2013-04-18

Automated Construction Worker Performance and Tool-time Measuring Model Using RGB Depth Camera and Audio Microphone Array System

Weerasinghe, Ittepana Payagalage Tharindu Rasanga

UNIVERSITY OF CALGARY

Automated Construction Worker Performance and Tool-time Measuring Model Using RGB Depth Camera and Audio Microphone Array System

by

Ittepana Payagalage Tharindu Rasanga Weerasinghe

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CIVIL ENGINEERING

CALGARY, ALBERTA

APRIL, 2013

Abstract

Construction productivity improvement activities in the industry has seen many developments in terms of new tools, techniques and processes been introduced through research and practices. However, on site studies and literature surveys confirm that a huge potential still exists in the industry for further improvements.

The thesis discusses a novel method for developing a sustainable, reliable, integrated, automated and systematic mechanism to extract construction worker tool-time and performance information that assists project managers and planners, in developing strategies for improving labour productivity, labour allocation and developing administrative schemes related to labour performance by using audio and video surveillance techniques addressing the potential drawbacks from the manual observation on construction site.

The research proposes a low cost and composite range sensing device (Microsoft Kinect) consisting of RGB camera, depth camera, and microphone array to extract multiple sensing modalities from indoor construction. A user friendly and comprehensive framework is developed to consolidate location aware information of workers, other site personnel and construction activities in order to generate productivity related data. Additionally, this provides information on worker behavioral analysis (i.e. supervisory effect on performance) and breakdown of non-tool time activities which can be used for better labor allocation strategies. The proposed set of algorithms (i.e. worker recognition, construction activity detection, direction of arrival detection and tool-time analysis) has been validated in a real work environment and experimental results witnessed over 90% precision. In brief, validated results reflect the potential for automated assessment of

worker tool time in indoor construction environment and each of these implications make an important contribution to the body of knowledge in construction automation.

Acknowledgements

It is a humbling experience to investigate social structure and opportunity. This research is a constant reminder that whatever knowledge I produce is not limited to my own doing. I am grateful to those that have, all too willingly, given of themselves during my pursuit of a doctoral degree.

I owe my deepest gratitude to my supervisor, Dr. Janaka Ruwanpura for all the ideas, advice, suggestions, guidance and encouragement given to me throughout the course of my doctoral research. I strongly believe that he provided a new source of inspiration to me with his energy, enthusiasm and motivation during the PhD program.

I would like to thank all the members of my supervisory committee, Dr. Jeffery Boyd, Dr. Aymen Habib for providing guidance, insightful comments and valuable input throughout this lengthy research project. I am also grateful to Dr. Sudarshan Mehta and Dr. Mohamed Al-Hussein for agreeing to be in my examination committee and devote their precious time in giving me feedback on the work I have done.

I am especially grateful to Dr. Arjuna Madanayake who inspired this research and for his thought-provoking conjectures. I would also like to extend my gratitude to Dr. Kamal Ranaweera, Dr. Thushara Gunarathne and Mr. Nuwan Ganganath for constructive criticism, guidance, and invaluable inputs throughout the research.

There were many graduate students and research assistants who have assisted me in this research when I conducted the research amidst a challenging environment. Mr. Varuna Adikariwattage, Mr. Amila Silva, Mr. Sulochana Madanayake, Mr. Niroshan Withanage and Mr. Udara Waraketiya deserve recognition and I would like to express my heartfelt appreciation for their unconditional support during the data collection and

analysis. To the productivity improvement research students, your commitment to duty, academic abilities and friendship offered the best environment for completing my thesis.

This research was funded by a group of companies EllisDon, Graham, PCL, Ledcor, CANA, Stuart Olson, Revay, Natural Science and Engineering Research Council (NSERC), Calgary Construction Association, city of Edmonton and Lafarge. I would like to thank for their generous support and advices and this research would have been impossible without their support.

It is an honour to acknowledge the early guidance, encouragement, blessings, and support extended by my family. I must express my eternal love and deep respect for my mother and brother, who always carry a silent wish for my success, have been the motivation for my education. Without their vision of my future, a doctoral degree would have been completely impossible.

Last, but not least, my heartfelt thanks to my dearest wife and my best friend Vindhya for her love, understanding, kindnesses, encouragement and belief in my abilities. I thank her for taking this journey, ever so patiently, with me. She has been and continues to be one of the greatest gifts to me.

Dedication

This thesis is dedicated

to

My Ever-loving mother Maya

My Smartest brother Eranga

and

My Dearest wife Vindhya

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

2D	Two Dimensional
3D	Three Dimensional
4D	Four Dimensional
AACEI	Association for the Advancement of Cost Engineering International
AOA	Angle of Arrival
AR	Augmented Reality
BB	Bounding Box
BIM	Building Information Modeling
CAD	Computer Aided Design
CC	Cross Correlation
CII	Construction Industry Institution
CQS	Craftsmen Questionnaire Survey
DOA	Direction Of Arrival
ECC	Eccentricity
EOP	Exterior Orientation Parameters
FDR	False Discovery Rate
FDS	Foremen Delay Survey
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FIRBPF	Finite Impulse Response Band Pass Filter
FPR	False Positive Rate

FPS	Frames per Second
GCC	Generalized Cross Correlation
GDP	Gross Domestic Product
GMM	Gauss-Markov Model
GPS	Global Positioning System
GTT	Group Timing Technique
GUI	Graphical User Interface
IOP	Interior Orientation Parameters
Li-DAR	Light Detection and Ranging
MEX	MATLAB Execution
MLE	Maximum Likelihood Estimation
NPV	Negative Predictive Value
PHAT	Phase Transform
Q1	1-Quantile
Q2	2-Quantile (median)
Q3	3-Quantile
RFID	Radio-Frequency Identification
RGB	Red, Green, Blue
RGBD	Red, Green, Blue, Depth
ROC	Receiver Operating Curve
ROI	Region of Interest
rpm	Revolutions per Minute

RSSI	Received Signal Strength Indicator
SDK	Software Development Kit
SLR	Single Lens Reflection
SN	Sensor Networks
SNR	Signal-to-Noise Ratio
SPR	Single Photo Resection
STE	Short Time Energy
TDE	Time Delay Estimation
TDOA	Time Distance Of Arrival
TNR	True Negative Rate
TOA	Time Of Arrival
TPR	True Positive Rate
UWB	Ultra Wide Band
VGA	Video Graphics Array
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Networks
ZCR	Zero Crossing Rate

Chapter One: **Introduction**

This chapter briefly explains the rationale of the research, from its inception through gradual evolution towards its final outputs and findings. The latter half of the chapter is organized as follows: First, the states of knowledge in the automated construction worker activity analysis are briefly reviewed. Next, the research objectives and methodologies are presented. Finally, the thesis chapter organization is briefly discussed.

1.1 Rationale for Construction Activity Analysis

An activity analysis examines the amount of time workers spend on specific construction activities. Activity analysis, which involves a continuous and comprehensive process of benchmarking, monitoring, and improving the amount of time craft workers spend on diverse set of construction activities, can play a vital role in improving construction productivity, safety, and occupational health. In comparison with work sampling, detailed assessment together with continuous improvement significantly differentiates activity analysis and thus provides recommendations for activity monitoring, improvements, and improvement applicability (CII, 2010). Many companies have experienced the benefits of activity analysis in recent years and are now proactively working towards implementing it in their projects (ENR, 2011).

In spite of the above mentioned benefits of activity analysis, certain aspects of this method make it prohibitively expensive as well as cumbersome. For example, an observer is required for every construction activity, for the accurate and detailed assessment of work in progress, which can be expensive. In addition, several cycles of

operation need to be recorded due to the variability in how construction tasks are carried out, or in the duration of each activity. The substantial amount of information that needs to be manually collected and analyzed can also adversely affect the quality of the process. Furthermore, differing study techniques are a source of controversy with cross examination (Ganesan, 1984; Whiteside, 2006).

Given the importance of the above observations, the state-of-the-art practice of activity sampling (CII, 2010) suggests that sampling population's methods be defined, such as the Tour or Modified Crew Method. The phase of a construction project, the site arrangement and congestion of activities, and the timeframe of the study (e.g., whether the work begins before or after lunch) may vary the duration of these studies. This may require the entire site to be monitored by the observers for each form of activity. These approaches generally require routes and times to be initially selected in order to make sure results are representative of the actual work carried out. Given all of the above, there is a need for a process that can track construction workers and analyze their construction activities systematically and automatically across entire construction projects, in a reliable and low cost manner. Such a method will significantly eliminate the challenges related to sampling methods or manual observations.

Over the past few years, several research studies (Brilakis, Park, & Jog, 2011; Yang, Arif, Vela, Teizer, & Shi, 2010) have proposed interactive vision methods for tracking project entities such as people, materials, equipments and vehicles. The need for a low-cost tracking mechanism can be addressed by some of these methods. However, a solution for automated recognition of construction worker activities has not been

addressed in these studies. In particular, applicability of these approaches in real job site can be affected considerably by varying illuminations and static and dynamic occlusions.

This study proposes a novel method for automatically recognizing multiple interacting construction workers and their activities using Microsoft Kinect video and multi-channel audio data information to address above limitations.

1.2 Current Research in Automated Activity Analysis

Methods such as that described by Brilakis et al. (2011) do not propose a solution for activity recognition and real-time tracking of workers, but focus only on tracking construction entities. Previous studies have compared several existing, 2D vision-based tracking algorithms on construction sites and identified several challenges when workforce interactions occur (Arif & Vela, 2009). A recent study has developed a supervised tracking algorithm that requires the user to manually identify the construction workers, and subsequently a machine-learning algorithm learns and tracks the target (Yang et al., 2010). The authors indicated that under changes in lighting condition and with partial occlusions, conditions that are predominant on construction sites, the proposed algorithm fails. Peddi, Huan, Bai, and Kim (2009) have also proposed a blob-tracking algorithm to track productivity of workers by classifying their poses into categories such as effective, ineffective, and contributory. In this method the assumption is that workers are constantly moving and each stance belongs to a certain type of action. An action recognition based on human pose analysis, which is a similar concept, has been

proposed by Escorcia, Dávila, Golparvar-Fard, and Niebles (2012) using the Kinect device.

When it comes to activity analysis, the simple assignment of each static pose to an action can be a major limitation. Recently, a classification algorithm for recognizing human actions from colour and depth was proposed (Escorcia et al., 2012; Shotton et al., 2011; Sung, Ponce, Selman, & Saxena, 2011). In this work we are mainly concerned with forming a method that can detect construction workers and classify their activities using audio visual data and signal processing techniques. The need for a new method that can simultaneously perform tracking and activity recognition of multiple interacting construction workers is the primary objective of this study.

1.3 Research Concept and Overview

Our proposed system is inspired by the human sensory (i.e. vision and auditory/hearing) system. The human stereo visualization system tracks and recognizes objects in 3D space, while the auditory system detects pre-known sounds and localizes the acoustic source direction. Then the brain analyzes the relationships between visual and auditory outputs to generate accurate results. A similar concept is adopted and explained in this thesis.

We used the Kinect for Windows SDK version 1.6 for our study in order to extract RGB, Kinect skeleton figures and extensive depth information of the scene. The region of interest (ROI) for analyzing workers was filtered using skeleton information. Then an integrated approach was applied to detect the hardhat. This involves visual features of the image, characteristics of the hardhat such as unique shape and colour and a

logistic regression classifier. Kinect infrared camera generates a depth map of the environment and based on this information, 3D coordinate of each pixel can be determined with respect to the Kinect device. The rigid body transformation algorithm was then applied to transform the Kinect coordinate of the worker location to the building coordinate system.

Meanwhile, an audio pattern recognition process was followed with a microphone array audio input. This enabled recognition of pre-identified tool sounds in the audio frame. We proposed strong audio features including temporal, energy, spectral moments, and finite impulse response (FIR) filters combined with a probabilistic classifier to differentiate the correct tool sound from general construction noise. Further, the direction of the acoustic sound source was estimated by applying two different cross correlation techniques: standard cross correlation (CC) and the generalized cross correlation function with the phase transform filter (GCC-PHAT). This context-specific information, such as location, time, and identity, can be cross-referenced with various context parameters, such as the task currently being performed, to define highly specific information pertinent to the decisions at hand. Thus, continuous assessment of these work activities is then efficiently used to analyze the tool time (defined in section 2.1) and performance, which ultimately support labour related decisions. In addition, other temporal information such as location of supervisors and workers is used to differentiate non-working time into several pre-defined categories: supervisory instruction time, material handling, tool handling, etc.

1.4 Research Objectives

The primary objective of the research is to develop an integrated and automated mechanism to extract construction worker tool time and performance information by using audio and video surveillance techniques to address potential drawbacks in manual observation on construction sites. This will assist project managers and planners in developing strategies for improving labour productivity and labour allocation, and in developing administrative schemes related to labour performance.

There are many secondary objectives associated with this primary research objective that need to be addressed effectively in the course of the research study in order to achieve the final overall objective defined in the research. These secondary objectives will allow the primary research concept to be integrated, systematized, and well understood by the industry in the long run. Hence, equal importance must be given to these secondary research objectives, listed below, in the research execution phase.

1. Become familiar with the construction operations, tools and techniques, and common productivity improvement and control approaches adopted by the construction companies in Alberta.
2. Identify the most common and high impacting factors that can be utilized for an automated worker tracking system. This includes identification of a simple efficient method that is common to most construction fields. Further, identify a robust technique for automatic construction activity event recognition, addressing well known critical issues in a jobsite such as dynamic environmental conditions.

3. Develop a robust model to track construction workers and differentiate them based on their work type under various lighting conditions of indoor construction work environment. Further, this method must extend to simultaneously performing tracking of multiple interacting construction workers.
4. Develop a MATLAB model to recognize commonly used construction activities in order to measure the amount of time craft workers spend on different construction activities.
5. Develop an integrated system combining the above two systems to measure the performance and tool time of each craft worker.
6. Validate the model to analyze the likelihood of tracking workers, detecting construction activities and tool time and performance measurements in a real construction environment to ensure consistency of output.

1.5 Expected Outcomes and Research Benefits

This novel technology gathers necessary information about construction workers, related to their performance factors and other worker behavioral patterns in order to assist project managers and planners in developing strategies for improving labour productivity and labour allocation. Performance factors explain tool time and work rate, time spent on each task of every worker. Meanwhile supervisory effects on work performance, performance rate variation/distribution, and worker movement patterns are categorized into worker behavioral information that can be carefully utilized for better labor allocation strategies. Additionally, location information is effectively used to classify

non-productive time into pre-identified non tool-time class members, such as material handling, tool handling, supervisory instruction time, etc. Compared with manual observation, this continuous collection of data from jobsites improves the validity of data significantly.

1.6 Thesis Structure

This thesis has been structured into seven chapters, each of which is briefly explained below.

Chapter One: Introduction

A brief description of the inception of the research idea and the preliminary research approaches, and industry's need for the research with a comprehensive justification that is further detailed in the thesis.

Chapter Two: Construction Productivity and Tool-Time

An introduction to the productivity concept definitions, measuring methods, and common terminology with some highlights of the research conducted in construction productivity and related subject streams.

Chapter Three: Research Design, Methodology, and Theoretical Framework

This chapter explains the development of the research concept, research justification, rationale, and associated research tools and techniques. Research methodology associated with each research objective is also described in detail in this chapter. In addition, this chapter discusses the theoretical basis of the research concept and details of all possible

structured theories that reinforced the core research subject and related ideas.

Chapter Four: Construction Worker Tracking System

This chapter consists of a detailed description of the proposed methodology of the worker tracking model. This includes the background study of related work, concept and assumptions of work, and outlines the full scale methodology and implementation of the tracking model considering different site level scenarios.

Chapter Five: Construction Activity Recognition System

The chapter starts with the selection of an audio classifier to detect construction activities on a real job site. The methods and techniques used to develop the audio recognition model are discussed step by step in this section. Further, sound recognition and source localization techniques are thoroughly reviewed. A detailed model description including selected audio features, function development, and accuracy of the model is also completely reviewed.

Chapter Six: Integrated application and model validation

This chapter explains details of the total integrated system that measures worker tool time and performance. Major tasks and outputs of each application module are visualized and further discussed in this section.

At the same time, the validation techniques used to test the proposed audio and video integrated model and the test environment

condition are detailed in this chapter. A brief on the constraints and limitations experienced or potentially expected in a realistic implementation is also included in this chapter.

Chapter Seven: Conclusions and Recommendations

The conclusion ties together and integrates the research findings, limitations, and various issues covered in the body of the thesis, to make comments upon the meaning of all of it while answering research questions and reflecting thesis statements or objectives. This includes noting implications resulting from discussion of the topic, forecasting future trends, the need for further research, and recommendations and guidelines to assist the potential user implementations.

Chapter Two: **Construction Productivity and Tool-Time**

The chapter gives an overview of productivity concept definitions, measuring techniques, and common terminology with some highlights of the previous research conducted on construction productivity and associated subject streams. Related manufacturing productivity improvement practices, which comprised the majority of the theoretical aspects of the research, have also been reviewed in this chapter.

2.1 Construction Industry & Productivity Trends

During last two decades, a decreasing trend in construction productivity in North America has been emphasized by many researchers (Dozzi & AbouRizk, 1993; Hewage & Ruwanpura, 2006; McTague & Jergeas, 2002). Within the same time period in the other parts of the world a similar declining trend has been observed, indicating the universal nature of the issue.

This declining trend in labour productivity in the US construction industry has been highlighted by Schneider (2003) with the growing trend of labour productivity in non-farm industries over the decade from 1993-2003. Other studies of construction productivity have also shown a decline in the period from 1990 to 2000 compared to other industries (McTague & Jergeas, 2002), including studies and analysis carried out by Statistics Canada on labour productivity, trends in construction labour productivity, and the contributions of the construction industry in the overall GDP. However, external market conditions and political and economic changes after 2000 altered conditions in this industry drastically.

Dozzi and AbouRizk (1993) categorized construction productivity issues into two types: macro issues and micro issues. The macro level deals with contracting methods, labour legislation, and labour organization while the micro level deals with management and operations at the job site. Productivity and tool time definitions are stated below.

Productivity:

In broad terms, productivity can be referred to as production output per given units of input (i.e. man hours, machine hours, cost, energy, or materials).

$$Productivity = \frac{Output}{InputSource} \qquad (1)$$

In the construction industry, the most frequently used definition for productivity is the output per man hour of input, which has been extensively used by many researchers.

Tool time:

Tool time is generally defined as actual worker time spent on direct work related tasks that contributes to producing the final output.

Several researchers (Borcherding & Garner, 1981; Liberda, Ruwanpura, & Jergeas, 2003) conducted studies to investigate factors affecting productivity and also to identify the factors that have the highest impact on construction productivity.

Studies carried out on productivity performance in the construction industry reveal that the construction sector is considered a low productivity sector because of the large portion of its labour force who are unskilled workers, and because of its low usage of technology (Lim & Alum, 1995). Tool time analysis studies focusing on North

American construction industry (Choy & Ruwanpura, 2006; Gouett, Haas, Goodrum, & Caldas, 2011; Hewage & Ruwanpura, 2006; McTague & Jergeas, 2002; Ranasinghe, Ruwanpura, & Liu, 2012) have shown that the actual portion of direct tool time spent on a construction operation generally falls between 40% and 60% of the total worker time. Furthermore these studies detailed the different activities on which a worker spends his time. Thereby we can arrive at a conclusion that a significant amount of daily worker time on site is spent on non-productive activities, such as searching for material, being idle, and waiting for instructions. Thus the outcomes of a tool time measurement give vital information to plan the activities specifying what tasks should be focused on and what areas need to be pre planned to minimize wasted time.

2.2 Productivity and Tool Time Measurements

Given the nature of the construction trade, measurement of site overall productivity or the productivity of a specific activity at any point in time is not the easiest task to carry out. In order to examine the actual average time spent by workers on various tasks during a regular work shift, and to assess the percentage of worker time that directly contributes to the final output, the Construction Industry Institute (CII), Business round table, and AACEI have lead tool time studies for different construction activities.

There has been a constant dialog on the connection between tool time and activity productivity in the productivity associated research domain. Though some researchers have indicated that there is a direct relationship between the two, others dispute this, saying through their own research that there is no such strong relationship. Noor (1998)

stated that the tool time measurements are a mode of measuring productivity on site and further proposed many different methods that could be used to ensure an effective tool time measurement.

As the focus area of this research study is labour productivity, the crew hours spent completing a task was considered the input as per the definition of productivity. One must attempt to produce more for the same input hours, or try to reduce input hours to produce the same output in order to be more productive. It is evident that tool time measurement is an essential study for collecting important information for productivity planning and process control.

According to Noor (1998), productivity measurement techniques fall within a spectrum between two broad types of observational methods, namely continuous observation (e.g. direct observation and work study) and intermittent observation (e.g. audio-visual methods, delay surveys, and activity sampling). A review of the individual techniques that can be used to measure tool time in construction activity is presented in the following sections.

