BIOMECHANICAL MARKERLESS MOTION CLASSIFICATION BASED ON STICK MODEL DEVELOPMENT FOR SHOP FLOOR OPERATOR

LIEW YU LIANG

SCHOOL OF MECHANICAL ENGINEERING
UNIVERSITI SAINS MALAYSIA
2021

BIOMECHANICAL MARKERLESS MOTION CLASSIFICATION BASED ON STICK MODEL DEVELOPMENT FOR SHOP FLOOR OPERATOR

by

LIEW YU LIANG

(Matrix No.: 138288)

Supervisor:
Associate Professor Ir. Dr Chin Jeng Feng

This dissertation is submitted to

UNIVERSITI SAINS MALAYSIA

As partial fulfilment of requirement for the degree of

BACHELOR OF ENGINEERING (HONS.)

(MANUFACTURING ENGINEERING WITH MANAGEMENT)

School of Mechanical Engineering
Universiti Sains Malaysia

July 2021

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently in candidature for any degree.

Signed …YuLiang…… (LIEW YU LIANG)
Date…23 July 2021…….

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by giving explicit references. Bibliography/references are appended.

Signed …YuLiang…… (LIEW YU LIANG)
Date…23 July 2021…….

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed …YuLiang…… (LIEW YU LIANG)
Date…23 July 2021…….

ACKNOWLEDGEMENT

First, I would like to express my appreciation to the School of Mechanical Engineering in Universiti Sains Malaysia (USM) for providing me with an opportunity to work on the research project and provide online seminar support during the research. These assistances helped me cope with the research writing during such a challenging period in the pandemics. An appreciation was sent to the course coordinator, Dr Muhammad Fauzinizam Bin Razali, for arranging the thesis schedule and title distribution.

My deepest gratitude was expressed to my supervisor, Associate Professor Ir. Dr Chin Jeng Feng, for the guidance and comments of the research work throughout the study. The advice given allowed me to try the solutions and resolve the difficulties continuously. Without my dedicated supervisor, this research study would not be accomplished.

Not to forget all my friends who volunteered to participate in the motion video capturing. They provided the source of video subjects in the study. Their participation was so important to me as this research study required different persons performing the movement to collect as much data as possible. I would also like to thank my friends for the encouragement and exchange of knowledge in the programming field.

Most importantly, I stated my gratitude to my parents and family members for their constant warmth and unconditional love. Their presence always told me to stay strong through the low tides of the research without easily giving up.

Finally, I would like to thank everyone who has offered me any helps during my research study. Even a little assistance could significantly tune down the difficulty to finish this research study.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF APPENDIX TABLE

Page

LIST OF FIGURES

Page

LIST OF APPENDIX FIGURES

LIST OF SYMBOLS

Symbol	Description	Unit
a_{x_n}	Cumulative acceleration in the x-direction of *n*th body part	pixels/second^2
a_{y_n}	Cumulative acceleration in the y-direction of *n*th body part	pixels/second^2
i	Instance number of motion data	
j	Logarithm value for maximum of variable data at *n*th attribute	
k	Number of nearest neighbours included in the majority voting process	
m	Total number of frames divided by 30	
$\max(v_n)$	maximum value of variable data in the *n*th attribute	
$\min(v_n)$	minimum value of variable data in the *n*th attribute	
max_{new}	new minimum value after normalization, usually -1 or 0	
min_{new}	new maximum value after normalization, usually 1	
n	Body part indication number	
t	Time point in the video	seconds
u_{x_n}	Initial velocity of *n*th body part at x-axis	pixels/second
u_{y_n}	Initial velocity of *n*th body part at y-axis	pixels/second
$v_{i,n}$	Original variable data for *n*th attribute at *i*th instance	
$v'_{i,n}$	New normalized variable data for *n*th attribute at *i*th instance	
v_n	Variable data at *n*th attribute	
v_{x_n}	Cumulative velocity in the x-direction of *n*th body part	pixels/second
v_{y_n}	Cumulative velocity in the y-direction of *n*th body part	pixels/second
x_{1_n}	x-axis coordinate at the first instance for body joint *n*	
x_{i_n}	x-axis coordinate at i^{th} instance for body joint *n*	
x_{i-1_n}	x-axis coordinate at the previous instance for body joint *n*	
y_{o_n}	y-axis coordinate at the start for body joint *n*	

y_{1_n}	y-axis coordinate at the first instance for body joint *n*
y_{i_n}	y-axis coordinate at i^{th} instance for body joint *n*
y_{i-1_n}	y-axis coordinate at the previous instance for body joint *n*
μ_n	mean of all data in the *n*th attribute
σ_n	standard deviation of all data in the *n*th attribute
$\Sigma_{i=1}^{m}$	Summation of the values starting from the first instance until reaching the total number of frames divided by 30

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AP	Average precision
3DSuit	Three-dimensional Suit
CCMEI	Contour coding of motion energy image
CNN	Convolutional neural network
COCO	Common Objects in Context
COCO-WholeBody	Common Objects in Context with whole body information
CSV	Comma Separated Values
DLT	Direct linear transformation
DSLR	Digital single-lens reflex
DSN	Decimal scaling normalization
DTW	Dynamic time warping
FGME	Fine-grained motion estimation
fps	Frames per second
GB	Gigabytes
GPU	Graphical processing unit
HOG	Histogram of Oriented Gradient
IMUs	Inertial measurement units
KDD	Knowledge discovery in database
KNN	K-nearest neighbours
LLGMN	Log-linearised Gaussian mixture network
MDC	Minimum distance classifiers
MMN	Min-max normalization
MPII	Max Planck Institute for Informatics
PAF	Part Affinity Field
PCKh	Head-normalised probability of correct keypoint
OKS	Object keypoint similarity
OneR	One Rule
OpenCV	Open-Source Computer Vision Library

