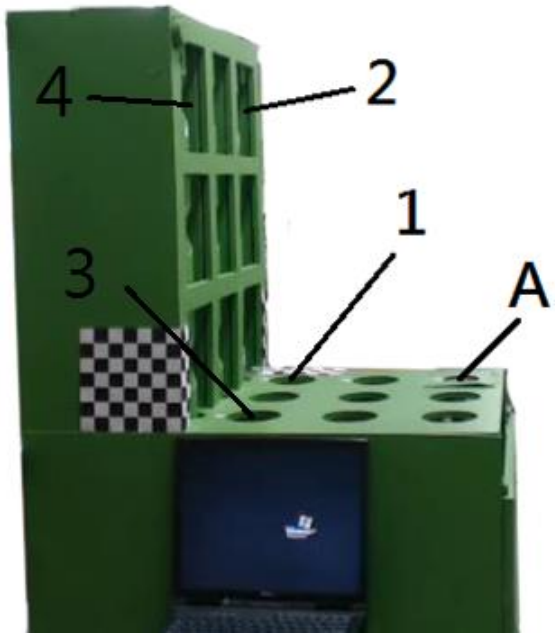


Figure 4-4 Laboratory task simulation apparatus. The participant grasps a ball from location A and moves it to either location 1 or 2, depending on the specified task.

The hands were videotaped from the side view (sagittal plane) with the palm of the right hand facing downward. The selected videos included two kinds of tasks, moving the tennis balls from location A to 1 and from location A to 2 (as shown in Figure 4-4). Videos were selected in which the hand of the subject was always in the sagittal plane (location from A to 1 and A to 2).

An Optotrak 3D infrared motion tracking system (MoCap) was used to track IR beacons attached to the hands and arms of the subject as shown in Figure 4-2. The MoCap system returns accurate estimates of 3D coordinates of the skeletal joints of the elbow, wrist, and hand based on the tracked trajectory of these markers. These 3D coordinates are specified relative to a world coordinate system whose x-y plane is not guaranteed to be parallel to the sagittal plane of the subject. To remedy this problem, we choose an initial video frame before the subject started moving the hand for camera calibration. We used six pairs of 3D coordinates of skeletal joints from the MoCap system and corresponding 2D image coordinates estimated using Openpose to

align the MoCap system relative to the video camera. The MoCap estimated 3D joint coordinates of the elbow, wrist, and hand estimated were projected to give corresponding 2D image coordinates. The same formulas in Equations (4.1), (4.2), and (4.3) were used to calculate the ground truth hand-pose angles. After obtaining the transformation matrix, we mapped the MoCap 3D coordinates into the image 2D coordinates and computed the ground truth angles using Equations (4.1), (4.2), and (4.3).

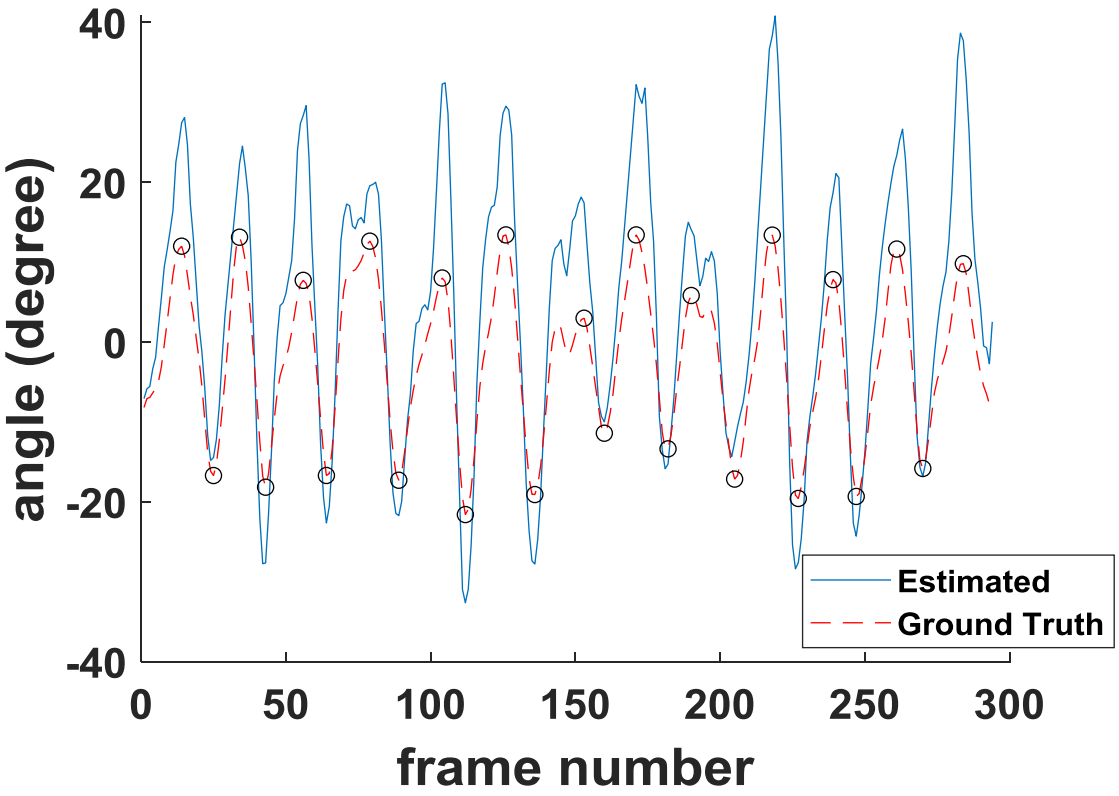


Figure 4-5 Angle Prediction by the Algorithm.

Although the hand flexion angles are estimated for each video frame, the hand state needs to be estimated only at *target frames* corresponding to the local maximum and the local minimum ground truth angles. These target frames where the extreme postures occurred were marked with circles in Figure 4-5. There were a total of 1464 target frames extracted from 61 laboratory videos.

The ground truth angles were obtained from MoCap measurements and the estimated angles were obtained from the videos using our computer vision algorithm.

4.2.2 Synchronization for Laboratory Videos

Because the initialization of the MoCap system and the video recorder were independent, there was a time offset between the recorded signals. This is shown in Figure 4-6 (a) where the solid line and the dashed line represent the MoCap (ground truth) and our algorithm estimated angles respectively. While these signals have the same frequency, the blue curve (estimated angle) lags behind the red dotted curve (ground truth angel) by about 4 cycles. To synchronize these two signals, we evaluated the cross-correlation of them and search for the position D of the global maximum of the cross-correlation coefficient. This is shown in Figure 4-6 (b). D is the estimated timing offset (in the unit of frames) between the ground truth and the estimated angles. The ground truth angle curve then will be shifted by D frames so that it is aligned temporally with the video observation, as shown in Figure 4-6 (c). The relationships between the two signals before and after shifting are shown in Figure 4-6 (d).