2.2.1 Direct observations

Direct observation, in simple terms, is monitoring work processes and productivity performance in the work phase, taking physical observations, and recording measurements. In this technique, observations are usually made throughout the work day by a trained observer making note of the time spent on various predefined tasks. The direct observation method provides detailed data for understanding productivity, yet has

some inherent and noteworthy deficiencies. A single observer is limited in the number of workers he can observe continuously and hence it is difficult for large projects to adopt this approach. If more than one crew is to be monitored, then more observers are required, resulting in higher costs. Since a single observer is recording different time against the tasks of a group of workers throughout the whole work day, this study has been recognized in the industry as the most arduous and time consuming observation tool.

The main criticism of the direct observation technique is that the observer's presence at the observation location leads workers to change their usual work behaviour, knowing that they are being monitored, which in turn influences the actual observation.

2.2.2 Time motion studies

The first use of time motion study as a measurement to examine productivity related issues was in the manufacturing and service industries. The technique is well known for its suitability in cyclic tasks and hence has been extensively applied in both manufacturing and construction industries since its inception. In time motion studies, the worker time spent on different tasks in producing an output is measured and the results usually are used to improve the work processes and to suggest optimal work sequences. Time study was first introduced and successfully applied by (Gilbreth & Kent, 1911) as a part of their scientific management approach in the automobile industry and introducing new brick laying techniques in building construction.

In this technique, a selected activity or work process is divided into several direct tasks for which the time spent is continuously recorded. For work process modifications and many planning decisions the inferences from time motion studies have been great resources. This technique can be effectively used for close and minute detailed analysis of repetitive task time and motions, mostly in standard work practices and in modular form constructions.

2.2.3 Work sampling

Work sampling is a method that assesses the amount of productive, supportive, and non-productive time spent by workers performing their assigned activities, through periodic observations. Work sampling is defined as "An application of random sampling techniques to the study of work activities so that the proportions of time devoted to different elements of work can be estimated with a given degree of statistical validity" (American Institute of Industrial Engineers, 1989). It identifies trends that affect productivity, and a precise definition has been established by Jenkins and Orth (2004), which states "work sampling is a series of instantaneous observations of work in progress taken at random times over a period of time". The samples collected in such manner are compiled together at the end of the study to illustrate the percentage of the day spent by workers performing productive and non-productive work as per further explanations by Lindenmeyer (2001). The premise of work sampling is that if recurrent observations are made of an activity during the course of a workday, inferences can be made on the distribution of the time workers spend on their daily activities.

Activity sampling is also a different term used for the same process. Activity sampling is defined as a statistical technique that can be used as a means for collecting data.

For the convenience of the observer and due to the disparities in the requirement for the study, there were different modified versions of the work sampling technique widely used in the industry. Some studies used the manual observation method while some observations were done through surveillance videos or other visual methods. As each method has its own merits and demerits, the method most appropriate to the site conditions and the research requirements should be used.

2.2.4 The group timing technique (GTT)

A modified form of the work sampling method is the Group Timing Technique (GTT), which is more suitable for processes that are repetitive and which have a short cycle time (Noor, 1998). Further, an interval of 0.5 to 3 minutes was suggested for observations that are made at fixed intervals within each cycle. With the GTT method, the activity of each crew member is recorded at the instant of the observation. The observation time is generally between 1-2 hours in duration for each work day. Hence compared to traditional activity sampling methods the GTT method is less time consuming (Noor, 1998).

2.2.5 Five minute rating method

The Five Minute Rating Method is another variation of the work sampling study. In this method, as its name implies, a crew should be monitored for a minimum of five minutes or for a length of time equal to the crew size, whichever is greater, with a fixed interval of 0.5 to 3 minutes. In this method, the performance of each crew member is recorded. This emphasizes the effectiveness of the crew. If a crew member is performing well, he is given a certain number of credits for his work. By analyzing the number of credits a crew does obtain and the number of credits the crew could have obtained, the crew's effectiveness is measured (Noor, 1998). Similar to the GTT method, the Five Minute Rating Method is less time consuming than the traditional methods. Its advantage over the GTT method is that it can be applied to any work process irrespective of it being cyclic or not (Noor, 1998).

2.2.6 Foremen delay survey (FDS)

Foreman delay surveys are used as a technique to investigate the real productivity impact factors on site from the foremen's point of view. The key idea of this method is to investigate the major reasons for productivity losses in an activity as determined by a person closely engaged in that activity. The mechanism is intended to enquire about the types of delays that affect worker performance from the perspective of the foreman, who acts as the interface between management and workers and is therefore considered to be the best source of information with less bias. Daily examination of foreman delay surveys provides a strong indication of the performance areas accountable for productivity losses.

Tucker, Rogge, Hayes, and Hendrickson (1982) state the rationale of using foreman delay survey method in productivity investigations: Compared to other observation methods, FDS is more reliable as a good foreman is likely to report data that reflects the actual site work and real issues, including any inefficiency in the administration of work. Other benefits of this method are its ease of use and its cost efficiency.

2.2.7 Craftsmen questionnaire survey (CQS)

The CQS technique primarily involves craft workers identifying causes for productivity losses, but it is not commonly used in the industry due to its potential for strong bias. Smith (1987) criticized this method for its unreliability and it is rarely used in the industry for the purpose of productivity investigation. Craftsmen's perspective on possible productivity loss or tool time loss impacting productivity may not be the most precise nor the most relevant point of view. Their exposure to the correct underlying information is limited and they lack the necessary management and technical knowhow regarding particular conditions to foresee the most relevant root causes for the issues. However a major advantage of this method is that it solicits the views of workers and makes them feel that they are contributing to the task at hand (Noor, 1998).

2.2.8 Audio visual methods

The audio visual method eliminates the distorted observations created by the observer's presence in other productivity measurement and observation methods (Silva & Ruwanpura, 2011). Audio visual techniques mainly include static time lapse photography

techniques and video surveillance techniques. In time lapse photography, conclusions are made by observing 3-4 second time lapsed photographs of the work (Noor, 1998). These methods eliminate the hassle of the observer's manual recording process. In video surveillance it is possible to archive required data and do the analysis in a later stage, and it also facilitates moving the video camera to capture better views and zooming have a clearer view of the work being done. Video surveillance also provides remote access through the internet from anywhere in the world. But video surveillance and time lapse photography, sometimes due to the lack of accessibility to the work crew, can lead to incorrect interpretations of the actual work. Further, this method involves higher initial cost and it is also prone to equipment failure (Noor, 1998).

2.3 Summary

Even in the present day there is still a high profile debate with regard to tool time measurement and its relationship to productivity. However, extensive tool time studies show a positive correlation between tool time and productivity of the task. Hence, onsite tool time measurements will benefit managers in strategic planning, labor allocation, and progress review.

Given the advantages and disadvantages of various tool time measuring techniques outlined in the previous sections, work sampling observation through video surveillance would be the best approach in manual observation addressing major deficiencies such as distorted observations.

Nevertheless, even in this method an observer has to manually assess tool time and non-tool time activities by analyzing time lapse images in archived video files. Time consumption of this method varies based on the crew size and clarity of images.

In brief, manual observations are labour intensive, time consuming, and worker tool time information is subject to human error and limitations in data. Given the operational imperative of construction projects and the ever increasing time pressures exerted on project schedules, the cost of employing personnel to conduct such observations, both in terms of the monetary cost of wages and the time value of observation that does not result in the physical growth of buildings (i.e. non value added), would deter companies from adopting such measurement techniques. As an alternative to the current system, this study introduces a novel framework of a fully automated tracking system using archived Kinect audio visual data for recognizing workers, construction activities, and location aware information. Further, this information will be used to estimate tool time and performance of each worker on a jobsite.

Chapter Three: **Research Design, Methodology, and Theoretical Framework**

This chapter elaborates on the research problem, design approach, theoretical frame work, layout, and all relevant research equipment, indicating parameters and advantages.

3.1 Introduction

Research design is often thought of as the detailed structure or the blue print of the research that defines the research layout. It integrates the research concept with methodology, measurements, and final research output in the most sensible manner. Research design formulates the relationship between individual research elements.

3.2 Research Problem

The key research findings of the preliminary phase of the research follow an extensive literature survey, on-site work process, and document analysis studies focusing on the construction industry in Alberta, and highlight a critical factor that contributes to poor productivity performances: Namely that there was no structured approach or standard method (widely accepted and followed by the construction industry) to automatically and simultaneously perform activity recognition of multiple, interacting construction workers in order to obtain productivity-related worker tool time and performance data.

It is a well-known fact that construction is considered a labour-intensive process and labour productivity is recognized as a very important factor driving the industry to higher profitability. Knowing actual, realistic worker tool time supports beneficial business decisions. Therefore, a new, structured methodology to capture construction

worker tool time data sets, while achieving low operation cost and effort for the system, is a dire necessity.

3.3 Related Secondary Problems

There are many minor or secondary issues associated with this key problem, and each issue needs close attention. The tools and techniques developed during the research primarily address these individual issues, which have been identified on construction sites.

- What type of tracking technologies can be used for worker recognition and differentiation? What level of features should be extracted to precisely detect workers under dynamic environments (i.e. dynamic movements, lighting levels, etc.)?
- What type of tracking technologies can be used for construction activity recognition?
- What type of construction activities would be tracked in a work site environment?
- What level of accuracy we can expect from the worker tracking system, and activity recognition and location prediction systems?
- How can we link recognized workers and detected activities (e.g. proximity of 3D positions, direction of arrival of tool sounds, or other methods) and up to what level of error values can be accepted?
- Up to what level of background noise will be allowed in order for the system to generate acceptable results?

- Can this system be developed as a low cost solution while achieving the highest precision level?
- What level of accuracy can be generated from the worker tracking system, activity recognition system, and integrated tool time and performance measuring system?

These were a few possible questions associated with the key research problem that we need to have a clear focus on when developing the research methodology and design. Investigation of possible solutions and workarounds for the above questions lead to the formulation of the content and the direction of the research.

3.4 Hypothesis

The research hypothesis we are going to test with the implementation of the proposed research outputs can be formulated considering the research design and our primary and secondary objectives. The hypothesis can be broadly formulated as the proposed worker tracking and activity analysis system that accurately measures and interprets the real world tool time performance at the site. The null hypothesis against which the research hypothesis is going to be tested can be termed as follows:

A systematic examination of multiple modalities (i.e. RGB, depth and multi-channel audio) using a combination of signal processing techniques and statistical approaches will not provide a direct way of tracking workers, recognizing construction activities and finding location aware information of workers and activities.

The null hypothesis can be tested against the alternative hypothesis which is,

A systematic examination of multiple modalities (i.e. RGB, depth and multi-channel audio) using a combination of signal processing techniques and statistical approaches will provide a direct way of tracking workers, recognizing construction activities and finding location aware information of workers and activities.

3.5 Theoretical Framework

The theoretical frame of a research design always discusses the logic of the research.

The overall formation of the research is depicted in Figure 3.1 and provides the holistic approach of the research in a more compact view.

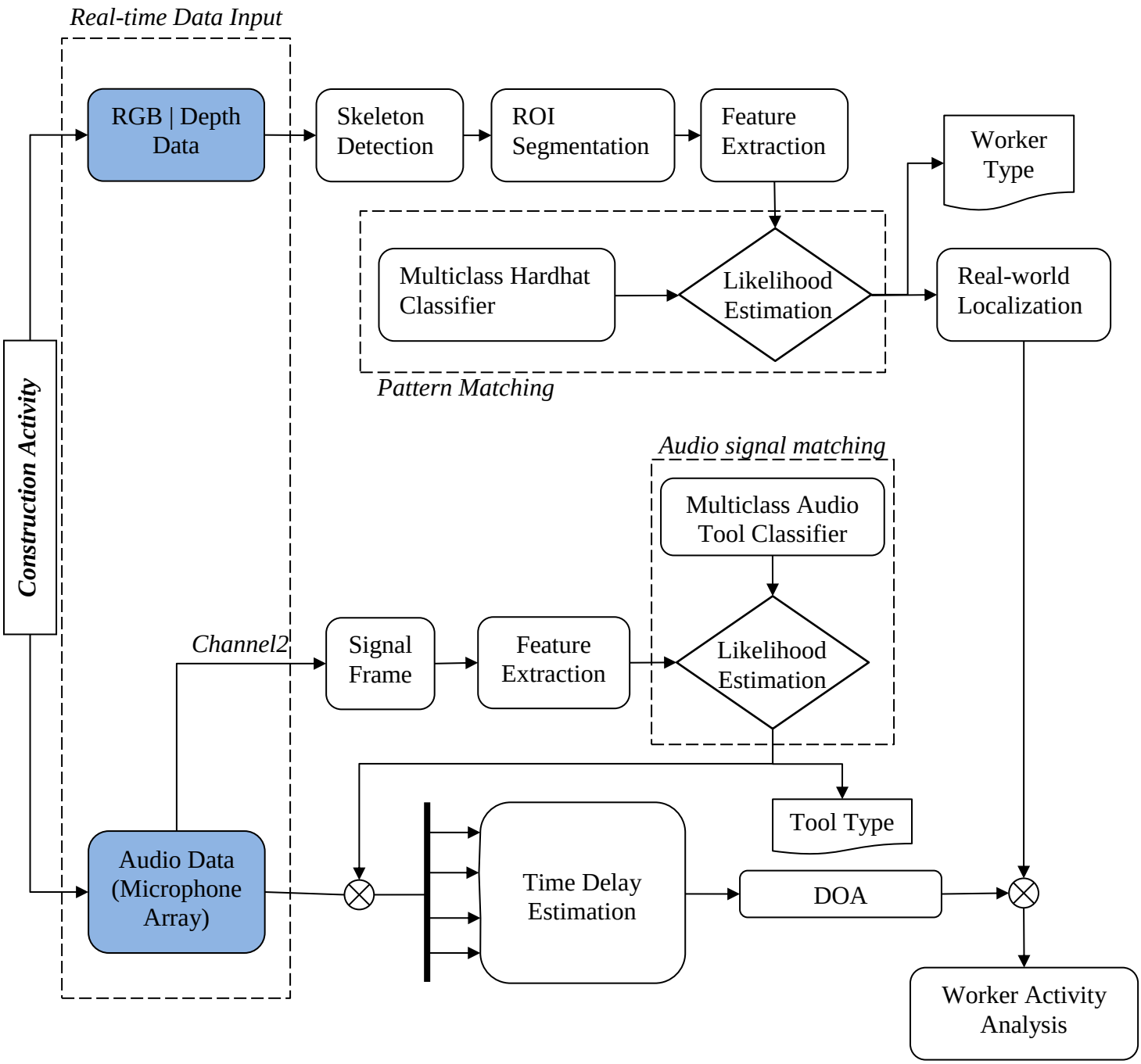


Figure 3.1: System process flowchart

Figure 3.1 shows the connection of interrelated concepts and the logical sense of the relationships between the variables and factors that have been deemed relevant/important to the problem. The proposed system mainly uses multi-channel audio, RGB video, and depth data to extract on-site information and it consists of two major components for data extraction and processing. We proposed the Microsoft Kinect device to extract these audio and video data. Different image processing and signal processing techniques have been used throughout the research study to accomplish various tasks. A detailed description of the properties and capabilities of the Kinect device will be further discussed in the next section.

As a universal safety precaution, all site personnel wear a construction hardhat when they are on site. Further, most construction companies use colour-coded hardhats for the site crew. The colour of the hardhat may vary based on the job type (i.e. yellow for labourers, red for supervisors, etc.). In this research we considered the "construction hardhat" as an object that represents the presence of a person at any given time. Hence, we proposed the hardhat as the key object for tracking workers on site. Moreover, special features of the hardhat such as its unique shape, high visibility, and established colour code system supported the object selection and increased the robustness of the proposed system. The study proposes image processing techniques to classify construction workers, while the audio classification process detects pre identified construction tool sounds on the job site. A likelihood estimation approach will be applied for both hardhat and audio tool classifiers. 3D location, the direction of arrival (DOA) of sounds, and recognition of tool sounds will be further analyzed for the activity breakdown process.

3.6 Data Analysis Methods

Data analysis is a process of transforming and modeling data to obtain useful information to derive the goal. Data analysis has multiple facts and approaches, encompassing diverse techniques under a variety of names, in different areas. In this study we mainly use image processing and signal processing data analysis methods, which will be defined in detail in the following chapters. We also use some statistical data analysis approaches to measure likelihood estimations such as logistic regression analysis, which is discussed in the next section.

3.6.1 Logistic regression

Logistic regression, being well suited for analyzing dichotomous outcomes, has been increasingly applied in many different areas. Further, this method is suitable for studying the relationship between a categorical or qualitative outcome variable and one or more predictor variables. We propose logistic regression analysis into our study in order to detect construction hardhats and to classify construction tool sounds in the noisy environment. The logistic model predicts the logit of Y from X. The logit is the natural logarithm of odds of Y. The simple logistic model can be expressed as:

$$\ln(\frac{\pi}{1-\pi}) = \log(odds) = \alpha + \beta x \qquad (2)$$

Hence, π= Probability (Y=outcome of interest |X=x):

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad (3)$$

where, π is the probability of the outcome or the event, under variable Y, α is the intercept of Y, and β is the slop parameter. X can be categorical or continuous, whereas Y is always categorical.

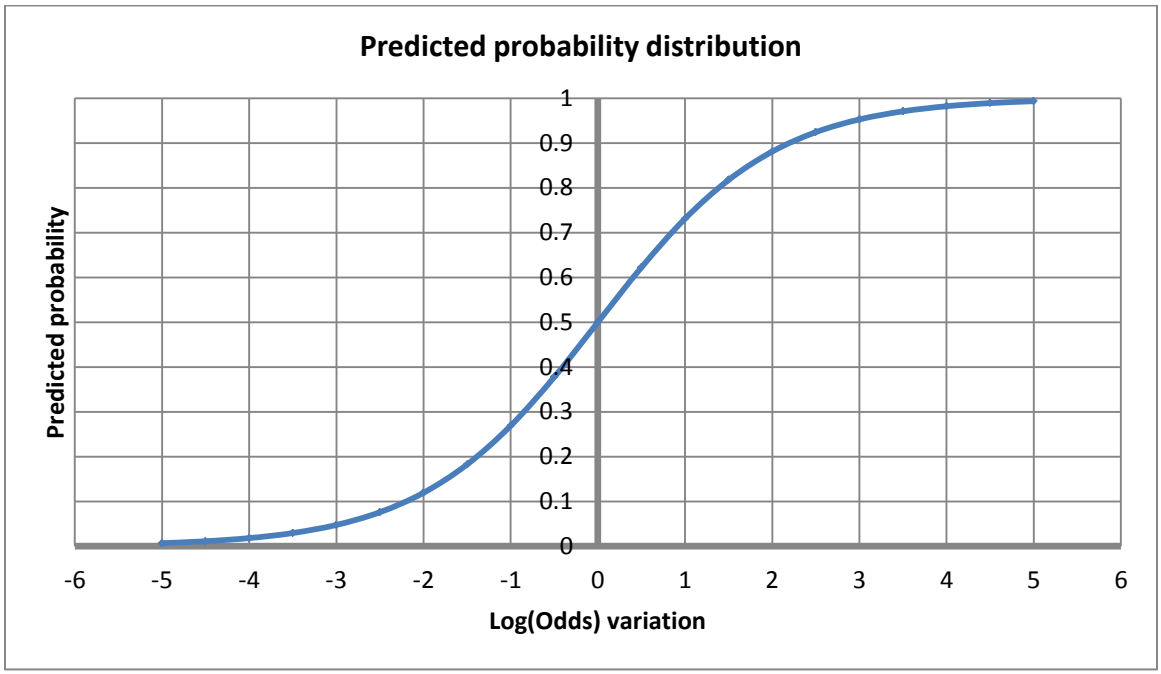


Figure 3.2: Predicted probability against Log (odds) variation

The regression line is monotonically increasing if β>0 while monotonically decreasing if β<0. The function takes on the value of 0.5 at the log (odds) ratio is zero or x=-α/β. Goodness-of-fit statistics assess the fit of a logistic model against the data. The model significance proposed by the Hosmer and Lemeshow (H-L) test and Nagelkerke R squared are used to assess the model fit to the data. We used two types of classification tables to show the validity of the predicted probabilities, such as the prediction success tables and the histogram of predicted probabilities. Receiver operating curve (ROC) is then applied to determine the optimum level for the cut-off value, which is a trade-off

between true positive rate (TPR) and false positive rate (FPR). Further details on these will be discussed in next chapters.

3.6.1.1 Modelling strategy

The following steps have been recommended for the logistic regression analysis (Peng, So, Stage, & St. John, 2002):

1. Perform a descriptive analysis of each predictor and its relation with the outcome variable. Initial analysis results provide much insight into potential variables for a feasible model.
2. Properly transform categorical predictors by a set of design variables by identifying unique features of different ranges.
3. Correctly identify the outcome event and model its probability by a series of univariate logistic regressions.
4. Based on univariate analysis, fit a multivariate logistic model using all predictors that are of importance.
5. Fit alternative models to data.
6. Compare the performance of alternative models in terms of statistical significance of predictor, goodness-of-fit statistics (Hosmer and Lemeshow test & Nagelkerke R squared), accuracy of prediction, and correlation of covariates. The superior model surpasses competing models in more areas than one.