OpenPose	Real-time multiple-person detection library
RAM	Random-access memory
RGB	Red, Green, Blue colour model
RMPE	Regional multi-person pose estimation
RNN	Recurrent neural network
sEMG	Surface electromyography
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support vector machine
Sklearn	Scikit-learn
TPU	Tensor processing unit
USM	Universiti Sains Malaysia
WEKA	Waikato Environment for Knowledge Analysis
ZeroR	Zero Rule
ZSN	Z-score normalization

ABSTRAK

Sistem klasifikasi pergerakan manusia menandakan teknologi industri era baru ini untuk memantau prestasi tugasan dan mengesahkan kualiti proses manual dengan menggunakan cara automasi. Namun demikian, kajian pada masa ini banyak bertumpu kepada tangkapan gerakan dengan penanda yang memerlukan peralatan mahal dan pemasangan sensor di bahagian badan subjek yang tertentu. Model klasifikasi pergerakan tanpa penanda masih berkurang perkembangan dalam bidang pembuatan. Oleh itu, penyelidikan ini bertujuan untuk membangunkan sebuah model klasifikasi pergerakan tanpa penanda untuk pekerja-pekerja berdasarkan timbunan model rangka dalam video pergerakan dan menentukan strategi perlombongan data yang terbaik bagi mengkategorikan kelas pergerakan manusia dalam industri. Lapan orang yang berumur 23 hingga 24 tahun sukarela untuk melibatkan diri dalam ekperimen melaksanakan empat jenis pergerakan, iaitu memindahkan kotak, memindahkan baldi, menyapu lantai dan mengepel lantai. Semua pergerakan dirakam dalam video secara berasingan mengikut jenis pergerakan. Semua rakaman video akan ditindihkan dengan model rangka yang terdiri daripada koordinat bahagian badan dan garisan penyambungan bahagian badan melalui algoritma pengaturcaraan. The algoritma menggunakan set data COCO dan modul OpenCV dalam Python untuk menganggarkan koordinat bahagian badan dalam model rangka. Data yang diekstrak daripada model rangka mengandungi halaju asal, halaju kumulatif dan pecutan kumulatif bagi setiap bahagian badan yang terlibat. Proses perlombongan data pergerakan termasuk normalisasi data, subsampel secara rawak dan klasifikasi data untuk mencari kriteria terbaik mengasingkan kelas pergerakan. Data vektor pergerakan dinormalisasi dengan tiga teknik normalisasi yang terdiri daripada normalisasi penskalaan perpuluhan, normalisasi min-maks dan normalisasi skor-Z untuk membentuk tiga set data bagi proses perlombongan data.

Ketiga-tiga set data dicubakan dengan lapan pengelasan untuk mendapatkan kombinasi algoritma pengelasan dan teknik normalisasi yang terbaik untuk mengelaskan data. Lapan algoritma pengelasan yang dikaji ialah ZeroR, OneR, J48 pepohon keputusan, kehutanan rawak, pepohon rawak, pengelas Bayes naif, jiran k-terdekat (k = 5) dan perceptron berlapisan. Keputusan pengkajian menunjukkan bahawa pengelas kehutanan rawak mencatatkan ketepatan pengelasan tertinggi dengan set data yang dinormalisasi oleh teknik min-maks, 81.75% bagi set data tanpa subsampel semula, 92.37% bagi set data yang melaksanakan cara subsampel rawak. Normalisasi min-maks hanya memberikan peningkatan keputusan yang tidak signifikan dengan menggunakan algoritma pengelasan sama, tetapi cara subsampel semula secara rawak meningkatkan ketepatan klasifikasi dengan margin besar. Cara subsampel secara rawak menyingkirkan data yang tidak relevan dan menggantikan nilai-nilai tersebut dengan salinan data yang lain. Teknik normalisasi dan pengelas perlombongan data yang paling sesuai telah ditambahkan ke dalam model klasifikasi pergerakan manusia untuk melengkapkan perkembangan model.

.

ABSTRACT

Motion classification system marks a new era of industrial technology to monitor task performance and validate the quality of manual processes using automation. However, the current study trend pointed towards the marker-based motion capture system that demanded the expensive and extensive equipment setup. The markerless motion classification model is still underdeveloped in the manufacturing industry. Therefore, this research is purposed to develop a markerless motion classification model of shopfloor operators using stick model augmentation on the motion video and identify the best data mining strategy for the industrial motion classification. Eight participants within 23 to 24 years old participated in an experiment to perform four distinct motion sequences: moving box, moving pail, sweeping and mopping the floor, recorded in separate videos. All videos were augmented with a stick model made up of keypoints and lines using the programming model. The programming model incorporated the COCO dataset and OpenCV module to estimate the coordinates and body joints for a stick model overlay. The data extracted from the stick model featured the initial velocity, cumulative velocity and acceleration for each body joint. Motion data mining process included the data normalization, random subsampling method and data classification to discover the best information for separating motion classes. The motion vector data extracted were normalized with three different techniques: the decimal scaling normalization, min-max normalization, and Z-score normalization, to create three datasets for further data mining. All the datasets were experimented with eight classifiers to determine the best machine learning classifier and normalization technique to classify the model data. The eight tested classifiers were ZeroR, OneR, J48, random forest, random tree, Naïve Bayes, K-nearest neighbours (K = 5) and multilayer perceptron. The result showed that the random forest classifier

scored the best performance with the highest recorded data classification accuracy in its min-max normalized dataset, 81.75% for the dataset before random subsampling and 92.37% for the resampled dataset. The min-max normalization gives only a slight advantage over the other normalization techniques using the same dataset. However, the random subsampling method dramatically improves the classification accuracy by eliminating the noise data and replacing them with replicated instances to balance the class. The best normalization method and data mining classifier were inserted into the motion classification model to complete the development process.