4.2.3 Industrial Field Videos

In addition to the videos recorded in the laboratory, we also used industry field videos recorded on-site in the State of Washington Department of Labor and Industries SHARP [93] project. Four kinds of tasks performed by different workers were video-recorded. The tasks included a laundry handler in a commercial laundry facility, a lumber handler in a sawmill, an assembler in an electronics plant, and a pharmacist in a large hospital pharmacy. These jobs had large variations in work posture, and each of them was performed in multiple locations in the same facility. Each job was recorded from two different angles using two synchronized camcorders. A time-sample observation method was used in which human analysts observed the hand postures at pre-selected

frames during a task. Custom data processing software was used for posture estimation by analysts to obtain a continuous angular scale. The software showed the preselected frame from two camera angles and asked the analyst to reproduce the wrist angle they see by clicking on the posture diagram.

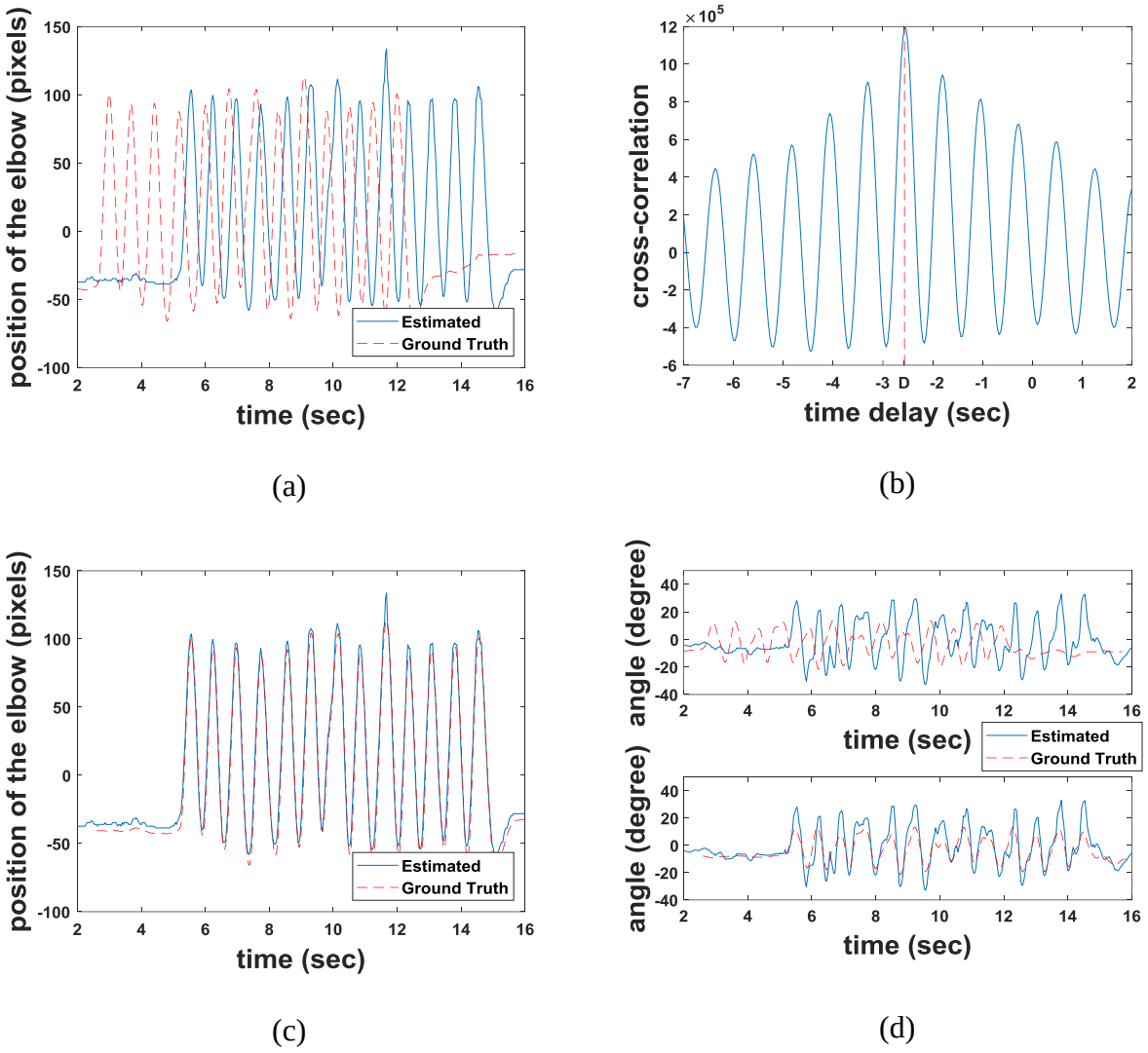


Figure 4-6 (a) Position-time curves of estimated and ground truth data; (b) Cross-Correlation of estimated and ground-truth signals; (c) Aligned position-time curves of estimated and ground truth data; (d) (Top) Angle-time curves of estimated and ground truth data; (Bottom) Two curves are aligned after shifting one signal by the time delay D.

We applied our algorithm to the same preselected frames and compared the posture estimate results. Due to the low resolution of the videos from the field studies, sometimes our algorithm

failed to detect a correct key-point location. We manually eliminated the frames in which the estimated joint locations were not detected. There were 262 side-view video frames selected. We considered the angles estimated observation by analysts as the ground truth and the angles estimated by our algorithm as the prediction.

4.3 Results

4.3.1 Angle Estimation Error

We defined an angle estimation error as the difference between the predicted angle and the ground truth (MoCap estimated, or human analysts estimated) angle. In Figure 4-7, a histogram of the angle estimation errors for the laboratory videos is plotted. The corresponding histogram of angle estimation errors for the industrial videos is plotted in Figure 4-8. The root-mean-square angle estimation error (RMSE) is the standard deviation of these distributions. In the case of laboratory videos, the RMS angle estimation error was 12.64 degrees. For the industrial videos, the RMS angle estimation error was 14.70 degrees.

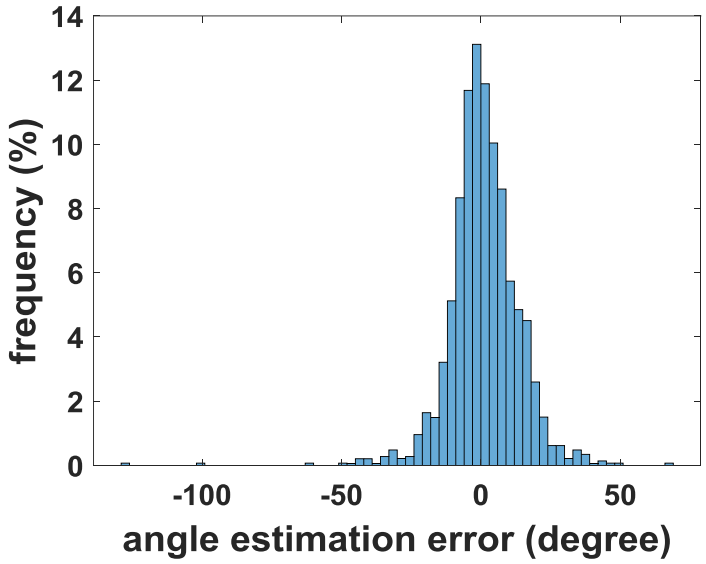


Figure 4-7 Histogram of the Angle Estimation Errors for Laboratory Videos.