The decision to accept one model to be the best fit model among its competitors should be justified with the above mentioned parameters and the sense of outcome.

3.7 Sampling

Sampling is the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population so that by studying the sample we may equitably generalize our results back to the population from which they were chosen. Since the main objective is to develop models to detect hardhats and pre-defined tool sounds, a population for each model can be defined as including all concerned types of items with the characteristics.

Advantages of the sampling method are:

1. Improves the accuracy/efficiency of estimation and overall validity of the research
2. Focuses on important subpopulations and ignores irrelevant ones

A poorly selected sample does not correctly represent the population and will have biases and hence impact the final results, constituting a common threat to the external research validity.

3.7.1 Sampling procedure specifications

During the research, the systematic sampling method was used to collect a sample that reflects the makeup of the population. Thus we arranged the target population according to an ordering scheme and then selected elements at random intervals through that ordered list. For instance, random observations were made at random time intervals, equally representing the complete depth from the camera to the hardhat under different lighting conditions to cover all possibilities that can occur in a real scenario.

3.7.2 Sample size

Sample size is another important aspect that determines the significance of the constructed model. Higher sample sizes increase the accuracy of the model. In terms of the adequacy of sample sizes, the literature has not offered specific rules applicable to logistic regression (Peng et al., 2002). However, several authors on multivariate statistics (Lawley & Maxwell, 1962; Marascuilo & Levin, 1983) have recommended a minimum observation/predictor ratio of 10 to 1, with a minimum sample size of 100 or 50, plus a variable number that is a function of the number of predictors.

We collected 250 sound samples for the audio classification model out of which 20% of the responses are positive for each tool. Meanwhile 100 image samples were extracted for the hardhat classification model, which comprised 65% of positive responses.

3.8 Data Collection

Data collection at the mathematical model construction stage and validation stage helps to prove the hypothesis formulated at the design stage. This validation-related data collection confirms the research concept and validates the ideas and concept proposed by the research. It is equally important to define the data collection methods, and to select the samples to satisfy the statistical requirement to represent the population. Knowing the final data analysis methods that are going to be used in the data analysis phase improves the selection of correct data collection methods, frequency, and the type of data to be collected. Both qualitative and quantitative type data were collected, mainly in the two

phases we have already discussed, that is, the preliminary research phase and the implementation or testing phase. It is also acknowledged that any qualitative data can be defined and manipulated numerically while quantitative data can also be based upon qualitative judgments. Data set preparation, construction, and validation for each model will be further reviewed in the subsequent chapters.

3.9 Microsoft Kinect

Low-cost range sensors are an attractive alternative to expensive laser scanners in application areas such as indoor mapping, surveillance, robotics, and forensics. A recent development in consumer-grade range sensing technology is Microsoft's Kinect sensor (Microsoft, 2012). Kinect was primarily designed for natural interaction in a computer game environment. However, the characteristics of the data captured by Kinect have attracted the attention of researchers from the field of mapping and 3D modelling. In this research study we use the Kinect device as the primary data extraction device. This is a low cost device ($100-150) that provides various sensor data streams such as colour, depth, 4-channel audio, infrared, and accelerometer data of the scene at a higher rate of frequency.

3.9.1 Kinect as a 3D measuring device

Kinect is a composite device consisting of an infrared projector of a pattern and infrared camera, which are used to triangulate points in space. It works as both a depth camera

and a colour (RGB) camera, which can be used to recognize image content and texture 3D points (see Figure 3.3).

Figure 3.3: Kinect sensor composition

As a measuring device, Kinect delivers five sensor data streams: infrared image, RGB image, depth image, 4-channel audio information, and accelerometer information of a scene in a higher rate of frequency. We propose Microsoft Kinect for Windows SDK version 1.6. This further provides skeleton tracking information, which is a technology breakthrough for human recognition systems.

3.9.2 Technical specification

Table 3.1 lists some important technical specifications of the Kinect device. Figure 3.4 and Figure 3.5 illustrate the physical capabilities of the Kinect device. These figures were extracted from Kinect for Windows human interface guidelines 1.5 (Microsoft, 2013).

Table 3.1: Technical specification of the Kinect sensor

Property	Specification
Field of View (Horizontal, Vertical, Diagonal)	57.5° H, 43.5° V, 70° D
Depth image size	VGA (640x480)
Spatial x/y resolution (@ 2m distance from sensor)	3mm
Depth z resolution (@ 2m distance from sensor)	1cm
Maximum image throughput (frame rate)	60fps
People recognition range	0.8m – 4.0m
Audio: built-in microphones	4 microphones
Power consumption	2.25W
Tilt range	-27 to +27 up & down
Operation environment (every lighting condition)	Indoor

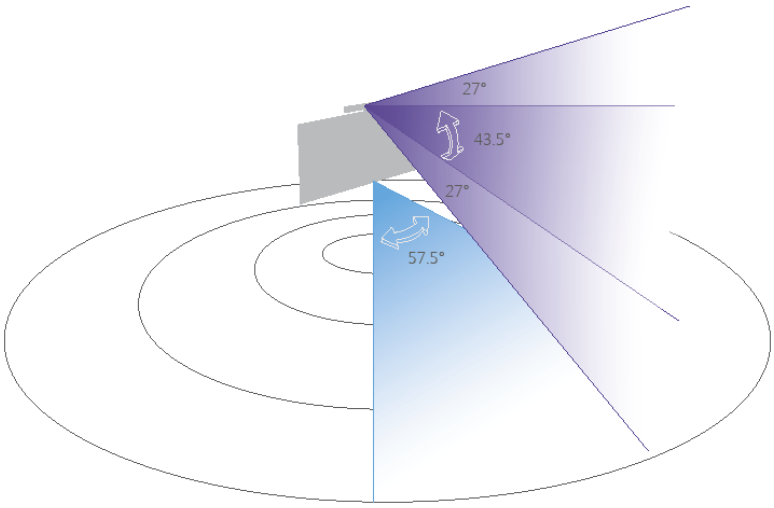


Figure 3.4: Kinect horizontal and vertical viewing angles (Microsoft, 2013)

As depicted in Figure 3.5, Kinect SDK supports people recognition and skeleton recognition up to 4m distance from the camera. However, the SDK supports extended depth data beyond the physical limits for other objects in the environment at reduced resolution and accuracy.

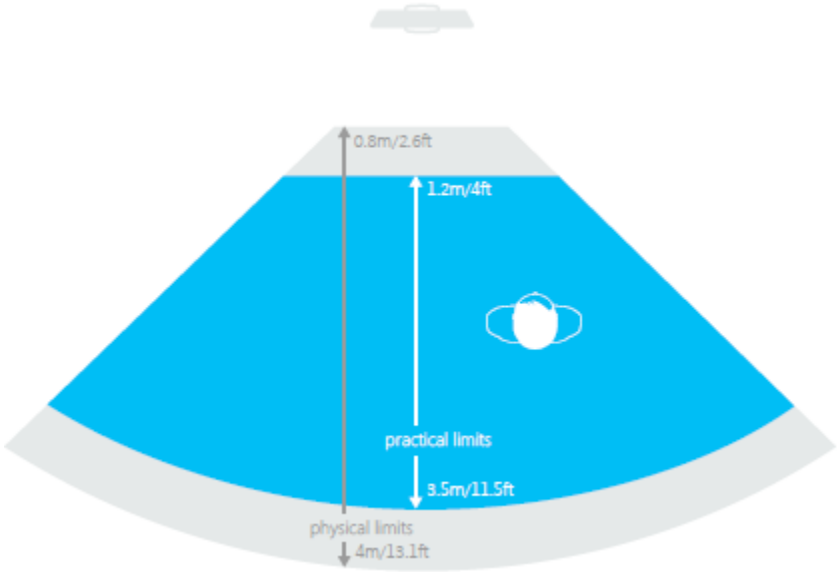


Figure 3.5: Kinect depth range limitation (Microsoft, 2013)

Kinect depth map is generated using an infrared structured light array system. Therefore, environments that have large amounts of natural light will make depth tracking less reliable. Given this, Kinect is well suited for monitoring indoor construction sites which has appropriate lighting conditions for the Kinect infrared depth camera. Generally, construction workers wear much more bulky clothing than typical individuals. Wearing or holding these items that drastically change the shape of a human may confuse skeletal tracking. Also, items or clothing material that is reflective will interfere with the IR reflection and make skeletal tracking less reliable. Further, using extreme tilt angles makes tracking less reliable.

Chapter Four: **Construction Worker Tracking System**

This chapter elaborates on the literature study of automated worker tracking systems, analysis for a technology selection, step-wise worker tracking procedure, hardhat classifier construction and camera calibration.

4.1 Introduction

In order to understand worker behavior and activities for improving the productivity of construction site operations, it is necessary to analyze observations of construction work in progress. For example, one way of improving current work practices is by observing work tasks and generating manual evaluations. This practice is commonly known as 'work sampling' (Borcherding, 1976). Location-aware computing offers significant potential for improving such manual processes and supporting important decision making tasks in the field.

Burrell and Gay (2001) discuss context-aware computing, which uses environmental characteristics such as the user's location, time, identity, and activity to inform the computing device so that it may provide information to the user that is relevant to the current context. Context-aware computing can thus potentially enable mobile users (e.g. construction inspectors, firefighters) to leverage knowledge about various context parameters such as their identity, current task, and location to ensure that they get highly specific information pertinent to the decisions at hand (Schilit, Adams, & Want, 1995).

Any technology that can reliably, accurately, and automatically record the location of construction resources for work sampling could significantly simplify previously conducted manual assessments and improve confidence in the measurements. Likewise, technological systems that track project critical resources (e.g., people, equipment, material) and provide information on resource utilization can enhance current work practices. The existence of a context-aware system in construction that tracks the location of construction resources, and identifies and measures the status of work tasks, would improve project performance.

In recent years, the need for tracking and localization has been rapidly expanding in many fields and the construction industry has also shown significant potential for use of these applications. Information technologies can benefit the industry by automating and integrating some of its predefined tasks/work functions common to most projects.

4.2 Literature Survey

This section provides an introduction to the automated systems that have been developed for construction worker performance related subjects in the construction field. Further, available techniques, concepts, and theories for tracking and localizing human figures and objects have also been reviewed.

4.2.1 Worker and equipment localization techniques used in the construction industry

In the construction field, worker performance monitoring is an essential part of a project as it assists project managers in formulating strategies and making decisions regarding

man-hour resource allocation in order to keep the project on track. There have been several techniques used for construction worker monitoring and performance measuring.

Many existing technologies for localization and tracking fall within the broader category known as sensor networks (SNs) or wireless sensor networks (WSNs). Radio-frequency identification (RFID) tags, ultra wide band (UWB) sensors, global positioning systems (GPS), and embedded sensor systems are the leading technologies. Sensor networks consist of a collection of sensing nodes used to compute position from location-based measurements via triangulation. When a resource is tagged with an electronic tag capable of generating the necessary signals, a sensor network provides location information of the tagged resource. The four predominant location-based variables of a wirelessly transmitted signal are the received-signal-strength indicator (RSSI), the angle of-arrival (AOA), the time-of-arrival (TOA), and the time-distance-of arrival (TDOA). Given measurements of one of these variables by a collection of distributed sensor nodes, triangulation leads to estimation of the associated signal source position. Apart from the SNs, range camera technology and computer visioning techniques have also been applied for localization and tracking. The following sections discuss the nature of SNs and their application in construction and non-construction fields. Advantages and disadvantages of these techniques have also been reviewed.

4.2.1.1 Radio frequency identification (RFID) technology

RFID technology is used to track and locate objects and people in a pre-defined physical space. The technology consists of a RFID tag that is attached to the object being tracked

and a RFID reader that identifies the tag when it is within the range of the reader. An RFID tag transmits a signal through radio waves that in turn are received by an RFID reader. Further, received RFID signals are triangulated and the location of the RFID tag transmitter can be determined.

A few years ago, researchers and industry professionals explored and demonstrated the application of RFID tags in construction (Goodrum, McLaren, & Durfee, 2006; Grau, Caldas, Haas, Goodrum, & Jie, 2009; Jaselskis & El-Misalami, 2003; W. Wu et al., 2010). This technique of monitoring RFID tags has also been used to track workers, equipment, and construction materials on a job-site (Sattineni & Azhar, 2010). Recently building information modelling (BIM) technology has emerged as the industry standard in the architecture, engineering, and construction (AEC) sector. Sattineni and Azhar (2010) carried out a research study and combined the two technologies to monitor the movement of RFID tags in a BIM environment. The research study also suggested that using this method will improve the safety and productivity of construction workers on a construction job site.

Pros and cons are widely discussed in the application of RFID technology in the construction field (Kiziltas, Akinci, Ergen, Pingbo, & Gordon, 2008). Although RFID has some attractive features, such as non-line-of-sight readability, low cost and compactness, and contactless communications (Lim, Choi, & Lee, 2006), several drawbacks of human tracking have also been identified. A construction job site is considered a dynamic environment since it builds up day by day, and the region of interest for tracking equipment and workers changes rapidly. In these circumstances it is very difficult to

implement an RFID system because it needs significant infrastructure, regular installation, and maintenance. Generally, an average worker's effective work area on site spans a fairly large area, hence active RFID tags have to be used because passive RFID transmits signals only a short distance. Nowadays, improvements of power and signal strength of RFID are popular research areas. However, increase in strength also increases the size of the RFID tag, and this will be a critical deciding factor since RFID tags are used in worker tracking by embedding tags into associated equipment (i.e. hardhat, construction boot, wristband, etc.). Another disadvantage of RFID is performance reduction in proximity of metals and liquids, in particular when used at higher frequencies (Kiziltas et al., 2008). In addition, the signal coverage of the RFID tags drops to 20% - 25% of the nominal reading range in open air environments (Kiziltas et al., 2008).

4.2.1.2 Ultra wide band (UWB) technology

UWB technology has also been adopted in the construction research field for evaluating real-time resource location tracking of workers, equipment, and materials in outdoor and indoor environments (Cheng, Venugopal, Teizer, & Vela, 2011; Cho, Youn, & Martinez, 2010), including construction (Teizer, Lao, & Sofer, 2007).

This has a similar arrangement to a RFID network system: a tag is attached to the object that requires location tracking. As each tag emits a UWB signal, location is calculated using both the time difference of arrival between different sensors (the

receivers) and the angle of arrival at each sensor. Each sensor employs a minimum of four UWB receivers, which allows the angle of arrival to be determined.

Compared to other localization technologies like RFID or ultrasound, UWB has shown to possess unique advantages including: longer range, higher measurement rate, improved measurement accuracy, and immunity to interference from rain, fog, or clutter (Cheng et al., 2011). Cheng et al. (2011) has carried out a case study of mobile resource tracking for analysis of work site operations using commercially available UWB technology in construction environments. The work demonstrates benefits of UWB technology and the applicability of UWB for the design of construction management support tools. Teizer, Venugopal, and Walia (2008) demonstrated how the UWB wireless sensing technology is capable of determining three-dimensional resource location information in cluttered construction environments to accomplish the work.

4.2.1.3 Global positioning system (GPS)

A Global Positioning System (GPS), being a satellite based navigation system, works very accurately outdoors but lacks support indoors and in congested areas. GPS based tracking systems are widely used in vehicle tracking and delivery monitoring systems in different fields.

For outdoor applications, positioning techniques have been investigated and validated in recent work reported in Behzadan and Kamat (2007). GPS technology is applied in the construction field for developing augmented reality (AR) platforms to monitor the visual progress of a job site (Khoury & Kamat, 2009b). However, GPS

technology is not suitable for indoor applications because it becomes highly ineffective. Similar to the RFID system, indoor GPS technology that consists of transmitters and receivers has been proposed in indoor construction for tracking mobile users (Khoury & Kamat, 2009c). The highlighted drawbacks of this technology are the higher infrastructure cost, regular operational cost, and range limitation when it comes to implementation on construction job sites.

4.2.1.4 Wireless Local Area Networks (WLAN)

The WLAN supports network communication over short distances using radio or infrared signals instead of traditional network cabling. This economical solution provides convenient connectivity and high speed links, and can be implemented with relative ease in software (Hightower & Borriello, 2001). Additionally, the WLAN sensing range covers a large area; the range of a typical WLAN node is about 100m (Wang & Liu, 2005) and is not restricted by line of sight issues. A WLAN can support a large number of nodes and vast physical areas by adding access points to extend coverage. Khoury and Kamat (2009a) carried out a case study of tracking workers in an indoor work environment using a WLAN based position system called the Ekahau Positioning Engine (Ekahau, 2012). Although the sensing range covers a large area, the accuracy of localization of this method does not provide a satisfactory resolution, which ranges from 1.5m to 2m.

4.2.1.5 Range camera technology

Low-cost range sensors are an attractive alternative to expensive laser scanners in application areas such as indoor mapping, surveillance, and robotics. A recent development in consumer-grade range sensing technology is Microsoft's Kinect sensor (Microsoft, 2012). Kinect was primarily designed for natural interaction in a computer game environment. However, the characteristics of the data captured by Kinect have attracted the attention of researchers involved in people tracking (Xia, Chen, & Aggarwal, 2011), action/pose recognition (Escorcia et al., 2012; Shotton et al., 2011; Sung et al., 2011), 3D reconstruction (Izadi et al., 2011), and mobile robot navigation (Benavidez & Jamshidi, 2011) in several fields. The Kinect consists of RGB camera, infrared depth camera, and four microphone array systems, which allow developers to use RGBD and audio inputs from the environment.

Escorcia et al. (2012) have developed a construction worker pose and action detection system using the Microsoft Kinect sensor. This worker-action detection system is based on machine learning techniques and discriminative classifiers. The system was tested for drywall construction and the experimental results showed that the method achieves an average precision of 85.28 percent.

Stereo cameras, and light detection and ranging (Li-DAR) devices and time-of-flight cameras such as SwissRanger SR4000 (MESA, 2013) are also used as range imaging techniques to find depth information of an image.

4.2.1.6 Video camera

Tracking construction workers and equipments using video cameras can make work sampling more objective by automatically recording and reviewing the performance of selected work tasks. Yang et al. (2010) carried out a research study and developed a supervised tracking algorithm that requires the user to manually identify the construction workers, and subsequently a machine-learning algorithm learns and tracks the target. Gong and Caldas (2010) presented a study on developing a video interpretation model to interpret videos of construction operations and automatically convert that into productivity information. It focused on the detection and tracking of project resources as well as work state classifications and abnormal production scenario identifications. Most related to this paper is the work in Cordova and Brilakis (2008), who detailed a static-camera, stereo-based tracking method for construction work sites.

Human activity recognition has been previously studied by a number of different authors. In this case, space time features are generally used to model points of interest in video (Dollár, Rabaud, Cottrell, & Belongie, 2005). Several authors have supplemented these techniques by adding more information to these features (Jhuang, Serre, Wolf, & Poggio, 2007; J. Wu, Osuntogun, Choudhury, Philipose, & Rehg, 2007). Moreover, Heydarian, Golparvar-Fard, and Carlos Niebles (2012) developed an automated visual recognition of construction equipment actions using spatio-temporal features and multiple binary support vector machines. In addition, some authors have designed activity recognition using filtering techniques (Rodriguez, Ahmed, & Shah, 2008), and sampling of video patches (Boiman & Irani, 2007).

Peddi et al. (2009) developed a human pose detection system using blob tracking, pose extraction, and pose classification to judge the productivity of the worker by classifying different poses into categories such as effective, ineffective, and contributory. The assumption is that workers are always moving and each pose belongs to a certain type of action. The simple assignment of each static pose to an action can be a major limitation when it comes to activity analysis. The key weakness of the current state of the technique is this can be used only for selected construction activities that need to have unique human work poses to recognize the performance. Similar to this study, attempts have been made to develop a framework for integrating posture analysis of workers and a predefined set of rules to categorize work tasks as ergonomic or non-ergonomic (Ray & Teizer, 2012).

4.3 Selection of tracking technique for the research study

As the preliminary literature reviews and process studies precisely indicate, a variety of sensors and sensing technologies with automated tracking capabilities are available for use in construction and infrastructure projects. Arguments for automated, remote tracking technology in construction are to increase tracking efficiency, to minimize the impact to current construction work environments, and to reduce labour costs. However, implementation costs associated with each technology add further constraints (Teizer, Caldas, & Haas, 2007), so this must be weighed against the potential benefits.