CHAPTER 1
INTRODUCTION

1.1 Overview

This chapter covers a brief background of shopfloor operator motion classification in the performance evaluation. Problem statements are discussed after extracting information from literature research. It leads to the proposed research objective to resolve the problem stated, followed by the scope of work to clarify the direction of research implementation.

1.2 Research Background

Human motion analysis is emerging as one of the crucial elements in the manufacturing industry to measure performance evaluation. With video technology advancing rapidly, human motion data has gathered many study interests from biomechanical experts and computer vision explorers. Motion analysis serves as an essential feature to detect and classify specific human motion in many applications such as sports performance analysis (Ferdinands, 2010), medical rehabilitation (H. Zhou & Hu, 2008), video surveillance (Garibotto, 2009) and virtual reality gaming (Kloiber et al., 2020). In the manufacturing field, the motion classification helps to verify the presence of action in an inevitable process by operators. Conversely, the absence of specific actions can lead to process defects and incompletion (Aehnelt et al., 2014). Furthermore, it also improves workplace safety by detecting actions that might lead to potential injury and supports robotic-human interaction with the simulation of human motion activity. (Han et al., 2012)

The book “Human Motion Sensing and Recognition: A Fuzzy Qualitative Approach” describes motion analysis as the combination of sensing the human body

and extracting the static or dynamic data in the form of gesture, behaviour, and actions from the human body (Liu et al., 2017). Fundamentally, human motion analysis systems consist of the description and recognition of human body motion. Motion analysis system can be expanded into several sub-parts with tracking, classification, quantification and prediction. Motion classification, which strongly associates with motion recognition, identifies the pattern of human body parts movement based on image sequences and eventually makes successful categorization (Hernandez et al., 2009).

Recent years' trend indicated a transition from the manual evaluation of human movement to a vision-based motion recognition method. Due to the subjectivity of human view, random error and motion variation can affect the human judgement accuracy. Therefore, plenty of researches dived deep into the potentials of using computer vision or artificial intelligence model to recognize human activities (Colyer et al., 2018; Kale & Patil, 2016; Yu et al., 2019). It aligned with the ongoing trend of utilizing computer technology to replace manual work to improve efficiency without compromising accuracy.

The history of the vision-based motion classification model started a few decades ago when fixed-axis and parallel projection assumption is discovered to calibrate feature points relative to the previous position for the human body parts (Webb & Aggarwal, 1981). The general framework of a vision-based motion classification model involves movement scene capturing, human tracking, humans and motion representation, motion recognition and classification into its respective class (Mohamed & Ali, 2013). The model generally processes each frame from the motion video according to its frame sequences. After a human is detected in a video frame,

segmentation of the frame image is applied to obtain the region of interest (Rittscher & Blake, 1999).

The motion can be visualized by augmenting a stick-figure model, volumetric model, 2D-blobs and geometric drawing (Aggarwal & Cai, 1999). Among these motion visualization methods, stick figure model offers a simple but efficient solution to estimate a human posture at a particular frame. The stick-figure model is a skeleton-like model comprising several keypoints, with each representing a body part. A keypoint is the coordinate of an observed body part in the image. A line connects each pair of body parts. These body parts act as moving joints, and their motion vectors are evaluated in comparison to their counterparts of the previous frame (Choi et al., 2012). By comparing the person's movement between each frame, the motion is categorized by one of the several activity recognitions approaches such as pattern-based recognition and dynamic-time wrapping (Liang Wang et al., 2003).

Due to variation in each person's physiques and way of executing activities, a motion recognition model might face a question of robustness and accuracy. The same question also applies to many noise factors in a factory setting such as location, lighting and clothing (Aehnelt et al., 2014). Therefore, it is vital to obtain diverse data for processing to improve the model's performance. A knowledge discovery process identifies the rules and information to differentiate the class of motion in the captured or real-time video from these data.

The knowledge discovery process (KDD), also known as knowledge discovery in database, is defined as the nontrivial research process, techniques and tools used to identify valid, novel, potentially useful, and ultimately understandable patterns in data (Agrawal & Shafer, 1996). Known as a machine-learning process, it consists of a few steps from developing algorithm application to consolidating the discovered knowledge

(Cios et al., 2007). Data mining is a vital step in the process flow. It outputs the patterns from the large sets of well-prepared data using the classifier methods such as classification rules, decision trees and conditional probability (Barhate et al., 2018). The evolution of data mining from the 1990s led to its expanded range of applications, including face and body motion analysis (Mariscal et al., 2010). Many of the classifiers were developed as readily-used datasets and automated algorithms to evaluate the human face and body pose. Testing this concept into the manufacturing field might revolutionize the performance review procedure of workers in the factory.

1.3 Problem Statement

Lately, companies slowly learned the powerful tool of implementing automated biomechanical motion analysis to evaluate the operators' performance (Bortolini et al., 2020). However, the complicated experimental setup or physical cooperative equipment setting like motion capture markers limits the application of the current motion-tracking model despite its incredible effectiveness. Moreover, the marker or fiducials placed on the workers will affect their normal movement and ultimately, their performance.

Meanwhile, the vision-based markerless motion classification model has low number of researches done on its implementation in the manufacturing industry. Previous research on related scope mostly covered basic activities like running, standing, sports, and medical abnormality (Geng et al., 2016; Lu & Chang, 2012). Insufficient evidence of results hurt the confidence of factories to adopt the markerless motion classification system for identifying more complicated activities in the production line, especially considering the factory environment being full of background noise. Therefore, a research about markerless biomechanical motion

classification model with the computer vision and data mining techniques is potentially beneficial in manufacturing field.

1.4 Research Objective

There are two main objectives in this study which are listed as followed:

a) To develop a descriptive model of motion classification based on the overlay of stick-figure model on the operator's motion in video frames.

b) To evaluate the classification accuracy of motion classification model using data mining classifier algorithms and determine the most suitable data mining strategy.