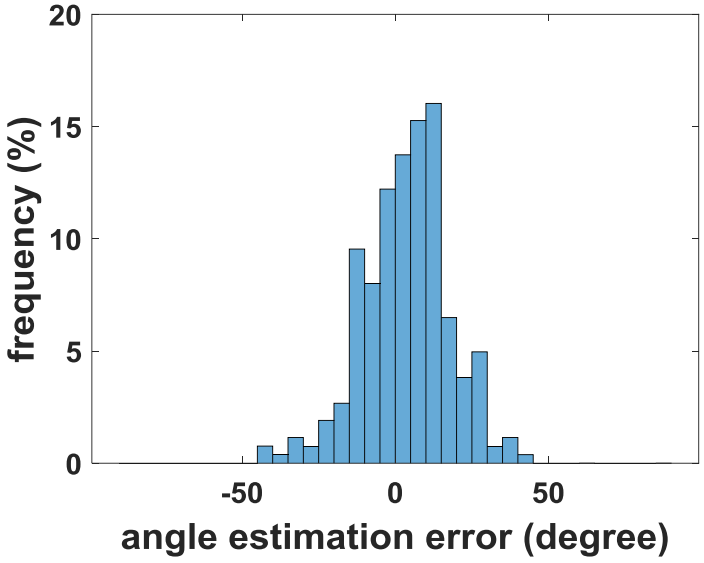


Figure 4-8 Histogram of the Angle Estimation Errors for Industrial Videos.

4.3.2 Averaged Per-Class Classification Rate

The confusion matrix for classifying the 1464 target frames from the 61 laboratory videos is shown in Table 4-1. The $(i,j)^{th}$ element of the confusion matrix, C_{ij} is the number of samples (video frames) that belong to the (ground truth) i^{th} class but were classified by our algorithm as the j^{th} class. Thus, the overall classification rate is calculated as:

$$r = \sum_{i=1}^{3} C_{ii} \Big/ \sum_{i=1}^{3} \sum_{j=1}^{3} C_{ij} = \sum_{i=1}^{3} C_{ii}/N \tag{4.4}$$

where N = 1464 is the total number of samples used. The entries in the parentheses below C_{ij} are C_{ij}/N (see Table 4-1).

Using the confusion matrix, one may also define a per-class classification accuracy

$$r_i = C_{ii} \Big/ \sum_{j=1}^{3} C_{ij} \qquad i = 1,2,3 \tag{4.5}$$

and averaged per-class classification rate

Table 4-1 Confusion Matrix for Laboratory Videos

Ground Truth				
	$\theta < -15^{\circ}$ Flexion	**131 (8.95%)**	**67 (4.58%)**	**0 (0.00%)**
	$15^{\circ} > \theta \geq -15^{\circ}$ Neutral	**105 (7.17%)**	**587 (40.10%)**	**102 (6.97%)**
	$\theta \geq 15^{\circ}$ Extension	**3 (0.20%)**	**105 (7.17%)**	**364 (24.86%)**
		$\theta < -15^{\circ}$ Flexion	$15^{\circ} > \theta \geq -15^{\circ}$ Neutral	$\theta \geq 15^{\circ}$ Extension
		Prediction		

Table 4-2 Confusion Matrix for Industrial Videos

Ground Truth				
	$\theta < -15^{\circ}$ Flexion	**27 (10.31%)**	**11 (4.20%)**	**1 (0.38%)**
	$15^{\circ} > \theta \geq -15^{\circ}$ Neutral	**12 (4.58%)**	**91 (34.73%)**	**33 (12.60%)**
	$\theta \geq 15^{\circ}$ Extension	**1 (0.38%)**	**6 (2.29%)**	**80 (30.53%)**
		$\theta < -15^{\circ}$ Flexion	$15^{\circ} > \theta \geq -15^{\circ}$ Neutral	$\theta \geq 15^{\circ}$ Extension
		Prediction		

$$r_{APC} = (r_1 + r_2 + r_3)/3 \tag{4.6}$$

r_{APC} tends to compensate for the imbalanced distributions of samples in different classes.

For laboratory videos corresponding to Table 4-1, the average per-class accuracy was 72.40%. For industrial videos, the corresponding confusion matrix is shown in Table 4-2. The average per-class accuracy was 76.03%.

4.3.3 Per-Class Sensitivity and Specificity

For each class (i = 1, 2, 3), we may convert the confusion matrix into a 2 × 2 matrix. For example, let i = 1, the confusion matrix in Table I can be converted into

$$\begin{bmatrix} TP_1 & FN_1 \\ FP_1 & TN_1 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} + C_{13} \\ C_{21} + C_{31} & C_{22} + C_{23} + C_{32} + C_{33} \end{bmatrix} = \begin{bmatrix} 131 & 67 \\ 108 & 1158 \end{bmatrix} \tag{4.7}$$

Then the sensitivity and specificity for class i may be defined as:

$$Sen_i = TP_i/TP_i + FN_i = C_{ii}/\sum_{j=1}^{3} C_{ij} \tag{4.8}$$

$$Spe_i = TN_i/TN_i + FP_i = \sum_{m\neq i}\sum_{n\neq i} C_{mn}/\sum_{m\neq i}\sum_{n=1}^{3} C_{mn} \tag{4.9}$$

The per-class sensitivity and specificity for the laboratory videos are list in Table 4-3.

Table 4-3 Sensitivity and Specificity for Laboratory Videos

Class	Flexion	Neutral	Extension
Sensitivity	66.16%	73.93%	77.12%
Specificity	91.47%	74.33%	89.72%

The per-class sensitivity and specificity for the SHARP industrial videos are list in Table 4-4.

Table 4-4 Sensitivity and Specificity for Industrial Videos

Class	Flexion	Neutral	Extension
Sensitivity	69.23%	66.91%	91.95%
Specificity	94.17%	86.51%	80.57%

4.3.4 Non-Sagittal Plane Tasks

The direction of the video viewpoint in this work was perpendicular to the sagittal plane of the subject. We also conducted experiments in which the camera viewpoint was not directly perpendicular to the plane of hand movement. This allowed us to investigate the difference between 2D and 3D angles estimated. We collected 1167 frames from 43 recorded videos for 15 participants (2 males and 13 females). As shown in Figure 4-4, two kinds of tasks, moving a tennis ball from location A to 3 and from location A to 4 (non-sagittal plane tasks), were studied. As discussed earlier, for the sagittal plane tasks, the RMS angle estimation error was 12.64 degrees and the average per-class accuracy was 72.60%. For the non-sagittal plane task, the RMS angle estimation error was 23.25 degrees and the average per-class accuracy was 74.26%. Although the non-sagittal plane task had a larger angle estimation error, the per-class accuracy was slightly better.