To be of interest to the construction industry, the tracking technology should meet as many of the following criteria as possible (Cheng et al., 2011):

- Cost and maintenance: Low implementation and maintenance cost for all the devices and infrastructure used in the job site.
- Reliability: Capable of accurately and precisely recording the activities that are associated with monitored work tasks.
- Data update rate: High data frequency provided in recording.
- Social impact: Less invasive technology, minimal impact to the construction site environment by adding devices.
- Device form factor: Small enough to fit in site environment or on any asset (as needed) without interrupting the completion of work objectives.
- Scalability: Robust in a variety of site layouts (open, closed, and/or cluttered space(s), and small to large spaces).

Existing research into low cost range cameras in construction applications has attracted attention and has focused on evaluating real-time resource location tracking of workers, equipment, and materials (Escorcia et al., 2012; Weerasinghe, Ruwanpura, Boyd, & Habib, 2012). Recent research has shown that the use of Microsoft's Kinect device (a low cost range camera in the current market) in construction offers a solution that meets the aforementioned requirements. Compared to sensor network technologies like RFID, UWB, or GPS, the Microsoft Kinect device has shown to possess unique advantages:

- video (30 fps) provides a rich context for late productivity analysis
- built-in microphone array system supports audio recognition and beam forming
- infrared camera provides depth information under various lighting conditions

- low implementation and maintenance cost and minimal infrastructure needed
- recorded video can be used for mitigating disputes and for training purposes

Requiring line of sight in video monitoring was the main limitation of video-based methods as compared to sensor networks. However, Katz, Saidi, and Lytle (2008) designed a multi-camera based system and used video footage from different angles to overcome the line of site and occlusion issue.

Table 4.1: Summary of indoor positioning (Khoury & Kamat, 2009a)

	Position Uncertainty	**Range**	**Calibration**	**Deployment and cost**
Indoor GPS	Low (1-2cm)	60m	Needed	Quite easy but very expensive
UWB	Medium (0-50cm)	10m	Not needed	Quite easy but very expensive
WLAN	High (1.5-2m)	10-100m	Needed	Easy and economical
Kinect	Very low (3mm)	0.8-4.0m	Not needed	Easy and economical

Nevertheless, video data processing and interpretation, which is still an arduous manual process, or at best a computer-assisted manual reviewing process, is the real barrier to wide application of video in construction. In this research we focus on developing an automated video interpretation addressing issues in manual data collection methods. Hence, Microsoft's commercially available, reasonably accurate, audio integrated Kinect device has been used in the research study.

4.4 Worker Tracking Using Kinect Skeleton Figures

The Microsoft Kinect sensor is one of the most anticipated devices and technologies with the potential to transform processes across the engineering and construction industries. Although the potential of Kinect is real, it does have limitations like any other technology. Without understanding and working with the limitations of Kinect, this technology may disappoint many before its true and significant capabilities are realized. In this section a novel method for reliable recognition of construction workers using colour and depth data from a Microsoft Kinect sensor is discussed. This proposed algorithm is based on skeletal tracking and colour segmentation and combines it with a powerful logistic regression analysis classifier, in which meaningful visual features are extracted based on the pre-defined rules to achieve accurate recognition.

Since hardhats are mandatory in the construction field, and since most contractors use colour coded hardhats to designate site personnel (e.g. labourer – yellow hardhat, supervisor – red hardhat, etc.), it is ideal to use hardhats as tracking objects to represent people on site (Weerasinghe et al., 2012). Unlike other existing methods, this allows for differentiation of people by detecting a single object and significantly reduces the computational cost and support requirements to achieve better performance in a real-time system.

4.4.1 Worker tracking framework

Figure 4.1 provides an overview of the proposed worker tracking system. The first key component of our system is the use of a Kinect sensor for capturing colour and depth

information of the scene. We used a set of MATLAB executable files (i.e. mex files) developed by Chikamasa (2012) to extract raw data files from the Kinect device. The human tracking and skeleton detection are more reliable on this approach since this algorithm uses depth images to extract undistorted silhouettes of human figures.

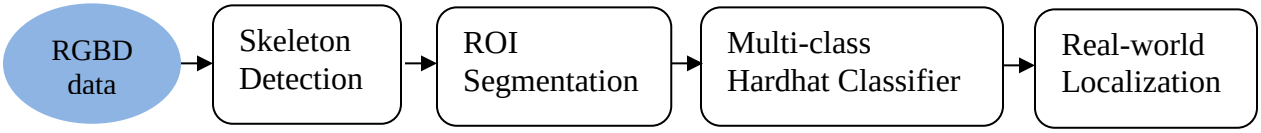


Figure 4.1: Overview of the worker recognition process

The second important component is the new visual representation for worker recognition and class differentiation. The key observation is that the typical hardhat of a person on a jobsite can be characterized by visual features of the hardhat. We encode this observation by adopting a logistic regression model classifier. The Kinect enables to determine the 3D coordinates (i.e. with respect to the Kinect device) for the skeleton figure based on the depth map information. In order to transform these Kinect coordinates into building coordinates which is used by surveyors and drawings in the project life cycle, the rigid body transformation algorithm is applied. For this purpose, camera calibration and single photo resection are followed. We exploit this powerful representation to build highly discriminative human recognition and differentiation classifiers. Finally, experimental results show that the proposed algorithm is able to classify worker recognition with a higher accuracy.

4.4.2 Skeleton detection

The first goal of our system is to track human figure silhouettes in the scene. The default skeleton tracking libraries provided by the Kinect SDK (Microsoft, 2012) tracks skeleton data that consists of 20 skeleton joints for each figure. In the current system, MATLAB can extract a maximum of two skeletons at a time. Figure 4.2 provides the detailed description of a skeleton extracted from the Kinect library. This research study mainly focuses on the upper body joints for the analysis. In particular, head and shoulder center joint are used to filter out the head interactive zone of the figure. A detailed review is provided below.

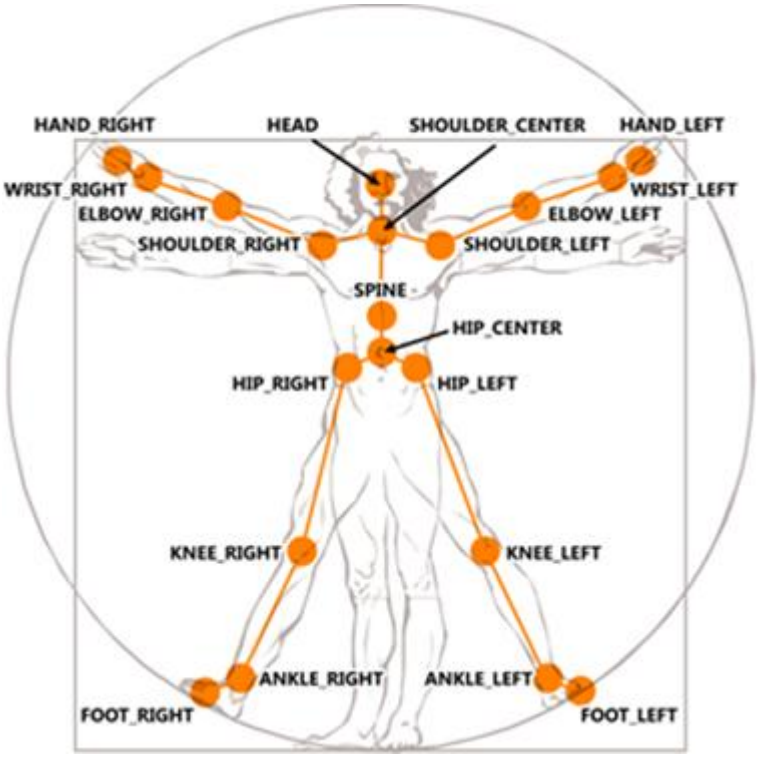


Figure 4.2: The 20 joints that make up a Kinect skeleton (Microsoft, 2012)

4.4.3 Region of interest (ROI) segmentation

In tracking terminology, defining and extracting the region of interest (ROI) is an important topic and accurate ROI reduces the computational cost and improves the robustness of the algorithm. Figure 4.3 illustrates the high level step-wise structure of the ROI extraction process. The goal of our system is to locate and track workers in the scene. For this purpose, we segment the workers from the background clutter and focus the figure and the relevant worker movements. In our system we use the player map provided by the Kinect SDK library for capturing active people on the scene. The player map is an index map that allocates an index number for each player in the image. Analyzing this index, silhouettes of human figures can be extracted. Figure 4.4 shows the results of ROI process: silhouette extraction, head interactive zone extraction and color segmentation.

The second important component is filtering the head interactive zone combining the skeleton data and player index. The effective radius from the head, which has a length equivalent to the distance between the head and the shoulder center, is selected and extracted head region is shown in the left bottom corner in the Figure 4.4.

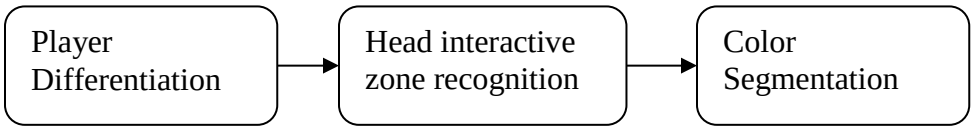


Figure 4.3: Structure of ROI selection process

In order to achieve the highest performance at a relatively low computational cost, the image colour segmentation (colour feature extraction) procedure is applied, which helps to detect four pre-identified, colour-coded hardhat shapes embedded in the image frame. The YCrCb space is used to differentiate the colour of the image. Y is the luma component and Cb and Cr are the blue-difference and red-difference chroma components. YCrCb is a practical approximation to colour processing and perceptual uniformity, where the primary colours corresponding roughly to red, green, and blue are processed into perceptually meaningful information. Furthermore, a colour can be selected using only two parameters ignoring luma component, while most of the colour spaces such as RGB need all three parameters to pick a colour code. However, alternative color spaces such as HSV and HSL can also be used for the color segmentation.

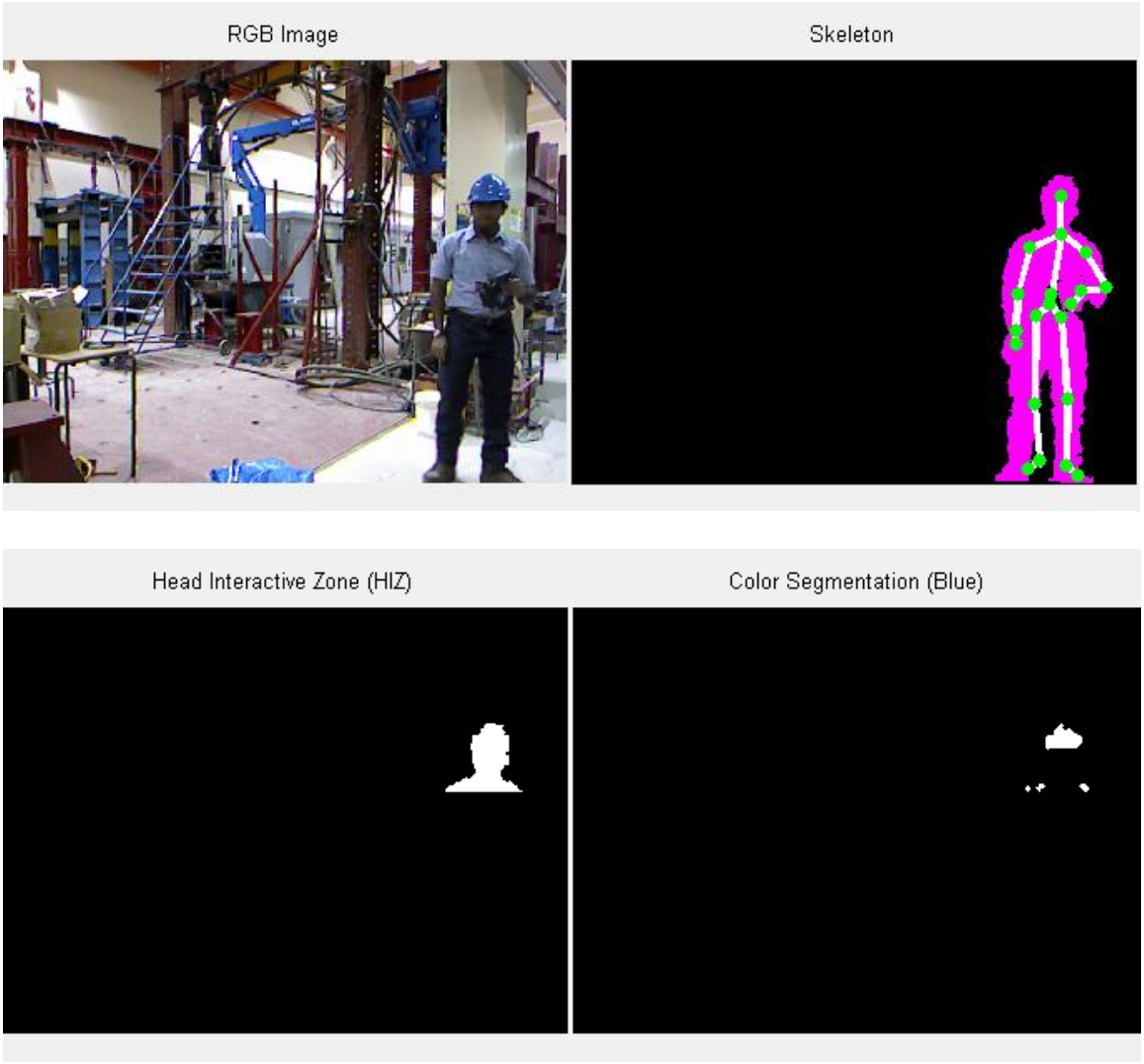


Figure 4.4: Results of ROI selection process

The colour parameter ranges for pre-identified colour coded hardhats (i.e. Red, Blue, Yellow and White) is shown in Table 4.2. These ranges were determined based on a trial and error study. It further allows users to change the range according to different colour specification requirements. The extracted four-colour silhouettes are referred to as

"blobs" and the shape parameters of these blobs are analyzed using a statistical approach as discussed in the following section.

Table 4.2: YCrCb colour ranges of selected 4 different coloured hardhats

	Y_min	Y_max	Cb_min	Cb_max	Cr_min	Cr_max
Yellow	60	235	16	70	135	170
Blue	75	235	138	240	16	119
Red	60	120	80	135	170	240
White	75	235	120	135	120	135

4.5 Hardhat Detection Classifier

A multivariate statistical model is developed for the prediction of construction hardhat in an image frame in the context of construction worker type assessment. Logistic regression was the preferred statistical procedure for this study because the technique is suited to models with a dichotomous outcome that results in hardhat or non-hardhat shape with multiple predictor variables that include a mixture of continuous and categorical parameters. A set of image features captured by different colour silhouettes are considered as the predictor variables for the logistic regression model. One of the reasons that logistic regression has become extremely popular is that if the logistic regression model is correctly specified, the parameters can be consistently estimated from a sample that is stratified by outcome y. A common study design is to obtain separate samples from the two outcome strata, cases (y=1) and controls (y=0), and compare these samples in terms of co-variables.

4.5.1 Test data

For this sampling analysis we collected a total of 100 random sample image frames from a recorded video of a moving worker with the following scenarios:

- Different poses of a worker with a coloured hardhat (i.e. red, blue, yellow, and white), changing the colour from time to time: 65 samples
- Different poses of a worker without hardhat: 35 samples

The video file is recorded using the Kinect device under a lighting level appropriate to the work space. Meanwhile, depth information and the skeleton information are also recorded in order to obtain the distance to the worker and head interaction zone. The following parameters of the blobs are considered predictor variables for the binomial logistic regression model (see Figure 4.5):

1. ***Eccentricity***: The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases; an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.) This property is a measure of roundness of the blob.
2. ***Blob area****:* This is a measure of the size of the blob.
3. ***Distance between blob centroid and skeleton head coordinate***: This is a measure of proximity of the blob to the head. A small distance has a higher potential to be a hardhat.

4. ***Distance to human figure***: According to the linear perspective rule, as objects become more distant they appear smaller because their visual angle decreases. The product of parameter two and four has a significant value to the model.

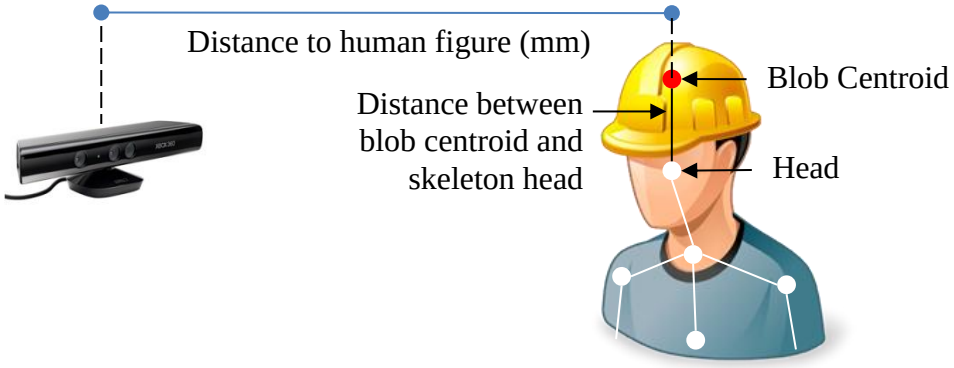


Figure 4.5: Predictor variables for hardhat classifier

4.5.2 Statistical analysis

Version 20 of the IBM SPSS Statistic software was used to analyze the logistic regression of this study. SPSS statistical software begins by conducting backward stepwise logistic regression, removing non-significant variables of that block before conducting backward stepwise logistic regression on the remaining variables.

With the model coefficients, the probability formula for hardhat occurrence in the image can be obtained. For the hardhat classifier model:

$$p(Hardhat) = \frac{1}{1+e^{-z}} \qquad (4)$$

where, $z = b_0 + b_1(valriable_1) + b_2(valriable_2) + \cdots + b_n(valriable_n)$, b_0 is the constant and b_1 to b_n are the corresponding parameter coefficients. Table 4.3 lists the variables and code used in the model.

Table 4.3: Predictor variables: Hardhat Classifier

Variable Name	**Code**
Eccentricity	ECC
Blob Area	AREA
Distance between blob centroid and skeleton head coordinate	D-CH
Distance to the human figure	D-FG

Classifier construction is based on five major criteria: Nagelkerke R squared value, model significance of Hosmer and Lemeshow (2000) test, significance level of each variable, variable correlation, and model overall accuracy.

Table 4.4 shows the output resulting from all the potential candidate predictor variables in the equation. The table also includes an additional combination of variables: the product of blob area and the square value of the distance to the human figure: AREADFG2. The accuracy level of all models tabulated below is based on the 0.5 cut off level. The optimum cut off level selection is determined using the ROC curve method after selecting the final model (Fawcett, 2006). Nagelkerke (1991) generalizes the definition of the coefficient of determination. R squared interpretation should be the proportion of the variation explained by the model. It should be between 0 and 1, with 0 denoting that the model does not explain any variation, and 1 denoting that it perfectly explains the observed variation.

The error is controlled by choice of a decision-making criterion, called level of significance of a variable. In this research we set p <0.10 to eliminate variables from the equation. In variable selection, the correlation value between variables should be smaller. In addition, higher model significance and the model overall accuracy reflect a better selection of variables.

An ideal modelling approach in logistic regression is to consider and contrast all models that are theoretically significant and practically important. As per Peng et al. (2002), we perform a descriptive analysis of each predictor and its relation to the outcome variable. Initial analysis results provide much insight into potential variables for a feasible model. Following figure illustrates the relationship of the selected predictor variables for the hardhat classifier and 3D scatter plot shows clear clusters for two scenarios (hardhat and non-hardhat events).

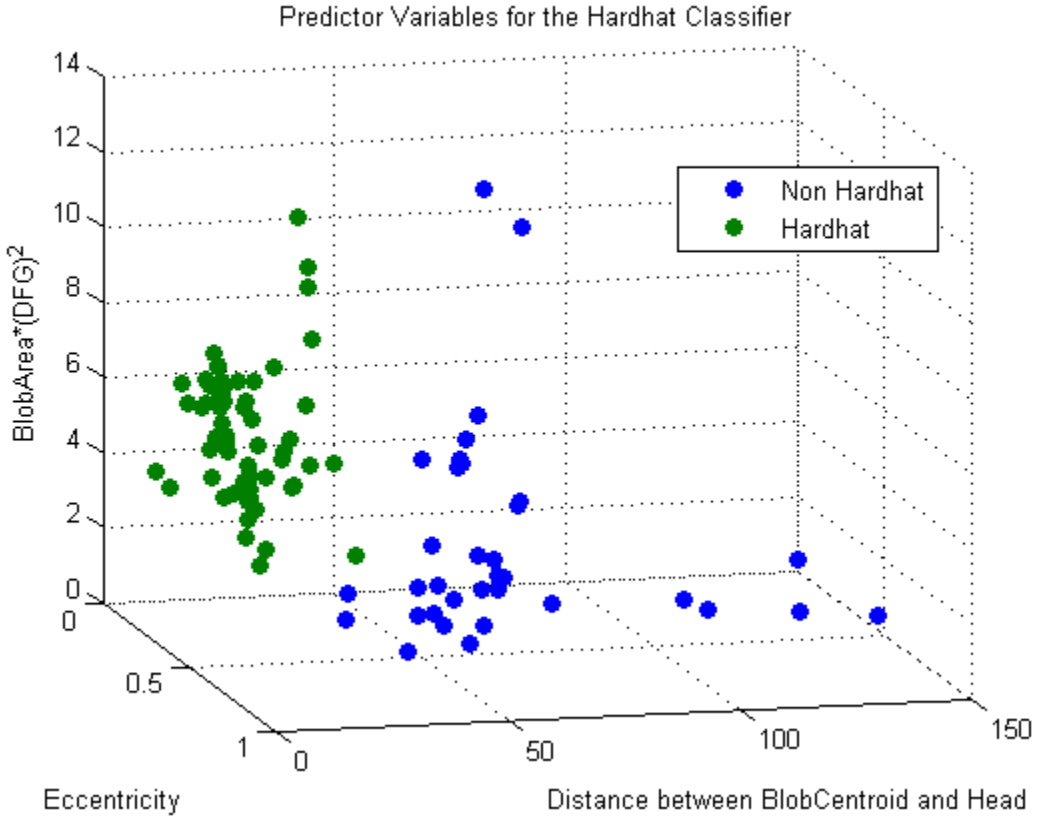


Figure 4.6: Relationship of predictor variables of hardhat classifier

Considering the relationship of predictor variables, the following alternative models have been constructed.