1.5 Scope of Study

A descriptive model of human motion classification analysis based on the stick-figure model augmentation in each video frame featuring an operator detected is developed to identify and verify the motion activity carried out by the production operator. Programming algorithms of stick-figure model overlay are developed with Python to estimate the position of body parts and calculate the vector variables of the motion. The video samples are collected with a camera using different subjects and four types of activities. The vector data is extracted into a data file format and pre-processed for data mining using the programming software of Python. The data mining process is implemented on the data file using the WEKA Experimenter interface to identify the patterns from the motion data. Different classifier methods with various settings are tested to determine the best classifier methods based on the classification accuracy.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

The study of human motion using computer graphics technology has been actively pursued since the evolution of the rotoscoping concept (Moeslund & Granum, 2001), which transferred realistic human motion into animation. This chapter reviews these studies from many aspects. Application and previous methods of motion classification in the manufacturing industry are discussed. It is followed by reviews on the current existing markerless motion classification model. Past study results of the human motion tracking and segmentation based on video frame are presented. Specifically, the visualisation using a stick figure model onto a human figure in the video frame is focused.

The second part of this research involves data mining. Consequently, the data extraction and data mining strategy are studied separately from the available research to gather the methodology and results for these two stages.

2.2 Motion Analysis in Manufacturing Industry

The rise of motion analysis technologies has enabled the industry to estimate the human pose and manufacturing activity more accurately. According to the survey from Menolotto et al. (2020), the industrial research about motion analysis primarily targeted the health and safety factors in the workplace, with over 60% of the reviewed literatures discussing this application. Conversely, productivity evaluation and task monitoring only occupied a low portion of research purposes in the same collection of studies (Menolotto et al., 2020). It indicated the unfairly low amount of attention to the manual process's motion tracking and performance evaluation of the production line. The limited application could cause the under-utilisation of motion data technology in the

industry. The most-used devices to capture and detect motion discussed in the same survey are the inertial measurement units (IMUs) sensors and camera-based motion capture system.

Kubota et al. (2019) tested the implementation of wearable inertial sensors with motion capture camera in the activity recognition of automotive activities. The activities were broken down into sequences based on walking, scanning, attaching and receiving to control the variation of actions. The experiment subject was entitled to simultaneously wear the Vicon markers as a motion capture marker and Myo armband, a surface electromyography (sEMG) sensor. The setup of markers and sensors were shown in Figure 2-1. Mobile characteristic of sensors and motion capture camera enabled flexible data collection regardless of places. However, the experimental setup would require expensive equipment and lengthy preparation time of wearing, which could cause inconvenience to the conventional production line. It also suffered from the common difficulty to detect fine-grained motions.

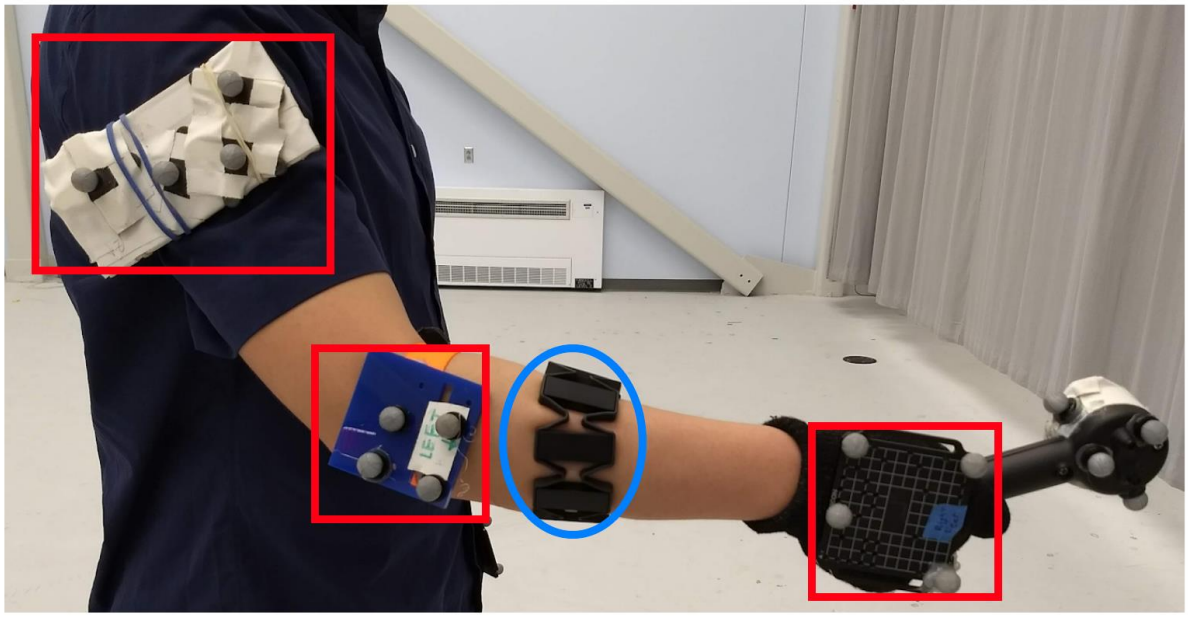

Figure 2-1: Sensors setup on the participant's arm with Vicon markers in the red box and Myo armband in the blue oval (Kubota et al., 2019).