4.4 Discussion

In this work, we present a computer vision-based method, utilizing keypoint detection, to measure the joint angle of the hand posture. This measurement can be used as an objective indicator to assess the physical exposure of workers in a factory. Our posture estimation measurements were automatic, therefore more objective and consistent.

From Table 4-1, note that there are C_{21} = 105 (7.17%) neural wrist angles incorrectly classified by our algorithm for the flexion. As shown in Figure 4-5, the computer vision estimated angles seem to be larger than those estimated using MoCap. This discrepancy may be due to the misalignment of the joint positions estimated by OpenPose and by the MoCap sensors. Reducing these estimation discrepancies may in turn reduces the angle classification error. In Table 3-1, note that there are 33 cases (12.6% of the samples) of neutral wrist angles misclassified as extensions.

However, since the ground truth was obtained from human observers, we attribute this discrepancy to human subjective judgment variations.

Applying the proposed method in the real work environment might be promising. Not limited to the hand posture, our method can also be extended to the application of measuring neck, shoulder, and elbow postures. For example, we can use the three landmarks: ear, the center of shoulders, and the center of hips detected by Openpose to replace the hand, wrist, and elbow in this experiment. The new three landmarks viewed from the sagittal plane of the subject will form the angle of the neck.

Due to the limitation of single-view videos, the assumption of the side view is necessary. In future work, depth image videos might be applied to remove this restriction.

A study in [71] compared the results between the video analysis and electrogoniometer method. They classified the wrist flexion-extension posture into five categories: $\theta \geq 45^{\circ}$, $45^{\circ} > \theta \geq 15^{\circ}$, $15^{\circ} > \theta \geq -15^{\circ}$, $-15^{\circ} > \theta \geq -45^{\circ}$ and $\theta < -45^{\circ}$, and achieve 57% agreement. For comparison, we use the same five-class categories, and the resulting confusion matrix is shown in Table 4-5. The agreement between MoCap and the computer vision method is 70.42%, which is 13% higher than the agreement between the electrogoniometer and observation method. Since the electrogoniometer provided 3D angles ground truth, we also compare the angles computed by original 3D coordinates of MoCap with the 2D angles of the computer vision method, and the confusion matrix is shown in Table 4-6. The overall agreement is 56.97%, which is consistent with the results in [71]. With similar performance, our method saves much time and labor since it is automatic.

The effect of viewing angle on wrist posture estimation has been addressed in [95]. A scatter plot of the angle computed using the 3D coordinates of MoCap vs. the angle computed using the

transformed 2D image coordinates is represented in Figure 4-9. We observe that the 3D angle is larger than the 2D angle in most cases. Furthermore, many 3D angles appear to be 0 degrees in 2D images. This observation may be explained using sketches in Figure 4-10. Note that due to projection, $\angle \mathrm{EW_{2D}H}$ will be larger than the $\angle \mathrm{EW_{3D}H}$. The flexion angle defined in this work (c.f. Figure 4-1 and Figure 4-2) is the complementary angle of the $\angle \mathrm{EWH}$. Hence $\theta_{2D} \leq \theta_{3D}$. Although we assume the camera is photographing from the side view, this phenomenon is mainly due to the supination and pronation of the wrist. It is also a main source of error.

We conclude that using a computer vision-based method to estimate wrist flexion/extension provides a quick assessment and requires no equipment attached to the subject. This non-intrusive, automated observation method allows long-term, large-scale collection of exposure data in longitudinal studies in the future. Combined with health outcomes, such studies will provide data-centric validation of exposure models developed to prevent job-related hand injuries.

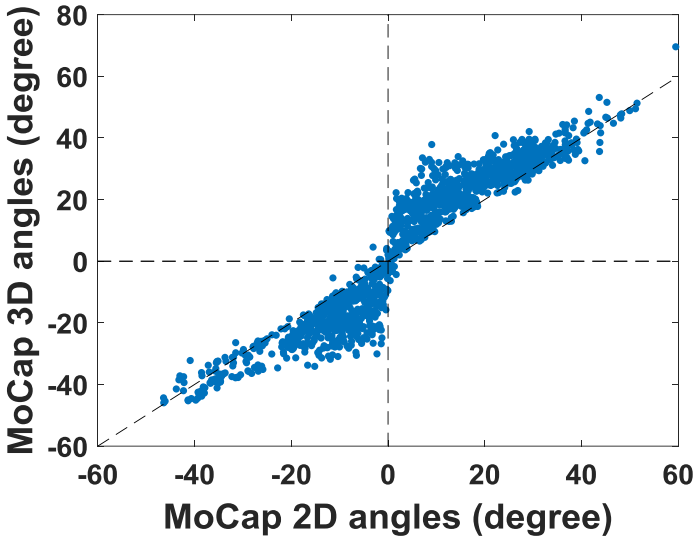


Figure 4-9 MoCap Estimated 3D Angles vs. 2D Angles.

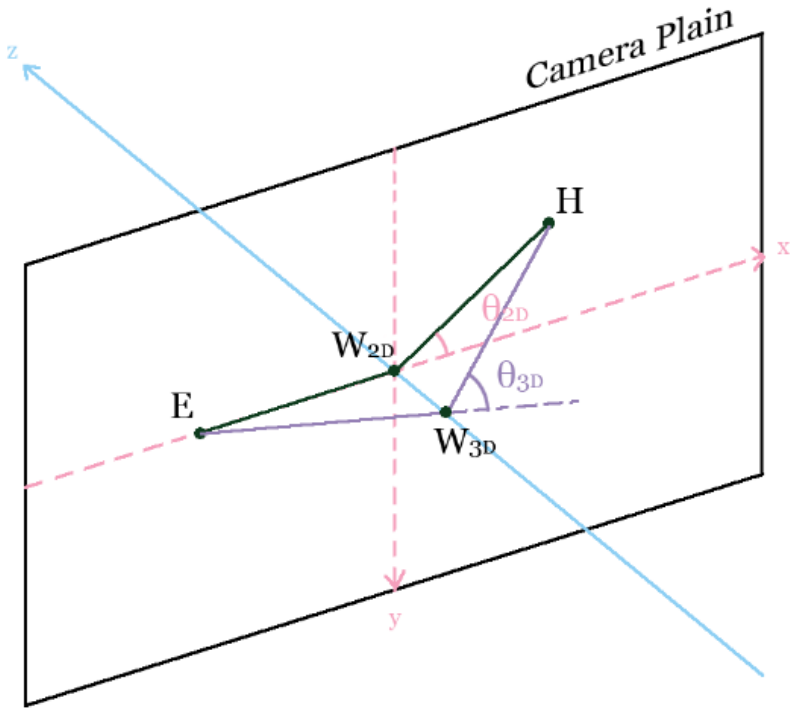


Figure 4-10 Comparison between the 3D and 2D angles. The 3D angle is larger than the 2D angle in this case.