Table 4.4: Model properties

No	Variables added	R2	Model Sig.	Max Var Sig.	Max Corr	Accuracy %
1	ECC, DHC	0.859	0.903	0.000	0.307	92.0
2	ECC, AREA, DHC	0.935	0.472	0.016	0.307	98.0
3	DHC, AREADFG2	0.940	0.989	0.003	0.436	95.0
4	ECC, DHC, AREADFG2	0.957	1.000	0.083	0.436	99.0
5	AREA, VP	1.000	1.000	0.988	0.143	100.0
6	AREA, DFG, VP	1.000	1.000	0.987	0.699	100.0

Figure 4.7 illustrates the comparison of different models with different predictor variables.

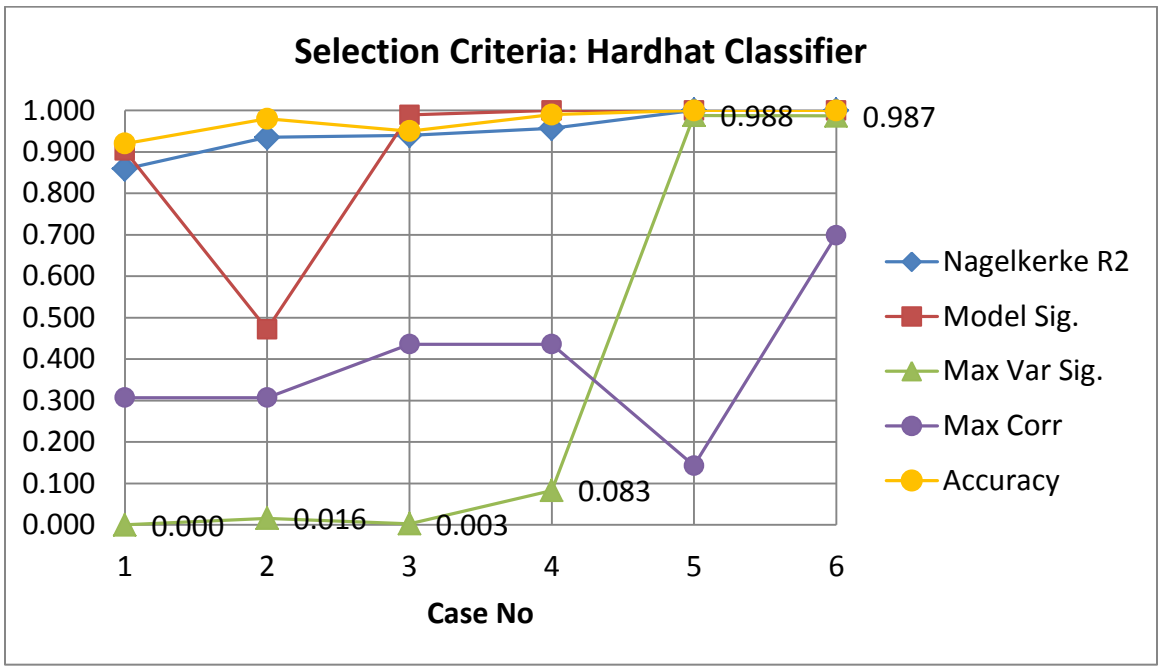


Figure 4.7: Comparison of model parameters (Hardhat classifier)

Observing the comparison illustrated above, case numbers 1 to 4 are the potential models for the hardhat classifier. Even though case number 5 and 6 reached the upper level of three criteria (i.e. Nagelkerke R squared value, model significance and accuracy), higher variable significance reduced the impact of the quality of the model. Hence it is rejected. Case number 4 with variables ECC, DHC, AREADFG2 is selected because of highest R2 and model significance while keeping the correlation of variables used in a moderate level.

Table 4.5: Variables in the equation (Hardhat Classifier)

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
ECC	5.459	3.148	3.007	1	0.083	234.899
DHC	-0.338	0.107	9.928	1	0.002	0.713
AREADFG2	1.102	0.430	6.562	1	0.010	3.009

The z for the hardhat probability formula is;

$$z = 0 + 5.549(ECC) - 0.338(DHC) + 1.102(AREADFG2)$$

The correlation matrix is depicted below. The maximum correlation amount is recorded between the distance to head from centroid and new variable (Area*distance to figure ^2).

Table 4.6: Correlation matrix of parameters (Hardhat classifier)

	ECC	DHC	AREA*DFG^2/1000
ECC	1.000	0.307	-0.157
DHC	0.307	1.000	-0.436
AREA*DFG^2/1000	-0.157	-0.436	1.000

Figure 4.8 shows the predicted probabilities of the constructed hardhat logistic model. The first 35 samples are non-hardhat samples. Only 3 samples reported above 0.1 probabilities while the maximum false probability lies just below 0.33 level. The second lowest probability of hardhat is 0.53, hence anywhere between 0.33 and 0.53 can be selected as the cut-off level. We selected the middle of this range 0.44 as the cut-off level and Table 4.7 tabulated the detailed accuracy of the model.

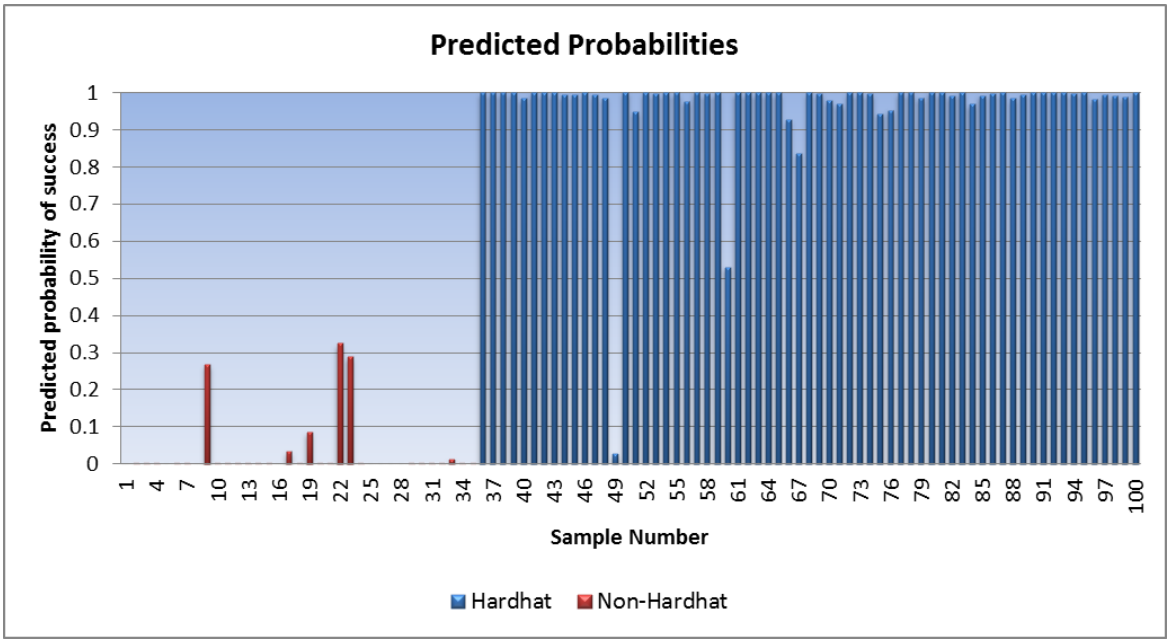


Figure 4.8: Predicted probabilities of hardhat classifier

Table 4.7: Model accuracy - Hardhat classifier

Observed		Predicted		
		Observed Hardhat		Percentage Correct
		0	1	
Observed Hardhat	0	35	0	100
	1	1	64	98.5
Overall Percentage				99
a. The cut value is 0.430				

4.5.3 Occlusion handling

Considering that most construction sites involve collaborative work, interactions between workers and machines will happen frequently. The interactions cause static or dynamic occlusion, result in difficulties for tracking. Therefore, occlusion handling still remains an

open problem in visual tracking. However, by tracking each worker, it is possible to use global information to tackle occlusion.

The Kalman filter is proposed as the solution to the occlusion handling of human figures. The Kalman filter estimates the state of a dynamic system from a series of incomplete or noisy measurements. This can use the previously estimated state to predict the current state.

The algorithm works in a two-step process: estimation and prediction. In the prediction step, the Kalman filter produces estimates of the current state variables, along with their uncertainties. Once the outcome of the next measurement is observed, these estimates are updated. Since we are focusing on the upper body of a human figure, real world skeleton joint coordinates are fed into the Kalman filter. Then these estimated coordinates are used to locate the worker once it is occluded for a short time period. Our system continuously estimates the location for two second period once the worker is being occluded. The true state at time k is modeled from the Kalman filter as given below.

$$x_k = F_k x_{k-1} + B_k u_k + w_k \qquad (5)$$

Where F_k is the state transition model which is applied to the previous state x_{k-1}, B_k is the control-input model which is applied to the control vector u_k, and w_k is the process noise $w_k \sim N(0, Q_k)$. Q_k, is the covariance of the process noise.

The observation z_k of the true state x_k at time k is made according to is made according to following equation.

$$z_k = H_k x_k + v_k \qquad (6)$$

Where H_k is the observation model which maps the true state space into the observed space and v_k is the observation noise $v_k \sim N(0, R_k)$, R_k, is the covariance of the observation noise.

The 3D position of the spine joint in the skeleton figure (see Figure 4.2) which is considered as moving in a linear path is fed if the tracking status is active, otherwise the prediction value of the previous frame is used if the tracking status is deactivated (i.e. figure is occluded). In the tracking subsystem, the Kalman filter block uses the locations of the skeleton figure detected in the previous frames to predict the locations in the current frame and Figure 4.9 illustrates a series of snapshots of handling occlusion.

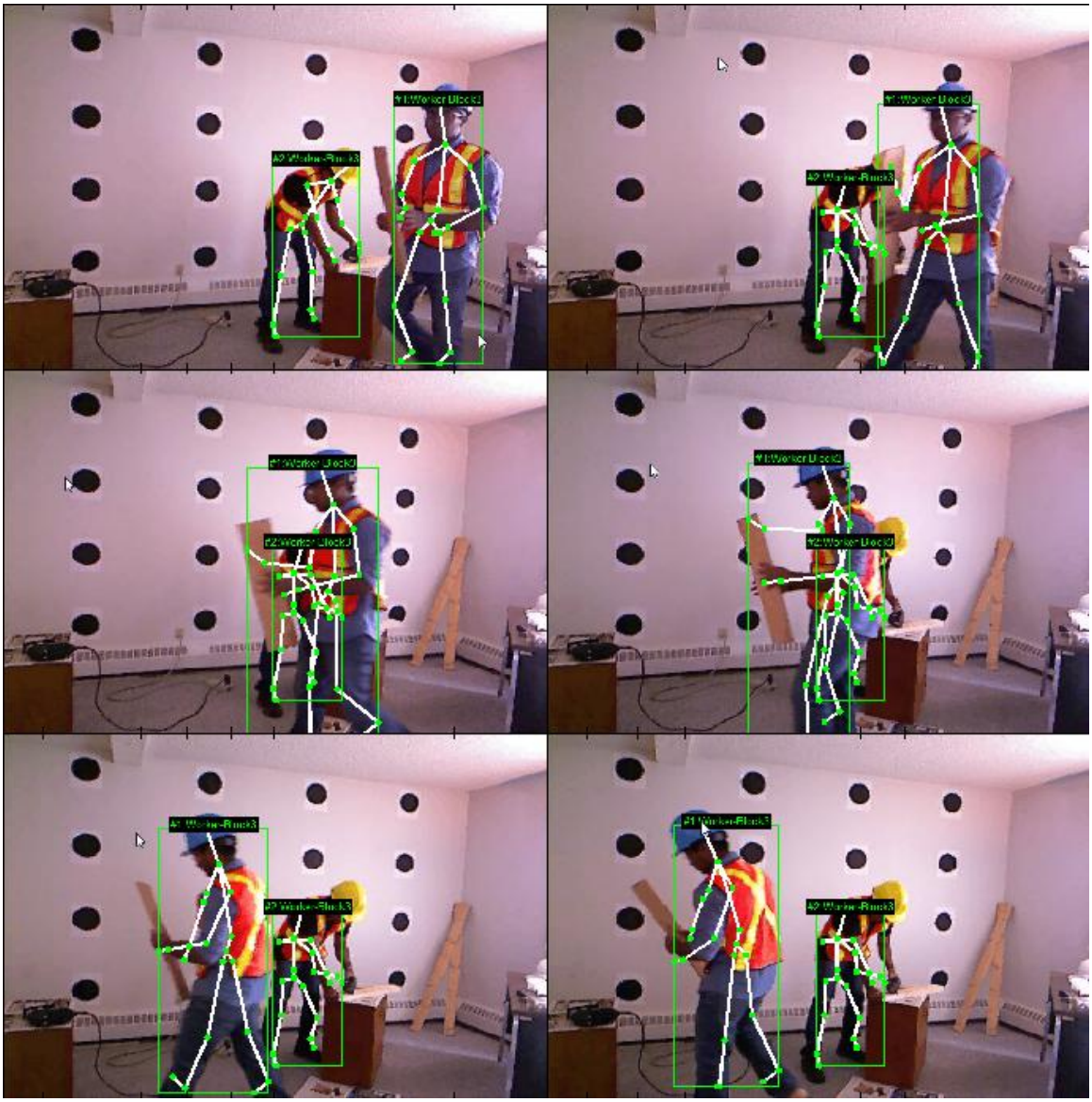


Figure 4.9: Instances of occlusion handling

Kinect infrared camera generates a depth map of the environment and based on this information, 3D coordinate of each pixel can be determined with respect to the Kinect device. In order to get the 3D coordinate with respect to the building coordinate system transformed from Kinect coordinate system, the rigid body transformation

algorithm is applied. IOP and EOP are the prerequisite for the rigid body transformation and these measures can be determined by following camera calibration and single photo resection procedures.

4.5.4 Camera calibration

Accurate camera calibration and orientation procedures are a necessary prerequisite for the extraction of precise and reliable 3D metric information from images. A camera is considered calibrated if the principal distance, principal point offset, and lens distortion parameters are known. The purpose of camera calibration is to determine numerical estimates of the interior orientation parameters (IOP) (Brown, 1971) and image coordinate corrections that compensate for various deviations from the assumed perspective geometry of the implemented Kinect camera. The IOP comprises the focal length (c); location of the principal point (x_p , y_p).

There is an extensive body of literature on the calibration of digital cameras, with topics ranging from overall reviews (Remondino & Fraser, 2006) to low cost digital cameras (Cronk, Fraser, & Hanley, 2006), and stability of parameters (Habib, Pullivelli, & Morgan, 2005).

Camera calibration requires control information, which is usually available in the form of a test field. Traditional calibration test fields consist of distinct and specifically marked points or targets. Establishing and maintaining a conventional test field, as well as carrying out the calibration procedure, require professional surveyors and photogrammetrists. Such requirements limit the potential use of high quality and low cost

digital cameras, and hence the well-known Bouguet (2010) calibration toolbox for MATLAB has already been used in many research studies (Smisek, Jancosek, & Pajdla, 2011; Zhu, Pan, & Luo, 2010).

We used a total of 16 monochrome test images of a planar checkerboard (55x55(*mm*) square size) for calibration of the Kinect RGB camera. Since the Kinect device requires a frame grabber, the following application was developed: Test images feature a planar checkerboard grid differently oriented in each image shown on Figure 4.10. This program features an algorithm that uses the extracted corner points of the checkerboard pattern to compute a projective transformation between the image points of the *n* different images. Afterwards, the camera's intrinsic and extrinsic parameters are recovered using a closed-form solution, while the third and fifth order radial distortion terms are recovered within a linear least-squares solution.

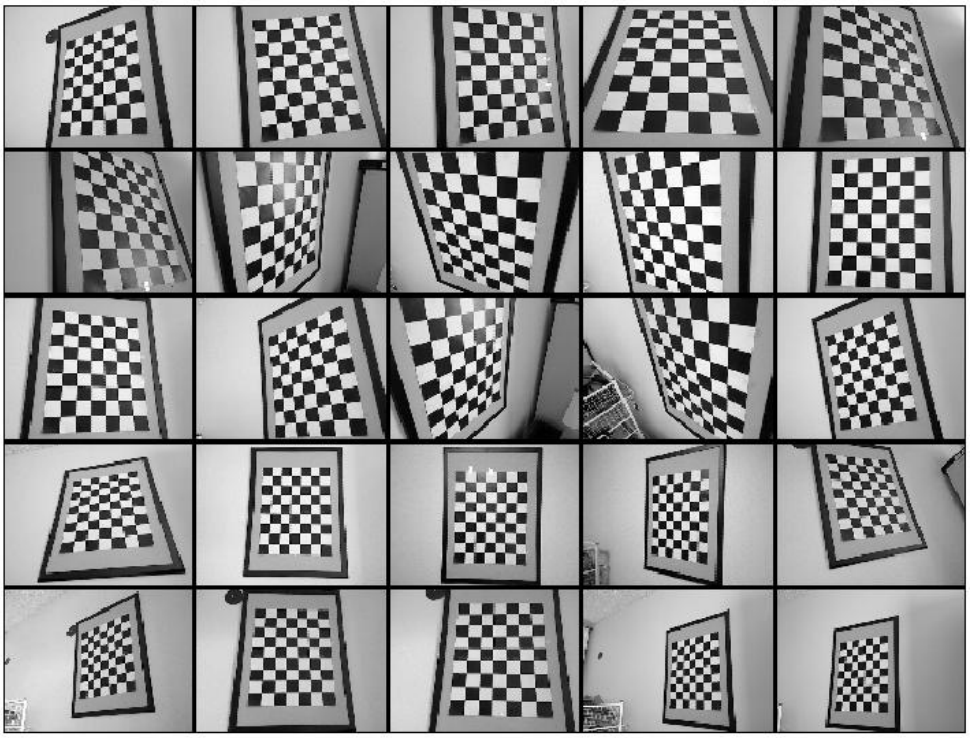


Figure 4.10: Camera calibration test images

Calibration should be accurately modelled and estimated, since a small error in these parameters causes large deviations from the actual position. Using the MATLAB software developed by Bouguet (2010), intrinsic parameters and distortion parameters were determined as tabulated in Table 4.8 and Table 4.9.

Table 4.8: Intrinsic Parameters of Kinect RGB

Description	X	Y
Focal Length (pixel)	526.01	527.13
Principal point (pixel)	331.89	261.26
Pixel residual error	0.31	0.23

Table 4.9: Distortion parameters (kc) of Kinect RGB

	Pixel value	Error
kc-1	0.19901	0.01538
kc-2	-0.42864	0.06047
kc-3	0.00166	0.00216
kc-4	-0.00012	0.00164
kc-5	0	0

Smisek et al. (2011) carried out a study to make a quantitative comparison of Kinect accuracy with a stereo reconstruction from digital SLR cameras. According to calibration results presented in the paper, these were approximately similar to our results. The paper also mentioned that the size of the pixel in the best resolution of Kinect RGB image is 2.8µm. Hence focal length can be calculated in mm as 2.95mm. The following figure visualizes the extrinsic parameters of the camera.