Fine-grained action typically involves the frequent interactions between human hands or fingers with objects. It is a standard feature in manual manufacturing activity. However, fine-grained action recognition is significantly more complex than gross actions such as standing and sitting down due to the higher level of features required. A mid-level action recognition approach was developed, with frame image being divided into several parts but only interaction parts being cropped out for data processing (Y. Zhou et al., 2015). Such an approach was emulated by other researchers for hand pose recognition using OpenPose (Kobayashi, 2019) and with a fine-grained motion estimation approach (FGME) for video frame interpolation (Yan et al., 2020). Both applied the convolutional neural network (CNN) based on the constructed bounding box of interacting body parts, as shown in Figure 2-2, to recognise the action. These approaches eliminated non-moving objects and background in the video frame but were experimented with similar background and action sequences. The effect of different environment and activity type on classification accuracy remains in doubt.

Figure 2-2: Interpolation result using FGME approach with a red box indicating cropped interaction part (Yan et al., 2021)

The alternative method of motion analysis in the industry used the advanced Kinect sensors and camera similar to the Microsoft Kinect system in gaming application. Kinect V2 sensors, combined with an RGB camera, an infrared camera and an infrared emitter, outputted three different images, with one of them being the depth image. The depth image shaded the image contours with different colours based on the distance between sensors and camera (Caruso et al., 2017). Besides the depth map, a

skeleton-like stick figure model can also be extracted from the Kinect sensors via a skeleton model descriptor. Kinect system could achieve acceptable accuracy in motion detection and be capable of capturing 3D motions. However, the markers and Kinect equipment are less affordable and convenient to set up. It also had the potential problem of poor edge detection under dark-coloured background (Dutta, 2012).

Meanwhile, Yu et al. (2019) had used the single RGB camera-based 3D motion capture algorithm to obtain the motion data for biomechanical analysis of fatigue. The methodology mainly aimed to test the absence of marker which might affect the worker's movement. The study discovered that this 3D motion modelling method could estimate the worker's 3D joint locations within an error margin of 4cm, which was just acceptable. In addition, Akanmu et al. (2018) examined the sensor-based Inertia 3DSuit motion capture suit in the construction industry to investigate the effect of these motions on the fatigue of body joints and reassign the workforce for work optimisation. These two methods have a common goal of studying construction activity but required expensive tools instead of the cheaper conventional camera to gather data.

2.3 Markerless Motion Classification Model

Without using motion capture markers, the multi-camera recording played the role to reconstruct the 3D view of moving human bodies. Nakano et al. (2020) tested the setup of multiple video camera from different angles to obtain the frames from various views before merging as 3D visualisation through the direct linear transformation (DLT) method. An alternate form of multi-camera setting practised the asynchronous way by applying audio synchronisation to the conventional video camera recordings and then sent for 3D mesh reconstruction using a feature-based approach (Hasler et al., 2009). However, the synchronised camera setup was a more modern trend

in markerless motion capture (Y. Wang et al., 2018). The feasibility of multi-camera setups is also a considerable concern in the factory setting with space limitations.

For the single video camera, Kanko et al. (2021) proved that the single 2D camera view with a deep-learning approach could produce the results of gait analysis and movement estimation as comparable as the marker-based model. The results were echoed by Wang et al. (2018) in comparing the accuracy of movement joint angles between marker-free and marker-based approach.

A study was carried out to adopt the single-camera to capture the video of general movements by infants and recognise the abnormality in those movements. The study outputted successful result to classify the abnormality. It followed a framework starting from feature extraction using computer vision, movement analysis using formula calculations and finally, the movement classification with a feedforward-type network known as a log-linearised Gaussian mixture network (LLGMN) (Tsuji et al., 2020).

A low-budget 2D-camera system developed by Zult et al. (2019) showed that the conventional video camera could extract the valid keypoints of body parts in the video frame based on the markers. The markers could be replaced by virtual coordinate points using a computer vision module such as OpenPose (Xu et al., 2020). The study by Kim et al. (2021) applied the OpenPose module, a type of artificial intelligence-assisted motion analysis system, to predict the knee and hip movement angles in a video captured using the smartphone camera. The validity of this OpenPose-based system with the post-processing automated algorithm showed early promise but may require further verification.

2.4 Human Motion Segmentation

Motion segmentation serves as the pre-processing stage of motion analysis to cluster the long frame sequences depicting human actions into several shorter, non-overlapping video segments. Subspace clustering works by searching subspace and cluster from a dataset and categorising data into new distinctive spaces based on similar features. For example, an approach proposed by Xia et al. (2018) discussed implementing a robust kernel low-rank representation method to combine with the sparse subspace clustering. Its sparse representation of motion data could be used for motion recognition, but it ignored the temporal correlation between successive frames.

Temporal data clustering aims to cut long sequential data into a set of non-overlapping parts. L. Wang et al. (2019) applied this clustering method in motion data segmentation. They emphasised that temporal information was crucial to achieving accurate model performance. The temporal clustering method had its drawback due to it being an unsupervised learning method. The transfer learning approach can overcome the unpredictable result of this unsupervised clustering method.

As the motion is analysed in the video format, a data mining strategy to consider time sequence in the video was discussed in several works (Mallikharjuna & Reddy, 2020; Tasoulis et al., 2013; Vijayakumar & Nedunchezhian, 2012). Mallikharjuna & Reddy (2020) explained the framework of video data mining with several stages. The video frames were splat into individual images before analysing and extracting data on all the individual images. Poms et al. (2018) added that the efficient video analysis extracted a fixed interval of frames without taking all frames to conserve the computational power of video data mining.

Transfer learning benefits from the prior knowledge from related source data to improve feature identification in the target data. Many recent studies practised transfer

learning based on existing datasets on the Internet to visualise the object motion. Several works have partially (T. Zhou et al., 2020) or fully adapted (J. Zhang et al., 2017) the transfer learning by applying deep neural network classifier parameters in these datasets to estimate the motion. Furthermore, the Haar cascade classifiers were tested to detect the human from the image data after being transformed into a thermal or grayscale map. It proved helpful, especially for detecting multiple people in the same image frame (Setjo et al., 2017). However, the classifiers can also incorrectly detect objects with similar characteristics as human.