Table 4-5 Confusion Matrix for Laboratory Videos with Five Categories

Ground Truth	Prediction: $\theta < -45^\circ$ Flexion	Prediction: $-15^\circ > \theta \geq -45^\circ$ Flexion	Prediction: $15^\circ > \theta \geq -15^\circ$ Neutral	Prediction: $45^\circ > \theta \geq 15^\circ$ Extension	Prediction: $\theta \geq 45^\circ$ Extension
$\theta < -45^\circ$ Flexion	**0 (0.00%)**	**3 (0.20%)**	**0 (0.00%)**	**0 (0.00%)**	**0 (0.00%)**
$-15^\circ > \theta \geq -45^\circ$ Flexion	**5 (0.34%)**	**123 (8.40%)**	**67 (4.58%)**	**0 (0.00%)**	**0 (0.00%)**
$15^\circ > \theta \geq -15^\circ$ Neutral	**3 (0.20%)**	**102 (6.97%)**	**587 (40.10%)**	**100 (6.83%)**	**2 (0.14%)**
$45^\circ > \theta \geq 15^\circ$ Extension	**1 (0.07%)**	**2 (0.14%)**	**104 (7.10%)**	**317 (21.65%)**	**37 (2.53%)**
$\theta \geq 45^\circ$ Extension	**0 (0.00%)**	**0 (0.00%)**	**1 (0.07%)**	**6 (0.41%)**	**4 (0.27%)**

Table 4-6 Confusion Matrix for Laboratory Videos with Five Categories and 3D MoCap Angles

Ground Truth	Prediction: $\theta < -45^\circ$ Flexion	Prediction: $-15^\circ > \theta \geq -45^\circ$ Flexion	Prediction: $15^\circ > \theta \geq -15^\circ$ Neutral	Prediction: $45^\circ > \theta \geq 15^\circ$ Extension	Prediction: $\theta \geq 45^\circ$ Extension
$\theta < -45^\circ$ Flexion	**2 (0.14%)**	**3 (0.20%)**	**0 (0.00%)**	**0 (0.00%)**	**0 (0.00%)**
$-15^\circ > \theta \geq -45^\circ$ Flexion	**5 (0.34%)**	**191 (13.05%)**	**217 (14.82%)**	**7 (0.48%)**	**0 (0.00%)**
$15^\circ > \theta \geq -15^\circ$ Neutral	**1 (0.07%)**	**30 (2.05%)**	**272 (18.58%)**	**40 (2.73%)**	**2 (0.14%)**
$45^\circ > \theta \geq 15^\circ$ Extension	**1 (0.07%)**	**6 (0.41%)**	**268 (18.31%)**	**366 (25.00%)**	**38 (2.60%)**
$\theta \geq 45^\circ$ Extension	**0 (0.00%)**	**0 (0.00%)**	**2 (0.14%)**	**10 (0.68%)**	**3 (0.20%)**

Chapter 5.

Conclusion

5.1 Summary

In this thesis, we focused on developing methods for repetitive hand motion analysis to automatically measure duty cycle (DC) and wrist flexion and extension from a 2D video for occupational health and safety research. Our goal was to reduce the usage of exposure assessment methods that require intrusive instrumentation or time-consuming human analyst training and interpretation.

In the application of measuring DC, the goal was to predict the hand activity level in the repetitive hand motion tasks. We proposed a machine learning algorithm using hand motion kinematics extracted from videos as the features and trained by the k-nearest neighborhood model to predict DC. Then, the corresponding hand activity level is estimated by an equation, which depends on the root mean square of the hand speed and DC.

In the application of wrist flexion/extension classification, the goal was to classify the wrist posture into one of the three categories: flexion, extension, and neutral. We proposed a computer vision method using the skeletal joints detected by *Openpose* to estimate the wrist flexion/extension angle between the hand and forearm, and classify the wrist posture as flexion, extension, or neutral based on the estimated angle.

5.2 Future Research Directions

We conclude this thesis by suggesting some future research directions.

- **Training a general model for predicting DC without the first cycle information**. Although the current proposed feature vector training method achieved promising

performance in predicting DC for the environment-controlled laboratory videos, it is hard to predict DC for the real-world factory videos accurately. The first cycle training method can be applied to different kinds of tasks but needs to be re-trained when we encounter a new task. To train a general model for predicting DC of different kinds of tasks, using the hand motion kinematics as the features solely might not be enough. Deep learning methods, such as *Convolutional Neural Network* (CNN), might be introduced to solve this problem.

- **Exploring more state-of-art human skeletal joints estimators**. The current approach for estimating the wrist flexion/extension angle uses *Openpose* to predict the 2D coordinates of the skeletal joints. Our application is limited to the images from the side-view of the target. If we can extend the 2D skeletal joints estimator to estimate 3D skeletal joints by combining different state-of-art techniques such as *Structure from Motion* (SfM), we might get rid of the side-view limitation and improve the accuracy of the wrist angle estimator.

Acknowledgments

This work was supported in part by a grant from the National Institute for Occupational Safety and Health (NIOSH/CDC), R01OH011024 (Radwin). The views expressed do not necessarily reflect the official policies of the Department of Health and Human Services.

References

[1] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288-303, 2010.

[2] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint Segmentation and Classification of Human Actions in Video," in *Int'l Conf. Computer Vision and Pattern Recognition (CVPR'11)* Colorado Springs, CO, 2011, pp. 3265-3272.

[3] N. Ikizler-Cinbis, and S. Sclaroff, "Object, Scene and Actions: Combining Multiple Features for Human Action Recognition," in *European Conf. Computer Vision (ECCV'10)*, K. Daniilidis, P. Maragos, and N. Paragios, Ed.: Springer Berlin Heidelberg, 2010, pp. 494–507.

[4] L. Wang, and D. Suter, "Learning and Matching of Dynamic Shape Manifolds for Human Action Recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646-1661, 2007.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving Human Action Recognition Using Fusion of Depth Camera and Inertial Sensors," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 1, pp. 51-61, 2015.

[6] A. Veenendaal, et al, "Dynamic Probabilistic Network Based Human Action Recognition," Xiv preprint arXiv:1610.06395 2016.

[7] M. Devanne, et al, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1340-1352, 2015.

[8] N. H. Dardas, and N. D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques," *IEEE Trans. Instr. Measurements*, vol. 60, no. 11, pp. 3592-3607, 2011.

[9] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision Based Hand Pose Estimation: A Review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52-73, 2007.

[10] T. S. Huang, Y. Wu, and J. Lin, "3d Model-Based Visual Hand Tracking," in *International Conference on Multimedia and Expo (ICME)*, 2002, pp. 902-905.

[11] M. Itoh, M. Ozeki, Y. Nakamura, and Y. Ohta, "Simple and Robust Tracking of Hands and Objects for Video- Based Mu1timedia Production," in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2003, pp. 252-257.