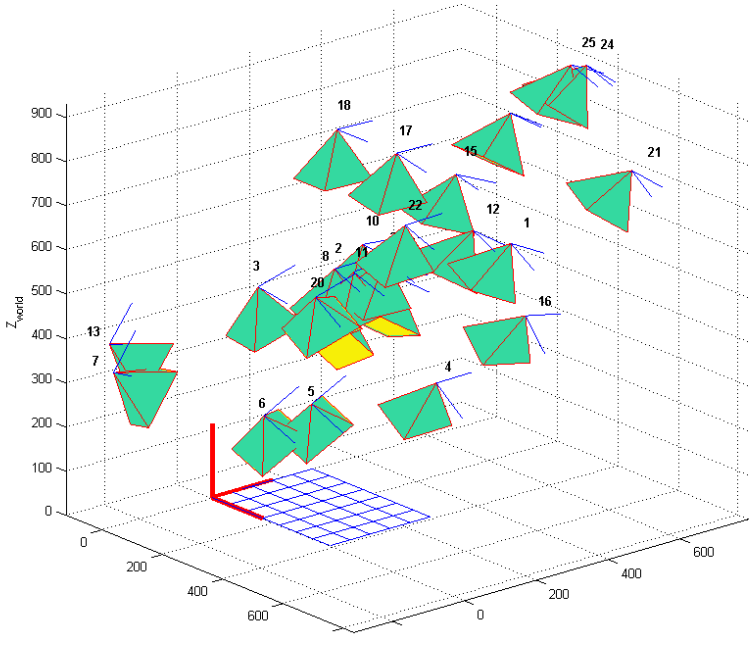


Figure 4.11: Visualization of extrinsic parameters (Bouguet, 2010)

4.5.4.1 Distortion parameters

Distortion parameters ($\Delta x, \Delta y$) compensate for all the deviations from the assumed perspective geometry: object point, perspective center, and the corresponding image point lie on a straight line. Previous studies indicate that radial lens distortion, de-centric lens distortion, atmospheric refraction, affine deformations, and out-of-plane deformations are the main potential sources of the deviation from a collinearity condition (Fraser, 1997). Since the research study is carried out in the indoor work environment and at close range distance, the atmospheric refraction distortion amount is considered a negligible parameter. In addition, as the Kinect RGB camera uses only a single lens, de-centric distortion will not be considered in this formula. Thus, apart from the barrel type radial lens distortion, all the other distortion parameters are considered negligible impact parameters to the final accuracy of the RGB image output.

- Radial lens distortion

The radial lens distortion causes non-linear geometrical distortion on images. In general, this distortion is caused by either large off-axial angle or lens manufacturing flaws.

a. Radial lens distortion occurs along a radial direction from the principal point.
b. Radial lens distortion increases as we move away from the principal point.

Figure 4.12 graphically describes the radial distortion (Δr) with respect to the image coordinate system.

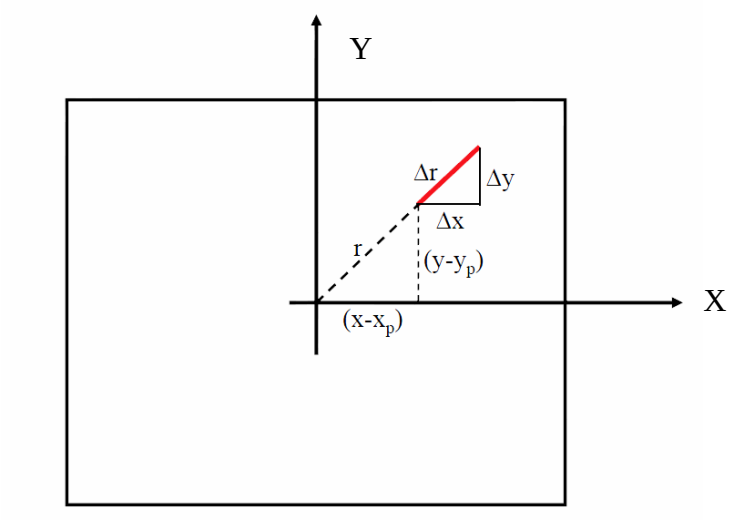


Figure 4.12: Radial lens distortion (Habib, 2008a)

The radial lens distortion model can be written as an infinite series, as given below:

$$\begin{aligned}\Delta x &= (x - x_p)(k_1 r^2 + k_2 r^4 + k_3 r^6 + ... + k_n r^{2n}) \\ \Delta y &= (y - y_p)(k_1 r^2 + k_2 r^4 + k_3 r^6 + ... + k_n r^{2n})\end{aligned} \qquad (7)$$

where, $k_1, k_2, ..., k_n$, are the radial lens distortion coefficients, $r = \sqrt{(x - x_p)^2 + (y - y_p)^2}$. The first five coefficients were already determined from the camera calibration procedure described in the preceding section. $k_i r^{2i}$ is considered as negligible when i >=3, hence second-order radial symmetric distortion parameters $k_1 = 0.19901$ and $k_2 = -0.42864$ were taken into consideration.

Thus the polynomial distortion model can be simplified as:

$$\Delta x = (x - x_p)(k_1 r^2 + k_2 r^4)$$
$$\Delta y = (y - y_p)(k_1 r^2 + k_2 r^4) \quad (8)$$

4.5.5 3D location determination

The primary objective of this worker tracking system is to generate spatial and descriptive information in terms of the construction site from the Kinect sensor. Kinect sensory information provides 3D coordinate information (w.r.t. its own coordinate system see Figure 4.13) of the scene by analyzing its depth map.

Figure 4.13: Kinect built-in coordinate system

In order to communicate with other major project information such as survey blueprints, drawings (i.e. construction, structural and architectural) and 4D CAD models these extracted information has to be transformed into a common platform, the building coordinate system. The rigid body transformation was proposed for this coordinate conversion (see Figure 4.14) and a detail description of rigid body transformation is discussed in the section: case study.

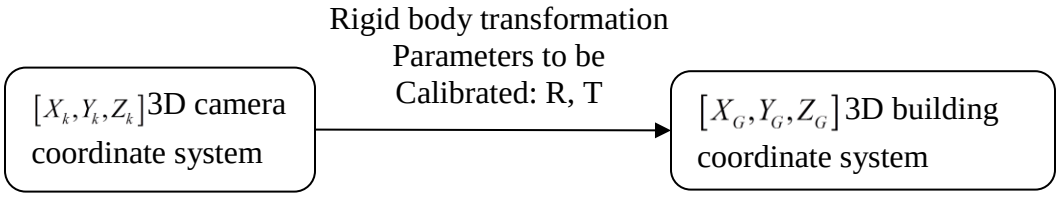


Figure 4.14: Rigid body transformation

The rigid body transformation includes rotation and translation matrices (exterior orientation parameters) of the original coordinate system (Kinect). Therefore, the single photo resection (SPR) is applied prior to rigid body transformation in order to determine the exterior orientation parameters of the Kinect device and these parameters will be applied in 3D transformation to convert the camera coordinate system to the building coordinate system.

4.5.6 Single photo resection (SPR)

Since the camera position and the rotation angles are fixed during the monitoring period in this study, exterior orientation parameters should be estimated using a single photo. This section describes the methodology of determining the extrinsic parameters: position (X_0, Y_0, Z_0) of the camera station and the orientation $(\omega, \varphi, \kappa)$ of the image coordinate system relative to the ground coordinate system using only a single image taken from the Kinect RGB camera. The mathematical model for this solution is developed based on the collinearity equations as described in the class notes of Habib (2008b). Collinearity equations can be applied when the perspective center, the object point, and the

corresponding image point are collinear (Habib et al., 2005). The collinearity equations are illustrated as:

$$x = x_p - c\frac{r_{11}(X - X_0) + r_{21}(Y - Y_0) + r_{31}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} + \Delta x \quad (9)$$

$$y = y_p - c\frac{r_{12}(X - X_0) + r_{22}(Y - Y_0) + r_{32}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} + \Delta y \quad (10)$$

where, x and y are the image coordinates, X , Y and Z are the ground coordinates, Δx and Δy are compensations for the deviations from collinearity (i.e. radial lens distortions, de-centric distortions, affine deformations are considered), x_p , y_p and c are the perspective point coordinates of the camera, X_0, Y_0, Z_0 are the ground coordinates of the exposure station (perspective center), $r_{11}, r_{12}, ..., r_{33}$ are the elements of the rotation matrix that are functions of ω, φ and κ.

It is assumed that the IOP (x_p, y_p, c) of the Kinect RGB camera and the ground coordinates of the control points (X, Y, Z) are pre-determined and treated as errorless. In order to solve the six unknown EOP, a minimum of three ground control points are required. However, all of the available control points shown in the test field have been used in order to increase the accuracy provided by the data redundancy. Therefore the least squares procedure was used to solve the unknown parameters of this over-determined system.

Collinearity equations are nonlinear with respect to the EOP, which are the unknown parameters of the single photo resection problem. Taylor's theorem is applied to linearize collinearity equations using approximate values for the unknown parameters.

$$x = x_0 + \left(\frac{\partial x}{\partial X_0}\right)dX_0 + \left(\frac{\partial x}{\partial Y_0}\right)dY_0 + \left(\frac{\partial x}{\partial Z_0}\right)dZ_0 + \left(\frac{\partial x}{\partial \omega}\right)d\omega + \left(\frac{\partial x}{\partial \varphi}\right)d\varphi + \left(\frac{\partial x}{\partial \kappa}\right)d\kappa$$
$$y = y_0 + \left(\frac{\partial y}{\partial X_0}\right)dX_0 + \left(\frac{\partial y}{\partial Y_0}\right)dY_0 + \left(\frac{\partial y}{\partial Z_0}\right)dZ_0 + \left(\frac{\partial y}{\partial \omega}\right)d\omega + \left(\frac{\partial y}{\partial \varphi}\right)d\varphi + \left(\frac{\partial y}{\partial \kappa}\right)d\kappa \qquad (11)$$

$dX_0, dY_0, \ldots$ etc. are the unknown corrections to be applied to the initial approximations.

x_0, y_0 are the corrected image coordinates. In the above linearized collinearity equations, the partial derivatives can be replaced by a simpler notation for handling convenience, as follows:

$$a_1 = \left(\frac{\partial x}{\partial X_0}\right), a_2 = \left(\frac{\partial x}{\partial Y_0}\right), a_3 = \left(\frac{\partial x}{\partial Z_0}\right), a_4 = \left(\frac{\partial x}{\partial \omega}\right), a_5 = \left(\frac{\partial x}{\partial \varphi}\right), a_6 = \left(\frac{\partial x}{\partial \kappa}\right)$$
$$b_1 = \left(\frac{\partial y}{\partial X_0}\right), b_2 = \left(\frac{\partial y}{\partial Y_0}\right), b_3 = \left(\frac{\partial y}{\partial Z_0}\right), b_4 = \left(\frac{\partial y}{\partial \omega}\right), b_5 = \left(\frac{\partial y}{\partial \varphi}\right), b_6 = \left(\frac{\partial y}{\partial \kappa}\right) \qquad (12)$$

Which yields,

$$x - x_0 = a_1 dX_0 + a_2 dY_0 + a_3 dZ_0 + a_4 d\omega + a_5 d\varphi + a_6 d\kappa$$
$$y - y_0 = b_1 dX_0 + b_2 dY_0 + b_3 dZ_0 + b_4 d\omega + b_5 d\varphi + b_6 d\kappa \qquad (13)$$

The Gauss-Markov model (GMM) is used to solve for the unknown EOPs. The general form of the GMM is:

$$y = Ax + e \qquad (14)$$

Where y is the observation vector, A is the design matrix, x is the unknown vector and e is the noise contaminating the observation vector.

Table 4.10: Detail description of least squares parameters

	Description	size
y	Vector of differences between the measured and computed image coordinates using the approximate values for the unknown parameters	$n \times 1$
A	Design matrix composed of the partial derivatives	$n \times m$
x	Vector of unknown corrections to the approximate values of the unknown parameters	$m \times 1$
e	Error vector	$n \times 1$
$\sigma_0^2 P^{-1}$	Variance covariance matrix of the noise vector	$n \times n$

The GMM can be illustrated in matrix form as given below.

$$y = \begin{bmatrix} x_1 - x_{1_0} \\ y_1 - y_{1_0} \\ x_2 - x_{2_0} \\ y_2 - y_{2_0} \\ \vdots \\ x_{n/2} - x_{n/2_0} \\ y_{n/2} - y_{n/2_0} \end{bmatrix}_{n\times 1} = \begin{bmatrix} a_{1_1} & a_{2_1} & a_{3_1} & a_{4_1} & a_{5_1} & a_{6_1} \\ b_{1_1} & b_{2_1} & b_{3_1} & b_{4_1} & b_{5_1} & b_{6_1} \\ a_{1_2} & a_{2_2} & a_{3_2} & a_{4_2} & a_{5_2} & a_{6_2} \\ b_{1_2} & b_{2_2} & b_{3_2} & b_{4_2} & b_{5_2} & b_{6_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{n\times 6} \cdot \begin{bmatrix} dX_0 \\ dY_0 \\ dZ_0 \\ d\omega \\ d\varphi \\ d\kappa \end{bmatrix}_{6\times 1} + \begin{bmatrix} e_{x_1} \\ e_{y_1} \\ e_{x_2} \\ e_{y_2} \\ \vdots \\ e_{x_{n/2}} \\ e_{y_{n/2}} \end{bmatrix}_{n\times 1} \qquad (15)$$

The following set of equations is applied to solve the unknown parameters (x), residual error distribution ($e \sim (0, \sigma_0^2 P^{-1})$), variance covariance ($D(x)$) and variance component (σ_0^2) of the above problem. n and m are sizes of the observation vector and unknown vector.

$$
\begin{aligned}
x &= (A^T PA)^{-1} A^T Py \\
D(x) &= \sigma_0^2 (A^T PA)^{-1} \\
\tilde{e} &= y - Ax \\
\sigma_0^2 &= (\tilde{e}^T P\tilde{e}) / (n-m)
\end{aligned}
\qquad (16)
$$

4.5.6.1 Pixel to image coordinate transformation

A digital image is essentially a matrix, consisting of a certain number of rows and a certain number of columns (see left image in Figure 4.15). In pixel coordinate system the origin is located at the left upper corner of the image. However, in image coordinate system is defined by central rows (x axis) and central columns (y axis) (see right image in Figure 4.15).

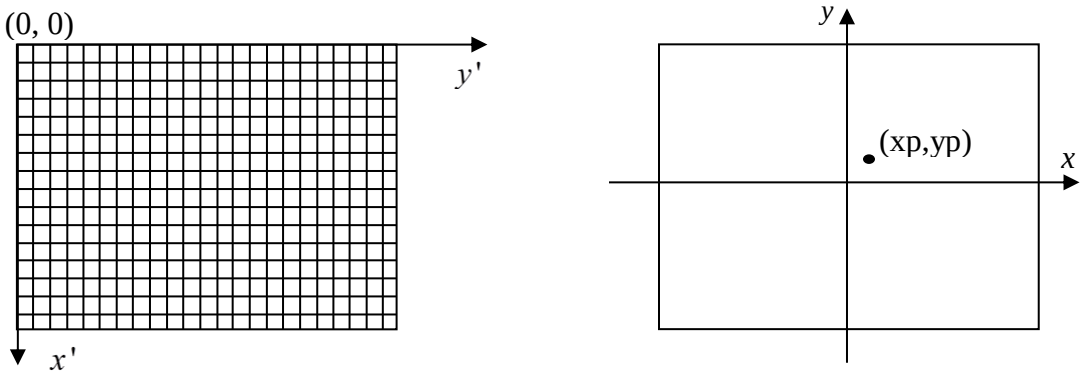


Figure 4.15: a) pixel coordinate system and b) image coordinate system

Pixel-to-image coordinate transformation is done by:

$$
\begin{aligned}
x &= (y' - n_c / 2.0) \times y\_pix\_size \\
y &= (n_r / 2.0 - x') \times x\_pix\_size
\end{aligned}
\qquad (17)
$$

where, n_r is number of rows, n_c is number of columns, x_pix_size is pixel size along the row direction, y_pix_size is pixel size along the column direction, and x', y' are image pixel coordinates.

4.5.7 Case study

An indoor laboratory similar to a construction environment was selected as the test environment of the study. We set up the apparatus and marked a set of ground control points (known 3D coordinates) in the field as required for the study. To model the SPR we developed an application and captured the image as shown in Figure 4.16. Semi-automated target extraction is introduced to the system which tracks checkerboard grid lines (code developed by: Bouguet (2010)), circular shapes (code developed by: Shoelson (2008)) and manually selected arbitrary target points. The circular shaped target points (see Figure 4.16) are tracked in the entire figure using Hough transform algorithm. The lower left corner is considered the origin of the building coordinate system for the purposes of system validation. Hence, a total of 16 observed control points were used for the SPR model.

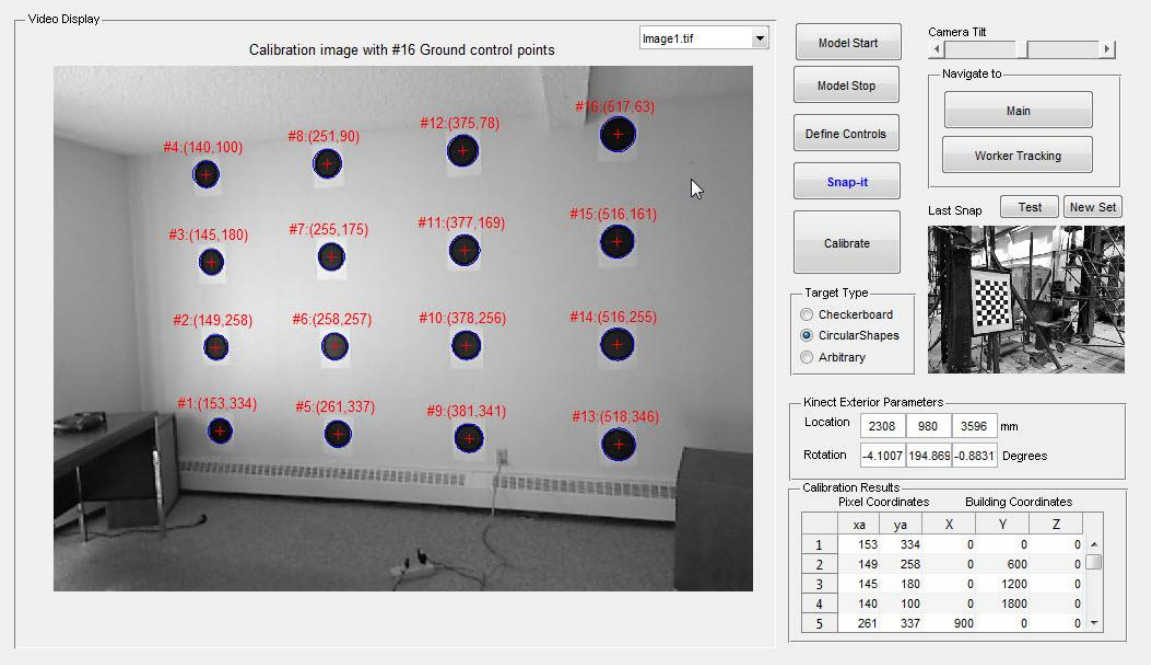


Figure 4.16: Single photo resection (SPR) application

Increased redundancy and higher observation points provide a key to optimized levels of accuracy and robustness in the SPR process. The SPR process has been carried out in the test environment and the redundancy (r) calculation is shown below:

Number of observations, [n Number of points] × 2 (x and y coordinates) =16×2=32

Number of unknown parameters ($dX_0, dY_0, dZ_0, d\omega, d\varphi, d\kappa$), m = 6

Redundancy, r = (Number of observations - Number of unknowns) = 26

The pixel coordinates extracted from Hough transform algorithm are then transformed into image coordinates as mentioned in 4.5.6.1. Further, corrected x and y coordinates are determined considering radial distortion factors as discussed above.

In order to find initial approximations of EOPs we assumed a vertical photograph (i.e. ω, φ=0) and estimate the κ, Xo and Yo from 2D similarity transform model given by:

$$X = a_0 + ax - by$$
$$Y = b_0 + bx + ay \qquad (18)$$

where, (x, y) are the image coordinates and the (X, Y) are corresponding ground coordinates. The angle κ is calculated as $\kappa = \tan^{-1}(\frac{b}{a})$. Hence the rotation matrix can be calculated. The initial approximations of Xo, Yo are taken as the a_0, and b_0 values of the above equation, and for Zo, an average of Z in ground control points measured from the Kinect depth map is used. Appendices addressing the following topics are attached:

1. Measured image coordinates and the approximations to the unknown parameters
2. The modified exterior orientation parameters after each iteration
3. The final adjusted values for the exterior orientation parameters
4. An estimate of the variance component of each iteration
5. The posterior variance-covariance (dispersion) matrix of the parameters
6. The residuals associated with the image coordinate measurements

As a result of the SPR analysis the following results have been generated: estimate of the variance component of the last iteration is 2.56E-05, final adjusted values for the exterior orientation parameters are shown in the Table 4.11 and the posterior variance-covariance (dispersion) matrix of the parameters for the last iteration is given in the Table 4.12. These measurements can be used when reviewing the quality of the results generated

from the model and the accuracy of the control points in the field. Covariance values of rotation angles in the diagonal of Table 4.12 are less than or marginal to the selected accuracy threshold level, 5.0E-05. In brief, SPR has provided a moderate accuracy level in generating exterior orientation parameters of the Kinect device.