Qu et al. (2020) suggested an automated human segmentation by mapping motion data into a hidden space map to apply character function without using body position coordinates. Choudhury et al. (2018) proposed a process flow to detect and partition the image region of human with background elimination by deviation thresholding and using holistic human descriptor to represent human silhouette orientation histogram. The shortcomings, however, were the ability to segment the human partially blocked by obstacles or workpiece and under the changing light in the environment.

Rubino et al. (2015) introduced semantic motion detection, which identifies the matches of objects between two views using semantic information. Its underlying principle was similar to the convolutional neural network model that obtained pattern from training data to identify features in the target data. Simonyan & Zisserman (2014) proposed a two-stream convolutional network model that incorporated spatial and temporal networks. This model identified the moving action in the testing video with prior knowledge of training data from the optical flow model. Meanwhile, D. Zhou & He (2020) estimated the human body region in the image using the recurrent network model by transforming the image into a pose heatmap. The coordinates of body joints

would be evaluated from the heat map, and these coordinates are vital to constructing the stick figure model.

2.5 Visualisation of Stick Figure Model on Human Motion

Stick figure model, also called the skeleton model, is a skeleton-like structure to track the body motion pattern by representing the important body joints and connect these joints into a complete model (Guo et al., 1994). It is accomplished by annotations of keypoints from the body pose estimation.

The earlier work used hand-crafted features such as Histogram of Oriented Gradient (HOG) to construct the stick model. However, the accuracy of identified keypoints was below the acceptable range (Dalal & Triggs, 2005). The modern studies of human body position estimation in a media file are divided into two main categories: single-person and multi-body. For the single-person approach, there are two types of frameworks to locate the body parts. The frameworks are direct regression and converting into heatmap (Dang et al., 2019).

For direct regression, Chan et al. (2016) explored the simplified version of the 2D human motion stick model developed through a mathematical regression coefficient model. The simplified 2D stick model presented a more straightforward interpretation with the usage of joints as calculation points. The 2D stick model was constructed by several points of body parts and lines, as shown in Figure 2-3.

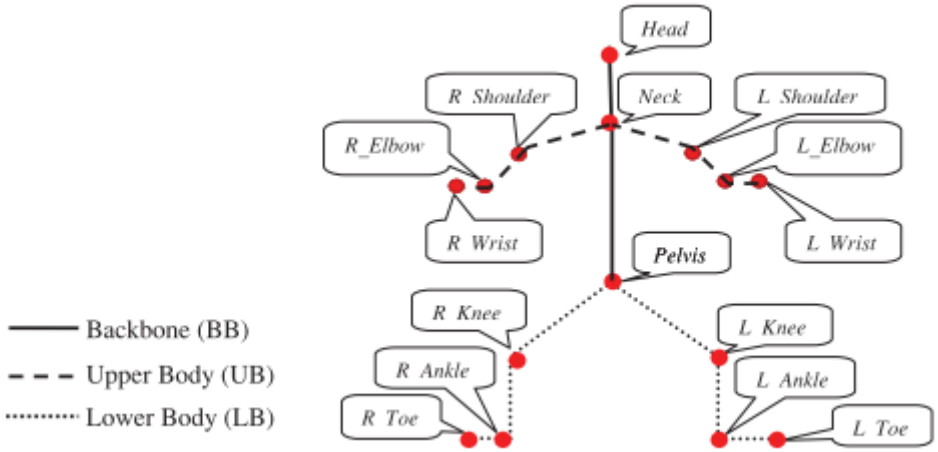

Figure 2-3: A Simple 2D Stick Model Comprising of Three Main Body Segments (Chan et al., 2016)

However, the regression-based stick model construction always requires additional procedures to map the feature points accurately onto the subject in an image. Carreira et al. (2016) introduced a corrective measure by providing a simple error feedback connection in the neural network model structure. The predicted error was fed back into the network like backpropagation to improve the prediction of keypoint locations progressively. Luvizon et al. (2019) presented an improvised method called *Soft-argmax* operation. This operation can convert the feature maps directly to joint coordinates by finding the maximum values from the target functions after being integrated into the deep convolutional neural network. This new method achieved a comparable result as the heatmap-based framework. However, the problem existed in expanding into multi-person cases, unlike the heatmap-based approach.

The foundation of the detection-based framework usually lies in the deep learning datasets that are pre-trained using thousands of human images. Sun et al. (2019) implemented an approach that used a convolutional neural network with two-strides convolutions to reduce the resolution and the main body that outputted the feature maps. At the end of the network, the regressor estimated the key point positions by evaluating

the loss function of the heat map using comparisons between predicted heatmap and ground-truth heatmap.

COCO and MPII are two of the most popular datasets, with them acting as benchmarks in several studies' experiments (Carreira et al., 2016; Papandreou et al., 2017; Sun et al., 2019). COCO datasets contained over 200,000 labelled pictures with keypoints and human instances to train the model. It functions by defining the object keypoint similarity (OKS) and calculating the mean average precision (AP) over 10 OKS thresholds as the primary comparison metric (Lin et al., 2014; Xiao et al., 2018). MPII datasets consist of over 40,000 body-pose pictures with a wide variability of appearance and activities in different environments. MPII datasets evaluate the presence of body pose by calculating the head-normalised probability of correct keypoint (PCKh), which indicates the current joints if the PCKh score falls within the calculated pixels of ground-truth position (Andriluka et al., 2014).