[12] M. Kolsch, and M. Turk, "Hand Tracking with Flocks of Features," in *Int'l Conf. Computer Vision and Pattern Recognition (CVPR)* San Diego, CA, 2005, p. 158.

[13] E. Lin, et al, "Hand Tracking Using Spatial Gesture Modeling and Visual Feedback for a Virtual DJ System," in *International Conference on Multimodal Interfaces (ICMI'02)* Pittsburgh, PA, 2002, pp.197-202.

[14] L. Liu, J. Xing, H. Ai, and X. Ruan, "Hand Posture Recognition Using Finger Geometric Feature," in *International Conference on Pattern Recognition (ICPR 2012)* Tsukuba, Japan, 2012.

[15] G. R. S. Murthy, and R. S. Jadon, "A Review of Vision Based Hand Gestures Recognition," *Int'l J. Information Technology and Knowledge Management,* vol. 2, no. 2, pp. 405-410, 2009.

[16] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient Model-Based 3d Tracking of Hand Articulations Using Kinect," in *Proc. British Machine Vision Conference*, Dundee, UK, 2011, pp. 1-11.

[17] S. Park, et al, "3d Hand Tracking Using Kalman Filter in Depth Space," *EURASIP Journal on Advances in Signal Processing,* vol. 2012, no. 1, pp. 1-18, 2012.

[18] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE Trans. PAMI,* vol. 19, no. 7, pp. 677-695, 1997.

[19] R. P. K Poudel, "3d Hand Tracking." vol. Doctor of Philosophy Poole, UK: Bournemouth University, 2014.

[20] S. M. M. Roomi, R. J. Priya and H. Jayalakshmi, "Hand Gesture Recognition for Human-Computer Interaction," *Journal of Computer Science,* vol. 6, no. 9, pp. 1002-1007, 2010.

[21] T. Sharp, et al, "Accurate, Robust, and Flexible Real-Time Hand Tracking," in *ACM Conference on Human Factors in Computing Systems (CHI'15)*, Seoul, Korea, 2015, pp. 3633-3642.

[22] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Real-Time Head and Hand Tracking Based on 2.5d Data," *IEEE Trans. Multimedia,* vol. 14, no. 3, pp. 575-585, 2012.

[23] L. Sun, and G. Liu, "Hand Tracking Based on the Combination of 2d and 3d Model in Gaze-Directed Video," in *Proc. Int'l Conf. Multimedia & Expo*, Barcelona, Spain, 2011, pp. 1-6.

[24] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, and P. Mark, "Robust Articulated Icp for Real Time Hand Tracking," *Computer Graphics Forum,* vol. 34, no. 5, pp. 110-114, 2015.

[25] J. P. Wachs, M. Kolsch, H. Stern, and Y. Edan, "Vision-Based Hand Gesture Applications," *Communications of the ACM,* vol. 54, no. 2, pp. 60-71, 2011.

[26] Y. Wu, J. Lin, and T. S. Huang, "Capturing Natural Hand Articulation Computer Vision," in *Int'l Conf. Computer Vision (ICCV'01)*, 2001, vol. 2, pp. 426-432.

[27] Y. Wu, J. Lin, and T. S. Huang, "Analyzing and Capturing Articulated Hand Motion in Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1910-1922, 2005.

[28] C.-H. Chen, Y. H. Hu, T. Y. Yen and R. G. Radwin, "Automated Video Exposure Assessment of Repetitive Hand Activity Level for a Load Transfer Task," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 55, no. 2, pp. 298-308, 2013.

[29] C.-H. Chen, Y. H. Hu, and R. G. Radwin, "A Motion Tracking System for Hand Activity Assessment," in *IEEE China Summit & Int. Conf. Signal and Information Processing* Xian, China: IEEE, 2014, pp. 320-324.

[30] C. H. Chen, "Vision-Based Human Hand Activity Monitoring and Assessment," in *Electrical and Computer Engineering*. vol. Doctor of Philosophy Madison, WI: University of Wisconsin - Madison, 2014.

[31] American Conference of Governmental Industrial Hygienists, "Hand Activity Level," in *Tlvs and Beis Based on the Documentation of the Threshold Limit Values for Chemical*

Substances and Physical Agents & Bioloegical Exposure Indices: American Conference of Governmental Industrial Hygienists, 2009, pp. 196-198.

[32] P. Drinkaus, R. F. Sesek, D. S. Bloswick, C. Mann, and T. Bernard, "The Hand Activity Level: Using Task Level Outputs to Evaluate Job Risk," in *National Occupational Research Agenda (NORA) Young/New Investigators Symposium* Salt Lake City, UT, 2003.

[33] P. Drinkaus, et al, "Job Level Risk Assessment Using Task Level Acgih Hand Activity Level Tlv Scores: A Pilot Study," *International Journal of Occupational Safety and Ergonomics,* vol. 11, no. 3, pp. 263-281, 2005.

[34] P. Spielholz, et al, "Reliability and Validity Assessment of the Hand Activity Level Threshold Limit Value and Strain Index Using Expert Ratings of Mono-Task Jobs," *Journal of occupational and environmental hygiene,* vol. 5, no. 4, pp. 250-257, 2008.

[35] S. Wurzelbacher, et al, "A Comparison of Assessment Methods of Hand Activity and Force for Use in Calculating the Acgih Hand Activity Level (Hal) Tlv," *Journal of occupational and environmental hygiene,* vol. 7, no. 7, pp. 407-416, 2010.

[36] J. K. Aggarwal, and M. S. Ryoo, "Human Activity Analysis: A Review," *ACM Computing Surveys (CSUR),* vol. 43, no. 3, pp. 1-43, 2011.

[37] M. B. Kaaniche and F. Brémond, "Gesture Recognition by Learning Local Motion Signatures," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2745-2752.

[38] D. Luo and J. Ohya, "Study on Human Gesture Recognition from Moving Camera Images," in *Proc. IEEE Conf. Multimedia and Expo (ICME)* Singapore: IEEE, 2010, pp. 274 - 279

[39] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 37, no. 3, pp. 311-324, 2007.

[40] Y. Wu, and T. S. Huang, "Vision-Based Gesture Recognition: A Review," *International Gesture Workshop,* pp. 103-115, 1999.

[41] Q.-Y. Zhang, M.-Y. Zhang, and J.-Q. Hu, "A Method of Hand Gesture Segmentation and Tracking with Appearance Based on Probability Model," in *Int'l Symp. Intelligent information technology application (IITA'08),* 2008, vol. 1, pp. 380-383.

[42] A. Franzblau, et al, "A Cross-Sectional Assessment of the Acgih Tlv for Hand Activity Level," *Journal of occupational rehabilitation,* vol. 15, no. 1, pp. 57-67, 2005.

[43] G. M. Hägg, and E. Milerad, "Forearm Extensor and Flexor Muscle Exertion During Simulated Gripping Work—an Electromyographic Study," *Clinical Biomechanics,* vol. 12, no. 1, pp. 39-43, 1997.