Table 4.11: Final adjusted values for the exterior orientation parameters

Unknown	Final Adjusted Values
X (mm)	2308
Y (mm)	980
Z (mm)	3596
ω (radians)	-0.0716
ϕ (radians)	3.4013
κ (radians)	-0.0154

Table 4.12: Posterior variance-covariance (dispersion) matrix of the parameters:

	X (mm)	Y(mm)	Z(mm)	ω(rad)	ϕ(rad)	κ(rad)
X(mm)	408.404	-47.410	-168.549	0.010	0.110	-0.002
Y(mm)	-47.410	718.573	4.320	-0.193	-0.013	0.014
Z(mm)	-168.549	4.320	98.469	5.92E-05	-0.047	0.001
ω (rad)	0.010	-0.193	5.92E-05	5.20E-05	2.86E-06	-3.79E-06
ϕ (rad)	0.110	-0.013	-0.047	2.86E-06	2.98E-05	-6.33E-07
κ (rad)	-0.002	0.014	0.001	-3.79E-06	-6.33E-07	2.15E-06

4.5.8 3D transformation

Figure 4.17 depicts the structure of the 3D transformation: O_G denotes the origin of the ground coordinate system and O_M denotes the origin of the Kinect coordinate system. First target locations are measured using the Kinect coordinate system and then they are

transformed to the ground coordinate system by applying the following mathematical model for the 3D transformation (Zeng, 2010):

$$\overrightarrow{X_G} = \overrightarrow{\Delta X_T} + R(\omega\varphi\theta)\overrightarrow{X_i} \qquad (19)$$

where, $\vec{X}_G$ is the target coordinate vector with respect to ground, $\overrightarrow{\Delta X}_T$ is translational (shift) vector, R (ωφθ) is the rotational relationship between the Kinect device and the ground coordinate system and $\vec{X}_i$ is the target coordinate vector generated from the Kinect device. Once the above equation is applied, ground coordinates of the Kinect device can be determined.

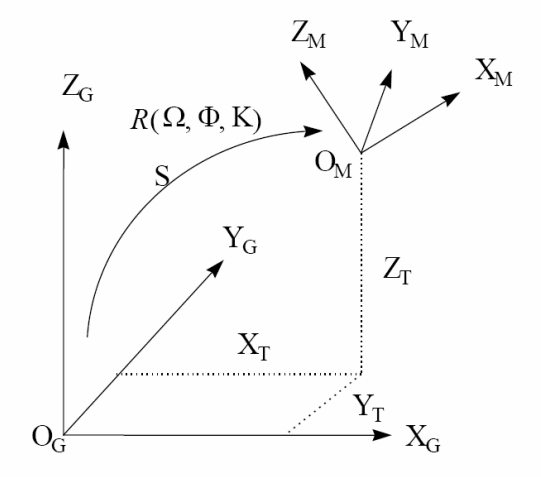


Figure 4.17: Rigid body transformation (Habib, 2008a)

4.6 Summary

In this study we focused on the problem of detecting and differentiating construction personnel in indoor dynamic construction environments. We used an inexpensive range camera (Microsoft Kinect) to collect RGBD data. This low cost sensor enables wide

applicability of the proposed method. The experimental result shows that we can use this system to track multiple workers on the job site. However, currently it permits tracking of only two human figures, which is a limitation of the MATLAB execution (mex) files used for this study. The preliminary experimental result shows that the overall accuracy of the hardhat classifier reaches 99.0% of the accuracy level while having 100% precision (see Table 4.7)

In order to reduce the computational cost while keeping the reliability of the system, the worker recognition process is applied to the first 10 images and thereafter it recognizes a skeleton. Subsequently the system confirms the category of the person it tracks (i.e. worker, supervisor, etc.) only by the skeleton location. This avoids a false alarm when a worker removes a hardhat for a short time period while working on site. Moreover, the system continuously tracks workers even under low light conditions once the category is determined.

Chapter Five: **Construction Activity Recognition System**

This chapter elaborates on the literature study of audio event classification, analysis for a construction tool set selection, detailed description of the tool sound detection system, training dataset preparation, step-wise comprehensive tool sound classifier and its accuracy with experimental results.

5.1 Introduction

Activity analysis, the continuous and detailed process of benchmarking, monitoring, and improving the amount of time craft workers spend on different construction activities can play an important role in improving construction productivity. As a workface assessment tool, activity analysis examines the proportion of time workers spend on specific construction activities. A combination of detailed assessment and continuous improvement significantly differentiates activity analysis from work sampling and can provide recommendations for activity monitoring, improvements, and improvement applicability. In recent years, many companies have experienced the benefits of activity analysis and are now proactively working towards implementing it in their projects (Escorcia et al., 2012).

5.2 Literature of Audio Event Classification

Humans classify audio signals, and their approximate direction and location, most of the time without a conscious effort. Recognizing a voice on the telephone, identifying the difference between a telephone ring and a doorbell ring, are not considered difficult tasks

for a human being. However, classification can become an issue if noise interferes, if sound signal is weak, or if a sound is similar to another sound. In this research study, a similar concept is applied to recognizing construction activities by identifying unique tool sound patterns and their direction in a construction job site.

Audio event classification/detection is receiving increased attention from the scientific community, particularly in the context of audio retrieval and indexing applications but also in the context of multimedia event detection applications where audio can be used as a complementary source of information. The use of audio sensors in surveillance and monitoring applications has proven to be particularly useful for the detection of events like screams or gunshots (Clavel, Ehrette, & Richard, 2005; Rouas, Louradour, & Ambellouis, 2006; Valenzise, Gerosa, Tagliasacchi, Antonacci, & Sarti, 2007). Such detection systems can be efficiently used to signal to an automated system that an event has occurred and, at the same time, to enable further processing like acoustic source localization for steering a video camera. Much of the previous work about audio-based surveillance systems has concentrated on detecting some particular audio events. Early research stems from the field of automatic audio classification and matching. In this way, audio sensors have been used in various contexts for recognizing sound events, such as vehicle class sound signature recognition (Huadong, Siegel, & Khosla, 1999), environmental sound recognition (Chu, Narayanan, & Kuo, 2009), etc. More recently, specific works covering the detection of particular classes of events for multimedia-based surveillance have been developed.

On top of this, sound source localization is another challenging task in these audio detection systems. The most popular technique for source localization in environments with small reverberation time is based on time difference of arrivals (TDOA) of the signal at a microphone array. These time delays are further processed to estimate the source location (K. Varma, Ikuma, & Beex, 2002).

In this study we propose a surveillance system that is able to accurately detect and localize construction tool sounds. The audio stream is recorded by a Kinect four channel microphone array. Audio segments are classified as predefined tools such as a jigsaw, a staple gun, an angle grinder and hammer, and background noise. Audio classified as noise is discarded. If a pre-defined tool sound event is detected, the localization module estimates the TDOA at each sensor pair of the array and computes the approximate sound source direction, adding the tool usage to a particular worker who is recognized using video processing, as described in the Chapter Four:.

5.3 Selected Construction Tool Set For the Study

A construction job site is a place where a vast range of tools and equipment is used to accomplish work. The research study selected four commonly used tools on a construction job site to identify and measure the tool-time and performance. They are jigsaw, staple gun, angle grinder, and hammer (see Figure 5.1). A jigsaw is a tool used for cutting arbitrary curves and custom shapes into a piece of wood, metal, or other material. It can be used in a more artistic fashion than other saws, which typically cut in straight lines only. A hammer is the most popular tool in any construction environment,

meant to deliver an impact to an object. This is a multi-purpose tool and hammers are widely used for many different applications: driving nails, fitting parts, forging metal, and breaking up objects. Another commonly available tool in construction job sites is an angle grinder. Common uses of the angle grinders are removing excess material from a piece or simply cutting into a piece. This is the most powerful and loudest tool used in this research study, with 11000rpm and powered by 6A.

Figure 5.1: Selected construction tools for audio classifier

A staple gun is a hand-held machine used to drive heavy metal staples into wood, plastic, or masonry. The most common uses of staple guns are to affix a variety of materials, including insulation, house wrap, roofing, wiring, carpeting, upholstery, and hobby and craft materials. The following table summarizes the properties and tasks of each selected tool.

Table 5.1: Properties of selected construction tools

Equipment Name	rpm	Ampere	Tasks
Mastercraft 4.2A Orbital Jigsaw	3600	4.2A	Cut arbitrary curves and custom shapes of wood, metal
Mastercraft Angle Grinder	11000	6A	Remove excess material from a piece, cut into a piece
Mastercraft Sure Shot Light-Duty Staple Gun	N/A	N/A	Affix a variety of materials: insulation, house wrap, roofing, wiring, carpeting, upholstery, etc.
Hammer	N/A	N/A	Drive nails, fit parts, forge metal and break up objects

5.4 Overview of Construction Activity Recognition System

The goal of our construction tool detection system is to segment the input audio stream into successive segments and to label these segments according to the five main classes – as four predefined tool classes and an ordinary class – that represents the environment's acoustic characteristics. The architecture of our audio event detection system includes a feature extraction module, a training module that is used to build the model of the two classes using a logistic regression model, and a classification module that, based on the previous models, labels the successive audio segments.

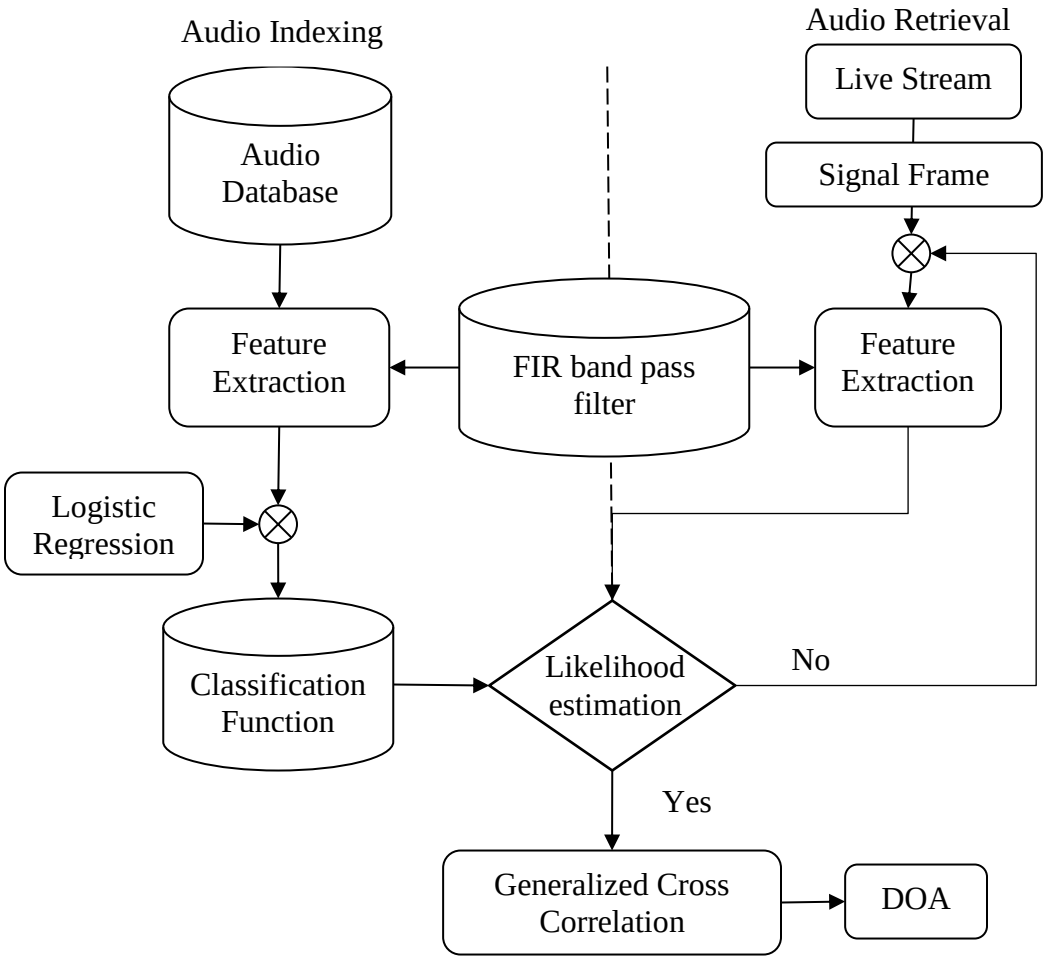


Figure 5.2: Structure of audio indexing and retrieval system

The sampling frequency becomes an important factor for time delay estimation (TDE) based methods, especially when the array is small in terms of distance between the microphones. This is because small distances mean smaller time delays and this requires higher sampling frequencies to increase the resolution of the delay estimates.

The scheme adopted here for construction tool sound recognition is based on a combination of finite impulse response (FIR) band pass filter and statistical approach, seeking to encode the most relevant information in a group of training samples that best

distinguishes them from one another. The classification algorithm finds the most likely class for a given input sound by presenting it to each of the binomial logistic regression models. The model with the highest maximum-likelihood score is selected as the representative class for the sound. As illustrated in Figure 5.2, frequency analysis begins with dividing the raw audio signal into frames. Given the maximum sampling frequency that can be recorded from the Kinect device, which is 16000 Hz, the frames are of 4096 samples (256ms) each, with 12.5% (512 samples or 32ms) overlap in each of the two adjacent frames. Next, a standard fast Fourier transform (FFT) algorithm is applied to each frame. The result is a set of 4096 FFT coefficients. According to the Nyquist limit, a Fourier transform produces a spectrum containing all frequencies from zero to half of the maximum sampling frequency: 8000 Hz. In other words, as the FFT phase information is symmetrical, we take the spectra for subsequent analysis, i.e., we consider only the first half of the FFT phase information where there is a vector with 2048 power spectrum components equally spaced in frequency from 3.9 Hz to 8000Hz at an increment step size of 3.9Hz.

5.5 Training Dataset

We developed a training dataset that consists of a total of 250 tool sound audio samples, including 50 samples of each selected tool sound (i.e. jigsaw, staple gun, angle grinder, and hammer) and another 50 samples of background noise of a typical construction job site, explaining that these collected spectrum samples are in the same class, i.e., from the same kind of tool, recorded under similar conditions. As described in the previous

section, 2048 power spectrum components are analyzed from the recorded audio file. The average sound spectrum distribution of each tool is illustrated in the following figures (Figure 5.3 to Figure 5.6).

Several featured frequency bandwidths are identified for each tool by analyzing frequency distribution of tool sounds. The selection of frequency band widths is based on the wave pattern shape characteristics, such as peaks and valleys. The selected bandwidths of each tool are highlighted in the colour strips in each figure.

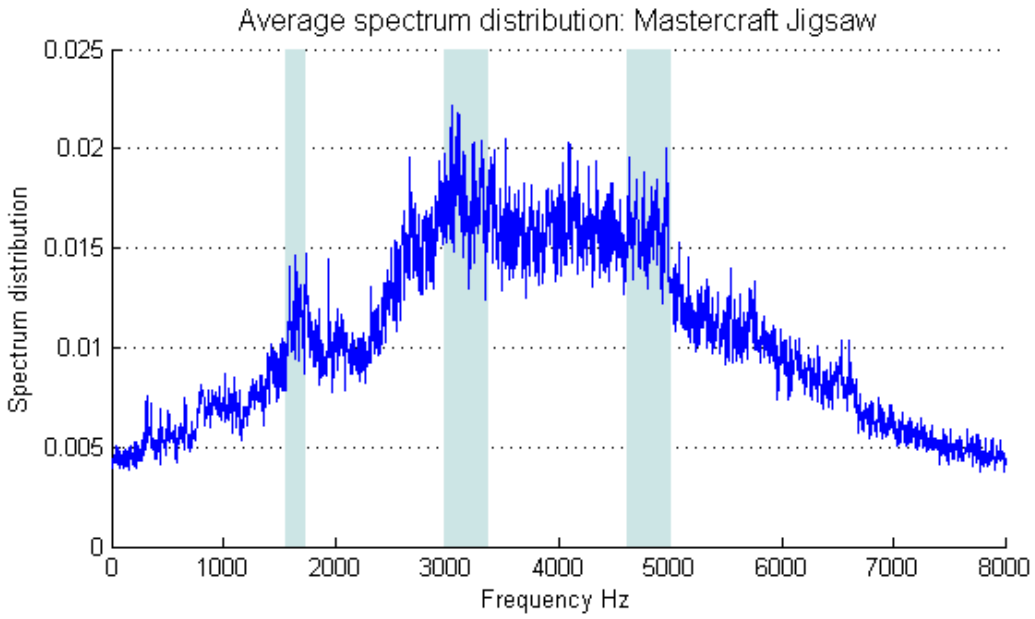


Figure 5.3: Average spectrum distribution: Mastercraft Jigsaw

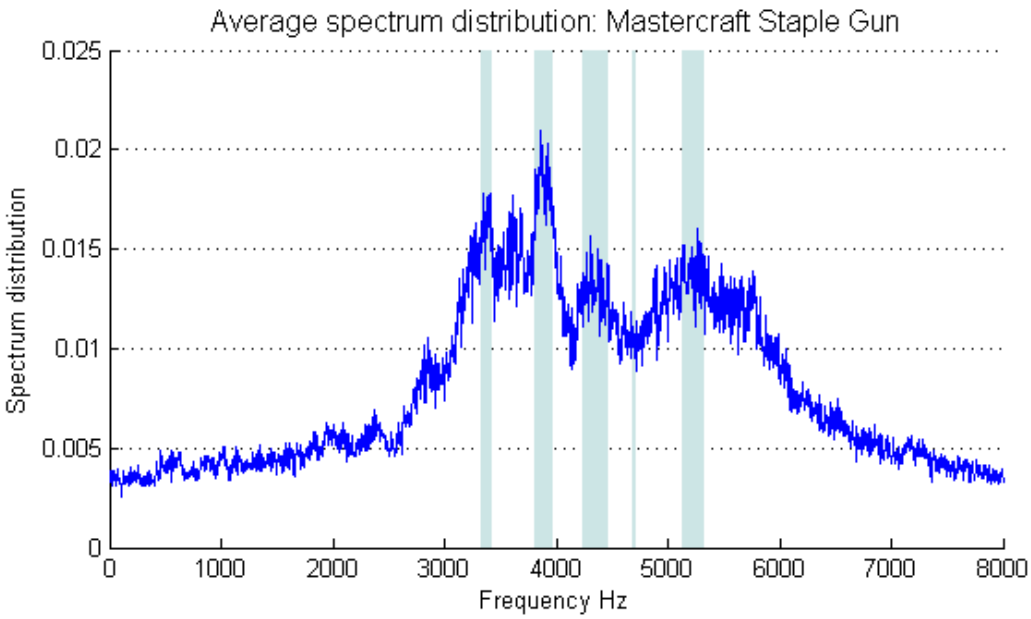


Figure 5.4: Average spectrum distribution: Mastercraft staple gun

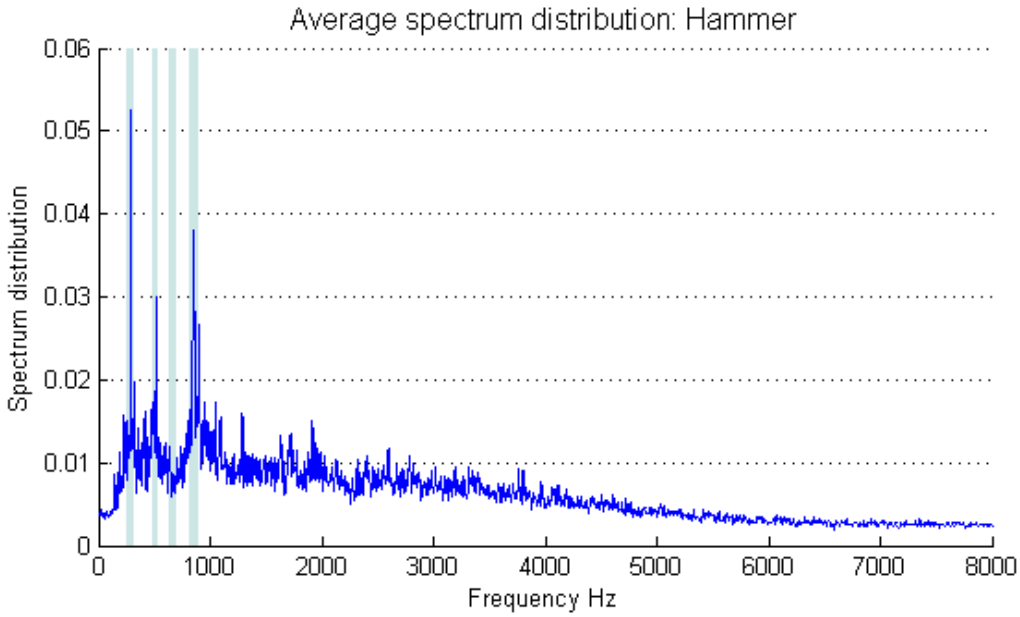


Figure 5.5: Average spectrum distribution: Hammer

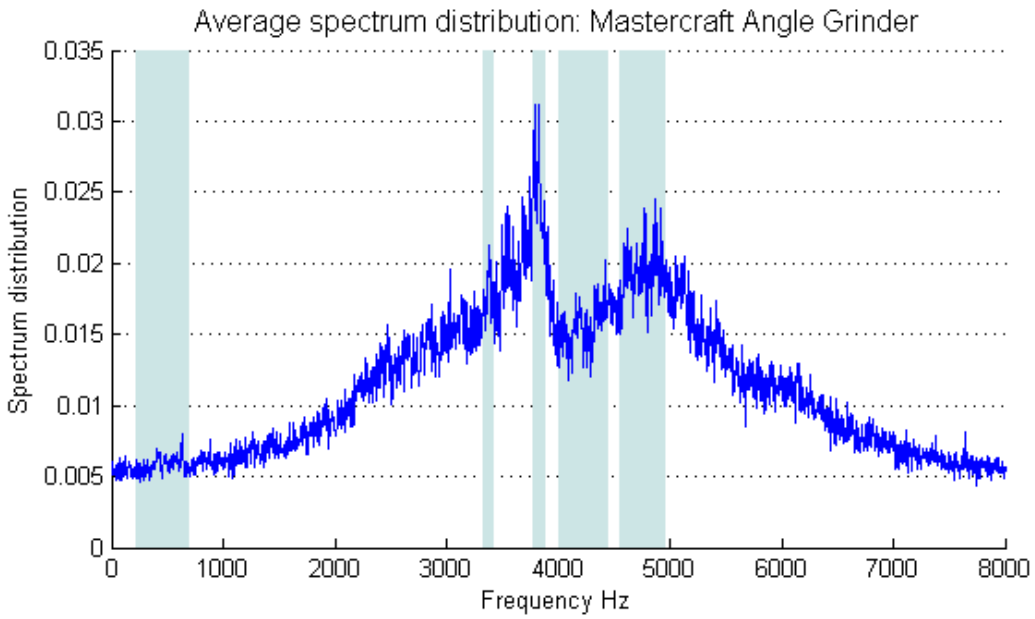


Figure 5.6: Average spectrum distribution: Mastercraft angle grinder

Filtering is the most common form of signal processing technique used to remove the frequencies in certain parts and to improve the magnitude, phase, or group delay in some other parts of the spectrum of a signal. We designed and implemented the FIR band pass filter to eliminate noise interference and to enhance the frequency characteristics of the signal to correctly recognize the construction tool type by its audio spectrum.