The COCO-WholeBody dataset was the extension of the COCO dataset, and it estimated the whole-body position with more attention details to each body part. Meanwhile, these datasets exist in multi-people detection variations such as MPII human multi-person dataset and COCO keypoint challenge dataset. Fang et al. (2017) proposed the regional multi-person pose estimation (RMPE) that used a top-down strategy by segmenting regions containing individual human and identifying the keypoints in each region representing a human body. The approach had a weakness of classifying people with overlapping regions and similar characteristics with the background. The RPME approach contrasted with another multi-person pose estimation strategy called the bottom-up strategy. The online-available OpenPose module applied the bottom-up Part Affinity Field (PAF) approach that recognised the keypoints first before associating with individual persons (Cao et al., 2021). The module only

recognises human pose but without any classification of activities which is the objective of our study.

2.6 Motion Data Extraction

Motion data can be obtained in many ways. The most convenient method is by using sensors. Most sensor-based motion data were extracted from the calibrated output of the sensors such as orientation, acceleration (Um et al., 2017), and movement intensity (Shi et al., 2020) before being sent to the following convolutional neural network (CNN) or recurrent neural network (RNN) for classification. Drumond et al. (2018) evaluated the subject movement against time with the joint rotation, orientation and angular speed using the IMUs sensors as a pre-processing step of the RNN. Nonetheless, sensor-based neural network classification requires a time-consuming process flow from preparing sensors to classification output.

The suitable parameters of stick-figure model data extraction usually revolve around joint motion and vector data. Grigg et al. (2018) used the joint angle data, while Elias et al. (2017) chose the feature vector values as the extracted motion attributes. Sedmidubsky et al. (2021) deduced that the raw skeleton model data of individual pose could be compared using similarity measures such as Euclidean distance. While the Euclidean distance can be computed using feature vector values, both attributes had difficulty distinguishing the similar or complex type of actions.

Fong et al. (2015) suggested a large set of attributes selected as shown in Table 2-1 for further classification. The attributes selected include axis position, velocity, and acceleration before using the summation formula to add the values together to consider the time series factor.

Table 2-1: Gesture Attribute Data Selection by Fong et al. (2015)

Position Data	Vector Data
(1) lh$x$: position of left hand (x coordinate)	(1) Vectorial velocity of left hand (x coordinate)
(2) lh$y$: position of left hand (y coordinate)	(2) Vectorial velocity of left hand (y coordinate)
(3) lh$z$: position of left hand (z coordinate)	(3) Vectorial velocity of left hand (z coordinate)
(4) rh$x$: position of right hand (x coordinate)	(4) Vectorial velocity of right hand (x coordinate)
(5) rh$y$: position of right hand (y coordinate)	(5) Vectorial velocity of right hand (y coordinate)
(6) rh$z$: position of right hand (z coordinate)	(6) Vectorial velocity of right hand (z coordinate)
(7) hx: position of head (x coordinate)	(7) Vectorial velocity of left wrist (x coordinate)
(8) hy: position of head (y coordinate)	(8) Vectorial velocity of left wrist (y coordinate)
(9) hz: position of head (z coordinate)	(9) Vectorial velocity of left wrist (z coordinate)
(10) sx: position of spine (x coordinate)	(10) Vectorial velocity of right wrist (x coordinate)
(11) sy: position of spine (y coordinate)	(11) Vectorial velocity of right wrist (y coordinate)
(12) sz: position of spine (z coordinate)	(12) Vectorial velocity of right wrist (z coordinate)
(13) lw$x$: position of left wrist (x coordinate)	(13) Vectorial acceleration of left hand (x coordinate)
(14) lw$y$: position of left wrist (y coordinate)	(14) Vectorial acceleration of left hand (y coordinate)
(15) lw$z$: position of left wrist (z coordinate)	(15) Vectorial acceleration of left hand (z coordinate)
(16) rw$x$: position of right wrist (x coordinate)	(16) Vectorial acceleration of right hand (x coordinate)
(17) rw$y$: position of right wrist (y coordinate)	(17) Vectorial acceleration of right hand (y coordinate)
(18) rw$z$: position of right wrist (z coordinate)	(18) Vectorial acceleration of right hand (z coordinate)
	(19) Vectorial acceleration of left wrist (x coordinate)
	(20) Vectorial acceleration of left wrist (y coordinate)
	(21) Vectorial acceleration of left wrist (z coordinate)
	(22) Vectorial acceleration of right wrist (x coordinate)
	(23) Vectorial acceleration of right wrist (y coordinate)

	(24) Vectorial acceleration of right wrist (*z* coordinate)
	(25) Scalar velocity of left hand
	(26) Scalar velocity of right hand
	(27) Scalar velocity of left wrist
	(28) Scalar velocity of right wrist
	(29) Scalar velocity of left hand
	(30) Scalar velocity of right hand
	(31) Scalar velocity of left wrist
	(32) Scalar velocity of right wrist

2.7 Data Mining Strategy for Motion Classification

Al-jabery et al. (2020) explained the criticality of the data preparation procedure and its general flow. The data preparation is purposed to clean the unimportant data and enhance the data quality before the actual data mining process. Its general flow consists of removing or replacing missing values, converting numeric class attribute into nominal class, removing redundant instances, detecting outliers and normalization. The data class is labelled with action descriptive words in the motion dataset that can save the effort to convert numeric attributes. The missing values in the time-series video data can be estimated using several approaches such as the deterministic approach, stochastic approach, and regression model (Fung, 2006). These approaches fill the missing values with the fit function or interpolation method that consider the complete time-series data. However, the missing value estimation of motion data may require different strategies that only look into a small group of frames near the time point because a full video data function demands high computational power.

Resampling is proved as one of the data preprocessing method that potentially improve the data mining process if the class distribution is imbalanced such as the research by Khaldy & Kambhampati (2018) to investigate the effect of resampling on the heart failure dataset. Common resampling method include the K-fold cross validation, bootstrap method and random subsampling method. Mehra & Agrawal

(2020) compared the random under-sampling and random over-sampling to evaluate their application in the imbalanced dataset. The random under-sampling removed part of the instances from majority classes, while over-sampling method duplicated several instances from minority classes to balance the class bias. In the motion dataset, the class bias is balanced but the noise instances in these classes might potentially raise some data classification issues. The resampling method can eliminate these instances without causing imbalanced data classes.