[44] W. Rohmert, "Problems in Determining Rest Allowances: Part 1: Use of Modern Methods to Evaluate Stress and Strain in Static Muscular Work," *Applied Ergonomics,* vol. 4, no. 2, pp. 91-95, 1973.

[45] W. Rohmert, "Problems of Determination of Rest Allowances Part 2: Determining Rest Allowances in Different Human Tasks," *Applied Ergonomics,* vol. 4, no. 3, pp. 158-162, 1973.

[46] S. Byström, C. Hall, T. Welander, and Å. Kilbom, "Clinical Disorders and Pressure-Pain Threshold of the Forearm and Hand among Automobile Assembly Line Workers," *The Journal of Hand Surgery: British & European,* vol. 20, no. 6, pp. 782-790, 1995.

[47] D. D. Wood, D. L. Fisher and R. O. Andres, "Minimizing Fatigue During Repetitive Jobs: Optimal Work-Rest Schedules," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 39, no. 1, pp. 83-101, 1997.

[48] J. R. Potvin, "Predicting Maximum Acceptable Efforts for Repetitive Tasks an Equation Based on Duty Cycle," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 54, no. 2, pp. 175-188, 2012.

[49] S. J. Moore and A. Garg, "The Strain Index: A Proposed Method to Analyze Jobs for Risk of Distal Upper Extremity Disorders," *American Industrial Hygiene Association Journal,* vol. 56, no. 5, pp. 443-458, 1995.

[50] C. Harris-Adamson, E. A. Eisen, J. Kapellusch, A. Garg, K. T. Hegmann, M.S. Thiese, ... & B. Silverstein, "Biomechanical Risk Factors for Carpal Tunnel Syndrome: A Pooled Study of 2474 Workers," *Occupational and Environmental Medicine,* vol. 72, no. 1, pp. 33-41, 2015.

[51] W. A. Latko, T. J. Armstrong, J. A. Foulke, G. D. Herrin, R. A. Rabourn, and S. S. Ulin, "Development and Evaluation of an Observational Method for Assessing Repetition in Hand Tasks," *American Industrial Hygiene Association Journal,* vol. 58, no. 4, pp. 278-285, 1997.

[52] R. G. Radwin, D. P. Azari, M. J. Lindstrom, S. S. Ulin, T. J. Armstrong and D. Rempel, "A Frequency-Duty Cycle Equation for the Acgih Hand Activity Level," *Ergonomics,* vol. 58, no. 2, pp. 173-183, 2015.

[53] Z. J. Fan, B. A. Silverstein, S. Bao, D. K. Bonauto, N. L. Howard and C. K. Smith, "The Association between Combination of Hand Force and Forearm Posture and Incidence of Lateral Epicondylitis in a Working Population," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 56, no. 1, pp. 151-165, 2014.

[54] S. Bao, N. Howard, P. Spielholz, & B. Silverstein, "Quantifying Repetitive Hand Activity for Epidemiological Research on Musculoskeletal Disorders–Part Ii: Comparison of Different Methods of Measuring Force Level and Repetitiveness," *Ergonomics,* vol. 49, no. 4, pp. 381-392, 2006.

[55] A. Garg, and J. M. Kapellusch, "Job Analysis Techniques for Distal Upper Extremity Disorders," *Reviews of Human Factors and Ergonomics,* vol. 7, no. 1, pp. 149-196, 2011.

[56] J. M. Kapellusch, A. Garg, S. S. Bao, B. A. Silverstein, S. E. Burt, A. M. Dale, ... & D. M. Rempel, "Pooling Job Physical Exposure Data from Multiple Independent Studies in a Consortium Study of Carpal Tunnel Syndrome," *Ergonomics,* vol. 56, no. 6, pp. 1021-1037, 2013.

[57] R. Wells, S. E. Mathiassen, L. Medbo & J. Winkel, "Time—a Key Issue for Musculoskeletal Health and Manufacturing," *Applied Ergonomics,* vol. 38, no. 6, pp. 733-744, 2007.

[58] T. Y. Yen, and R. G. Radwin, "Automated Job Analysis Using Upper Extremity Biomechanical Data and Template Matching," *International Journal of Industrial Ergonomics,* vol. 25, no. 1, pp. 19-28, 1999.

[59] O. Akkas, D. P. Azari, C-H. Chen, Y. H. Hu, S. S. Ulin, T. J. Armstrong, D. Rempel and R. G. Radwin, "A Hand Speed and Duty Cycle Equation for Estimating the ACGIH Hand Activity Level Rating," *Ergonomics,* vol. 58, no. 2, pp. 184-194, 2015.

[60] T. Y. Yen, and R. G. Radwin, "A Video-Based System for Acquiring Biomechanical Data Synchronized with Arbitrary Events and Activities," *IEEE Transactions on Biomedical Engineering,* vol. 42, no. 9, pp. 944-948, 1995.

[61] C. Harris, E. A. Eisen, R. Goldberg, N. Krause, and D. Rempel, "Workplace and Individual Factors in Wrist Tendinosis among Blue-Collar Workers - the San Francisco Study," *Scand J Work Environ Health,* vol. 37, no. 2, pp. 85-98, 2011.

[62] R. G. Radwin, and M.-L. Lin, "An Analytical Method for Characterizing Repetitive Motion and Postural Stress Using Spectral Analysis," *Ergonomics*, vol. 36, no. 4, pp. 379-389, 1993.

[63] C.-H. Chen, A. Akkas, Y. H. Hu, and R. G. Radwin, "The Accuracy of Conventional 2d Video for Quantifying Upper Limb Kinematics in Repetitive Motion Occupational Tasks," *Ergonomics*, vol. 58, no. 12, pp. 2057-2066, 2015.

[64] J. M. Rubin, and W. A. Richards, "Boundaries of Visual Motion," MIT, Cambridge, MA, Artificial Intelligence Lab Memos 835, 1985.

[65] C. Rao, A. Yilmaz, & M. Shah, "View-Invariant Representation and Recognition of Actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 203-226, 2002.

[66] C. Couvreur, and Y. Bresler, "On the Optimality of the Backward Greedy Algorithm for the Subset Selection Problem," *SIAM Journal on Matrix Analysis and Its Applications*, vol. 21, no. 3, pp. 797-808, 1999.

[67] H. Bay, T. Tuytelaars and L. Van Gool, "Surf: Speeded up Robust Features," in *European conference on computer vision*, 2006, pp. 404-417.

[68] M. A. Fischler, & R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, 1981.

[69] J. A. Barondess, "Epidemiologic evidence," in *Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities*, Washington, DC, USA: National Academy of Sciences, 2001, ch. 4, pp. 85-183.

[70] WSIB, "By the Numbers: 2017 WSIB (Ontario) Statistical Report," 2017.

[71] C. D. McKinnon, S. Ehmke, A. M. Kociolek, J. P. Callaghan, and P. J. Keir, "Wrist Posture Estimation Differences and Reliability Between Video Analysis and Electrogoniometer Methods," *Human Factors*, p. 0018720820923839, 2020.