Switonski et al. (2019) explored markerless motion extraction data mining in the research regarding motion capture data. The researcher utilised dynamic time warping (DTW) to classify the human motion data into gait patterns. The model identified the variation in the orientation of motion capture and subject for motion recognition in time-series data. It obtained the angles in the joint data and evaluated the closest probability of classification with the minimum distance classifiers (MDC). MDC was combined with k-nearest neighbour (KNN) classifiers to extract the advantages of both types of classifiers for better accuracy and consistency. Schneider et al. (2019) also applied the DTW approach after annotating the skeleton model based on the OpenPose module dataset to evaluate the warping distance. Before the classifier was applied, the image data in coordinates were normalised to condense the data range into a smaller number. Then, the warping distance of time-series data was classified using nearest neighbour classifiers. The result still had limitations such as reliance on the representativeness of the dataset, poor precision of recognition when noise reduction is needed, and motion capture marker setup required.

Qian et al. (2010) tested the multi-class support vector machine (SVM) classifiers by extracting the centroids and instantaneous speed of human motion after eliminating the background. The frame sequence comparison outputted the contour

coding of motion energy image (CCMEI) with square-to-circular coordinates transformation that changed the plane coordinates into polar coordinates. SVM also acted as the classifiers in the study by Choi et al. (2013) to classify the gait motion pattern. The parameters used include the joint angle and distances between body parts. While SVM is an excellent option to recognise motion with great accuracy, plenty more classifiers are yet to be tested in motion classification.

Yang & Zhao (2011) adopted the decision tree classifiers to determine firefighters' motion class, but the attributes were composed of string-type descriptions and not in numeric form. H. Zhang et al. (2012) tested three classifiers of Naïve Bayes, SVM and random forest in classifying six different motions for an interactive system. The result deduced that the Random Forest classifier scored the highest classification accuracy using position and vector data. Li et al. (2020) also examined the motion recognition model using the random forest algorithm using the normalised joint coordinates difference between keyframes. Fong et al. (2015) agreed with the result of the random forest classifier being the best performance using position and vector data from the skeleton model. Its classification accuracy was higher than the neural network approach and other traditional classifiers.

2.8 Summary

The markerless motion classification model plays a vital role in the manufacturing industry to monitor tasks compliances and control process quality based on the information from past reviews. The majority of motion analysis methods in the current industry focused on the marker-based model, but the setup cost and time proved the vast obstacles. Many previous reports have proved a markerless motion classification model with single 2D cameras with comparable results as complicated

and expensive 3D camera systems. The subject's body motion tracking without markers can be segmented by converting the video into individual frames. Each frame is transformed into a depth map or heat map to detect the presence of the human. The stick figure model augmentation is developed through programming with the help of existing datasets such as COCO and MPII datasets to estimate the positions of each joint. The pre-processing of data mining includes the extraction of position and vector data from stick model and calculation of time-series factor by cumulative sum. Data mining experiments should be carried out for testing whether Random Forest classifiers performed better than other data mining classifiers in motion classification, as suggested by previous studies.

CHAPTER 3
METHODOLOGY

3.1 Overview

This chapter explains the development process of the automated markerless motion classification model. The experimental activity capture via video recording was carried out to obtain the video samples. The stick figure model was overlaid onto the human motion in the capture video frame-by-frame using open-source Google Colab software with the downloaded COCO datasets. The motion data composed of all body joint positions were retrieved from the stick model and used for calculating vector motion velocity and acceleration between frames. The calculated data were extracted into a dataset file to implement the data preprocessing before the data mining process. The data preprocessing step involves the normalization of motion data and the data cleaning procedure. The study experimented with different normalization techniques and classifiers in the data mining experiment to identify the best and normalization methods and classifiers for motion classification. The classifiers were tested on the motion dataset using WEKA software, and the results were compared among the tested classifiers. The knowledge discovered from the data mining experiment was combined into the model to complete the development process.

3.2 Experimental Motion Selection

As the research focused more on the operator's motions from the production site, the experiment was designed to involve activities likely to be observed in the factory. The motion activities featured in the experiment were decided to be moving carton box, moving pail, sweeping floor and mopping the floor. The descriptions of each motion activity were included in Table 3-1. The activities were also illustrated in Figure 3-1 using a sample video frame for each participant's action.

Table 3-1: Descriptions for each type of experimental motion activities

Motion Activity	Description
Move box	Bend down the body, lift the box with two hands, stand upright, walk a few steps, bend down the body, put down the box, resume to a standing position
Move pail	Bend down the body, lift pail by its handle with one hand, stand upright, walk a few steps, bend down the body, put down the pail, resume to a standing position
Sweeping	Grasp a broom with one hand, move the broom down until its brush touching the floor, pull the broom to sweep the dirt, lift the broom up
Mopping	Grasp a mop with two hands, slightly bend the body, move the mop in a direction while it touches the floor, change mop movement to the opposite direction

(a)

(b)

(c)

(d)

Figure 3-1: Experimental motion classes featured in the motion classification, (a) moving box, (b) moving pail, (c) sweeping, (d) mopping

Eight participants between the age of 23 to 24 years old volunteered to participate in the motion video collection. Each participant was required to perform a set of afro-mentioned activities in different settings. Due to variation in different persons executing the action, it affected the pose recognition (Lau et al., 2009), so the movement classification experiment should not be limited to a single person. The different backgrounds and light conditions were applied in the video sample collection because various backgrounds in the video affected the motion recognition using a markerless system (Bosch et al., 2008). The indoor and outdoor environments were both used in the video recording as background. The outdoor used the natural light, while the indoor light condition had the options of being bright and dim.