[72] B. Juul-Kristensen, N. Fallentin, and C. Ekdahl, "Criteria for classification of posture in repetitive work by observation methods: A review," *International Journal of Industrial Ergonomics*, vol. 19, no. 5, pp. 397– 411, 1997.

[73] A. M. Dale, C. Harris-Adamson, D. Rempel, F. Gerr, K. Hegmann, B. Silverstein, S. Burt, A. Garg, J. Kapellusch, and L. Merlino, "Prevalence and incidence of carpal tunnel syndrome in US working populations: pooled analysis of six prospective studies," *Scandinavian journal of work, environment & health*, vol. 39, no. 5, p. 495, 2013.

[74] J. M. Kapellusch, F. E. Gerr, E. J. Malloy, A. Garg, C. Harris-Adamson, S. S. Bao, S. E. Burt, A. M. Dale, E. A. Eisen, and B. A. Evanoff, "Exposure-response relationships for the ACGIH threshold limit value for hand-activity level: results from a pooled data study of carpal tunnel syndrome," *Scandinavian journal of work, environment & health*, vol. 40, no. 6, p. 610, 2014.

[75] B. A. Silverstein, L. J. Fine, and T. J. Armstrong, "Occupational factors and carpal tunnel syndrome," *American journal of industrial medicine*, vol. 11, no. 3, pp. 343–358, 1987.

[76] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of electromyography and kinesiology*, vol. 14, no. 1, pp. 13–23, 2004.

[77] G. C. David, "Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders," *Occupational medicine*, vol. 55, no. 3, pp. 190–199, 2005.

[78] G. David, V. Woods, G. Li, and P. Buckle, “The development of the Quick Exposure Check (QEC) for assessing exposure to risk factors for work-related musculoskeletal disorders,” *Applied ergonomics*, vol. 39, no. 1, pp. 57–69, 2008.

[79] G. Li and P. Buckle, “Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods,” *Ergonomics*, vol. 42, no. 5, pp. 674–695, 1999.

[80] R. J. Shephard, “Limits to the measurement of habitual physical activity by questionnaires,” *British journal of sports medicine*, vol. 37, no. 3, pp. 197–206, 2003.

[81] S. Bao, N. Howard, P. Spielholz, B. Silverstein, and N. Polissar, “Interrater reliability of posture observations,” *Human factors*, vol. 51, no. 3, pp. 292–309, 2009.

[82] W. Lee, J.-H. Lin, and S. Bao, "Inter-rater reliability of an inertial measurement unit sensor-based posture-matching method: A pilot study," *International Journal of Industrial Ergonomics*, vol. 80, p. 103025, 2020.

[83] B. Juul-Kristensen, G. Å. Hansson, N. Fallentin, J. H. Andersen, and C. Ekdahl, “Assessment of work postures and movements using a video- based observation method and direct technical measurements,” *Applied ergonomics*, vol. 32, no. 5, pp. 517–524, 2001.

[84] A. Hua *et al.*, “Evaluation of Machine Learning Models for Classifying Upper Extremity Exercises Using Inertial Measurement Unit-Based Kinematic Data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2452–2460, 2020.

[85] W. Yang, D. Yang, Y. Liu, and H. Liu, “Decoding Simultaneous Multi-DOF Wrist Movements from Raw EMG Signals Using a Convolutional Neural Network,” *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 5, pp. 411–420, 2019.

[86] M. C. Schall Jr, R. F. Sesek, and L. A. Cavuoto, "Barriers to the adoption of wearable sensors in the workplace: A survey of occupational safety and health professionals," *Human Factors*, vol. 60, no. 3, pp. 351-362, 2018.

[87] J. A. Diego-Mas and J. Alcaide-Marzal, “Using Kinect™ sensor in observational methods for assessing postures at work,” *Applied ergonomics*, vol. 45, no. 4, pp. 976–985, 2014.

[88] P. Plantard, H. P. Shum, A.-S. L. Pierres, and F. Multon, “Validation of an ergonomic assessment method using Kinect data in real workplace conditions,” *Applied ergonomics*, vol. 65, pp. 562–569, 2017.

[89] Z. Li, R. Zhang, C.-H. Lee, and Y.-C. Lee, "An evaluation of posture recognition based on intelligent rapid entire body assessment system for determining musculoskeletal disorders," *Sensors*, vol. 20, no. 16, p. 4414, 2020.

[90] A. M. Kociolek and P. J. Keir, “Reliability of distal upper extremity posture matching using slow-motion and frame-by-frame video methods,” *Human factors*, vol. 52, no. 3, pp. 441-455, 2010.

[91] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[92] J. Schmidt, D. R. Berg, H.-L. Ploeg, L. Ploeg, J. Schmidt, and D. R. Berg, “Precision, repeatability and accuracy of Optotrak ® optical motion tracking systems,” *Int. J. Experimental and Computational Biomechanics*, vol. 1, no. 1, pp. 114–127, 2009.

[93] S. Bao, B. Silverstein, N. Howard, and P. Spielholz, “The Washington State SHARP approach to exposure assessment,” in *Fundamentals and assessment tools for occupational ergonomics*, vol. 44, no. 1, pp. 22–44, 2006.

[94] O. Akkas, C.-H. Lee, Y. H. Hu, T. Y. Yen, and R. G. Radwin, "Measuring elemental time and duty cycle using automated video processing," *Ergonomics*, vol. 59, no. 11, pp. 1514-1525, 2016.

[95] M. H. Lau and T. J. Armstrong, "The effect of viewing angle on wrist posture estimation from photographic images using novice raters." *Applied ergonomics*, vol. 42, no. 5, pp. 634-643, 2011.

[96] M. Oudah, A. Al-Naji, and J. Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[97] D. Jiang *et al.*, "Gesture recognition based on binocular vision," *Cluster Computing*, vol. 22, pp. 13261–13271, 2019.

[98] A. S. Al-Shamayleh, R. Ahmad, M. A. M. Abushariah, K. A. Alam, and N. Jomhari, "A systematic literature review on vision based gesture recognition techniques," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28121–28184, 2018.

[99] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55-79, 2005.

[100] M. Eichner and V. Ferrari, "Better appearance models for pictorial structures," in *Bmvc*, vol. 2, p. 5. 2009.

[101] A. Toshev and G. Christian Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653-1660, 2014.

[102] Z. J. Fan *et al.*, "Quantitative exposure-response relations between physical workload and prevalence of lateral epicondylitis in a working population," *American Journal of Industrial Medicine*, vol. 52, no. 6, pp. 479–490, 2009.

[103] B. A. Silverstein *et al.*, "The natural course of carpal tunnel syndrome in a working population," *Scandinavian Journal of Work Environment and Health*, vol. 36, no. 5, pp. 384–393, 2010.

[104] A. Patrizi, E. Pennestrì, and P. P. Valentini, "Comparison between low-cost marker-less and high-end marker-based motion capture systems for the computer-aided assessment of working ergonomics," *Ergonomics*, vol. 59, no. 1, pp. 155–162, 2016.