5.6 Audio Feature Extraction

Developing an audio classifier to recognize tool sounds in a noisy construction environment is a challenging task. Hence, robust audio features should be identified in order to make a significant classification model to recognize tool sounds more accurately. A considerable number of audio features have been used for the tasks of audio analysis and audio retrieval. Traditionally, these features have been classified in temporal features,

e.g. zero crossing rate (ZCR); energy features, e.g. short time energy (STE); spectral features, e.g. spectral moments. In addition to the traditional features listed above we used a FIR band pass filter in our classification model.

5.6.1 Zero crossing rate (ZCR)

The zero-crossing rate is a measure of the number of times the signal crosses zero or changes the sign. Periodic sounds tend to have a relatively smaller ZCR than a noisy sound. Moreover, consistent and higher range of ZCR can be identified from jigsaw and grinder tool sounds, while a hammer generates a relatively low amount. It is computed at each time frame of the signal.

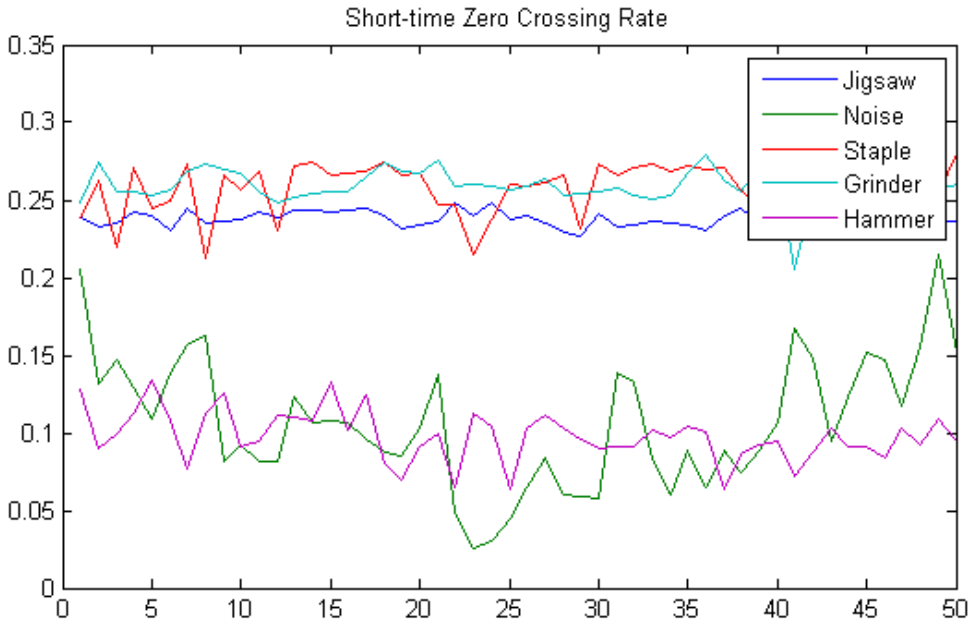


Figure 5.7: Zero crossing rate (ZCR)

5.6.1 Short time energy (STE)

STE describes signal energy at a given time and is alternatively referred to as loudness or volume of the signal. The following graph clearly shows that each tool follows a specific range of energy level while it operates.

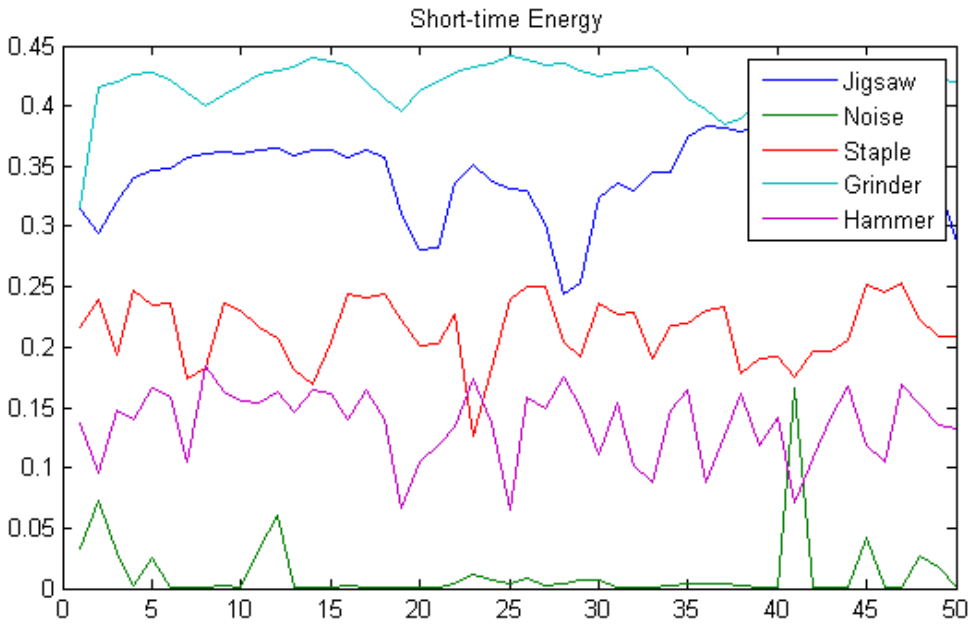


Figure 5.8: Short time energy (STE)

5.6.2 Spectral features

Four spectral moments are considered in this study: spectral centroid, variance, skewness, and kurtosis. The importance of each measure and feature of variations for each tool will be further reviewed in following sections.

5.6.2.1 Spectral centroid/mean

The spectral centroid, which is also called first spectral statistical moment or spectral mean, is a measure used in digital signal processing to characterise a spectrum. It

indicates where the "center of mass" of the spectrum is. Further, it has a robust connection with the impression of "brightness" of a sound.

$$Spectral\_Centroid(Hz), \bar{x} = \frac{\sum_{i=0}^{N} f(i)x(i)}{\sum_{i=0}^{N} f(i)} \qquad (20)$$

where, $f(i)$ represents the frequency value of bin number i, and $x(i)$ represents the center frequency of that bin. Figure 5.9 depicts variation of spectral centroids of collected data samples where staple sound reaches the top of the frequency line. Data samples of jigsaw, staple, and grinder are more consistent, while hammer sound ripples at the bottom with noise. As indicated in the figure, we can define specific frequency ranges for jigsaw and grinder models. Thus, the centroid parameter can be efficiently used in the classifier models either as continuous form or categorical form. Continuous variable is suggested for the staple gun, and categorical form is suggested for jigsaw and grinder sounds.

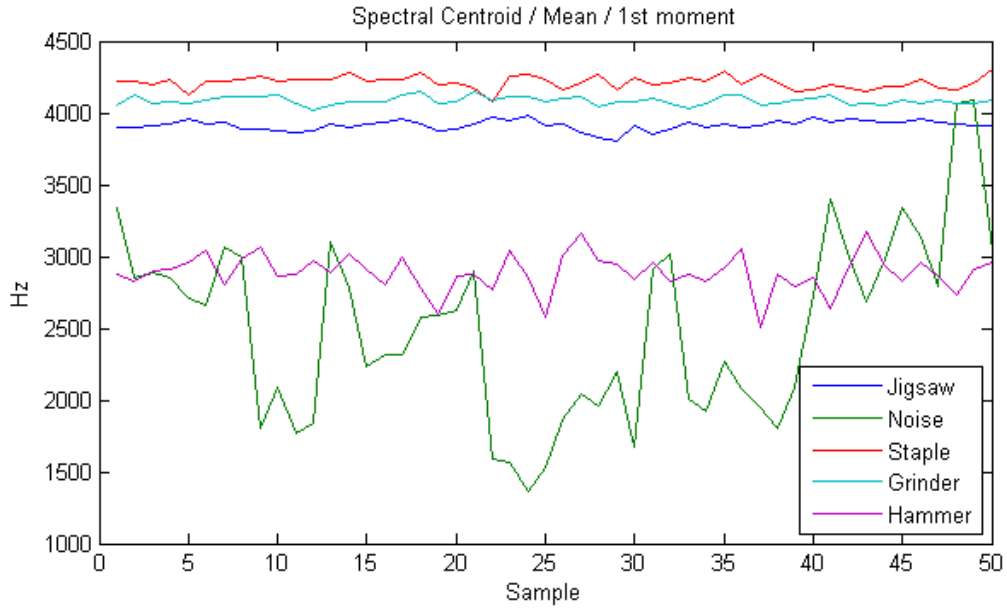


Figure 5.9: Spectral centroid/mean/1st moment

5.6.2.2 Spectral spread or variance

The spectral spread or variance is a measure of how far a set of frequencies is spread out its spectral centroid.

$$Variance, \sigma^2, m_2 = \frac{1}{\sum_{i=0}^{N} f(i)} \sum_{i=0}^{N} f(i)[x(i) - \bar{x}]^2 \qquad (21)$$

where, $\bar{x}$ is the frequency centroid and σ denotes the standard deviation. Figure 5.10 shows the variance distribution for each tool sample. Variances of the hammer and staple sounds are in the upper and lower boundaries respectively.

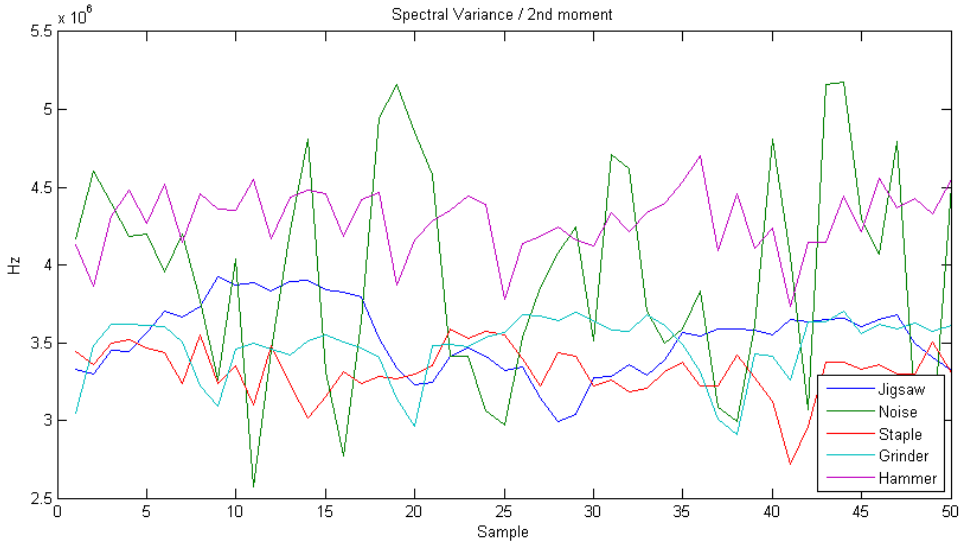


Figure 5.10: Spectral variance/spread/2nd moment

5.6.2.3 Spectral skewness

Spectral skewness is used as a measurement of the asymmetry of the frequency distribution. The skewness value can be positive or negative. Negative skew indicates that the tail on the left side of the frequency distribution function is longer than on the right side, and the bulk of the frequency values lie to the right of the mean. Positive skew explains the right tail is longer; the mass of the distribution is concentrated on the left of the figure. The following equation is used to calculate the skewness of the distribution:

$$Skewness = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{\sum_{i=0}^{N} f(i)} \sum_{i=0}^{N} f(i)[x(i) - \bar{x}]^3}{(\frac{1}{\sum_{i=0}^{N} f(i)} \sum_{i=0}^{N} f(i)[x(i) - \bar{x}]^2)^{3/2}} \qquad (22)$$

where, $\bar{x}$ is the frequency mean, m_3 is the third spectral moment, and m_2 is the variance.

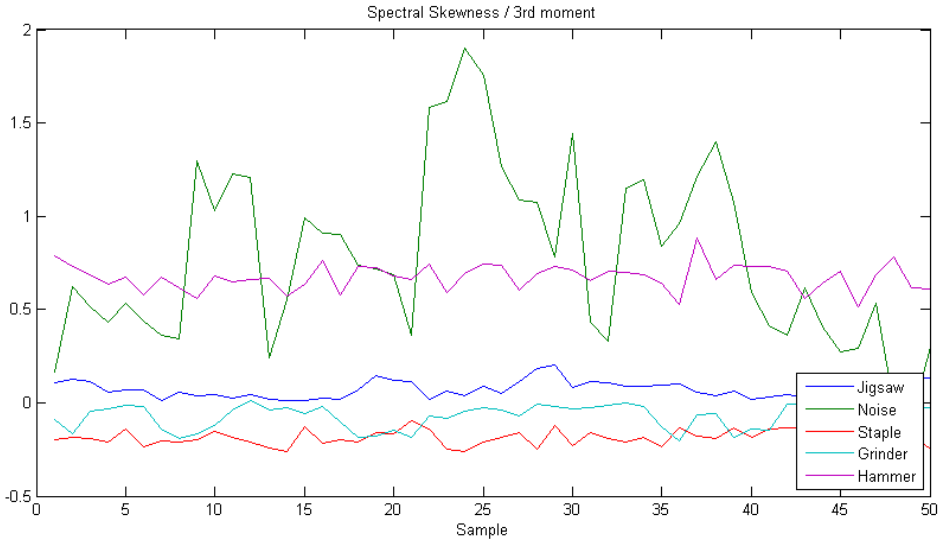


Figure 5.11: Spectral skewness/3rd moment

Hammer sound has higher positive skewness value. In other words, large components heavily reside at left side frequency band widths.

5.6.2.4 Spectral kurtosis

This is the fourth moment of the frequency distribution. Spectral kurtosis is a measure of the peakedness of the distribution and the heaviness of its tails. The formula below is applied to determine the kurtosis value of the distribution and Figure 5.12 graphically illustrates the kurtosis distribution for each tool. The kurtosis distribution does not significantly correlate with any tool sound.

$$Kurtosis = \frac{m_4}{m_2^2} = \frac{\frac{1}{\sum_{i=0}^{N} f(i)} \sum_{i=0}^{N} f(i)[x(i) - \bar{x}]^4}{(\frac{1}{\sum_{i=0}^{N} f(i)} \sum_{i=0}^{N} f(i)[x(i) - \bar{x}]^2)^2} \qquad (23)$$

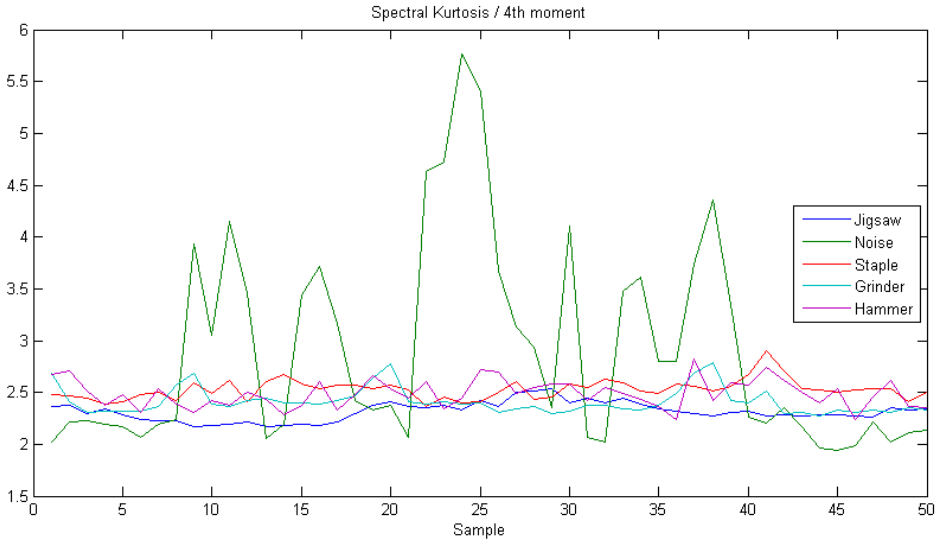


Figure 5.12: Spectral Kurtosis/4th moment

5.7 FIR Band Pass Filter (FIRBPF)

Information collected from data acquisition systems is mixed with a variety of environmental and equipment noises. Making a strategic decision always requires efficient information; therefore, the process of noise filtering is needed for information obtained from the data acquisition system. In this study, 227-tap band-pass FIR filters are designed to eliminate the noise interferences and pass-band frequencies of each tool discussed in following sections. The purpose of estimating the spectral density is to detect any periodicities in the data, by observing peaks at the frequencies corresponding to these periodicities. Several studies have already been done to explore the flexibility of FIR filter design (He, Zhang, & Zeng, 2011; Hong, Zhaonan, & Xiangli, 2010).

5.7.1 Fundamentals of band pass filter

To filtrate noises in a sound acquisition system, a band-pass FIR filter is designed, which is based on a distributed algorithm and numerous pass-band frequencies of selected construction tools. The design process is as follows: FIR filter is a linear time-invariant system and the input and output relationship in time domain is convolution computing, which is shown as formula:

$$y(n) = b_0 x[n] + b_1 x[n-1] + \ldots + b_N x[n-N] = \sum_{i=0}^{N} b_i \times x[n-i] \qquad (24)$$

where, N is the filter order. An Nth-order filter has (N+1) terms of coefficients; x(n) is the input sequence; y(n) is the output sequence; b_i are the filter coefficients that make up impulse response. The requirement of minimum attenuate in stop band of the band-pass FIR filter is 40 db. Figure 5.143 graphically explains the process of FIR filter for Nth order.

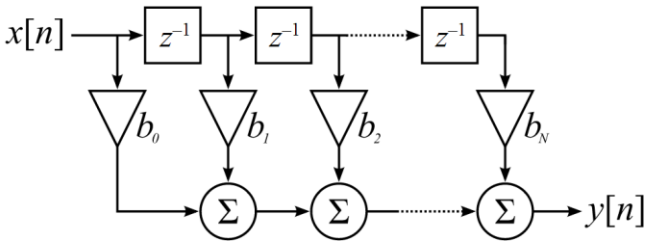


Figure 5.13: FIR Filter of Nth order

5.7.2 Design indexes of the FIR filter

Filter type: band-pass FIR digital filter; Sampling frequency Fs: 16000Hz; Filter order: 227; Stop-band minimum attenuation ds: 40db; Pass-band maximum attenuation dp: 1db.

Figure 5.14 graphically explains filter specifications used in the FIR band pass filter design such as stop band frequencies and pass band frequencies and attenuation levels.

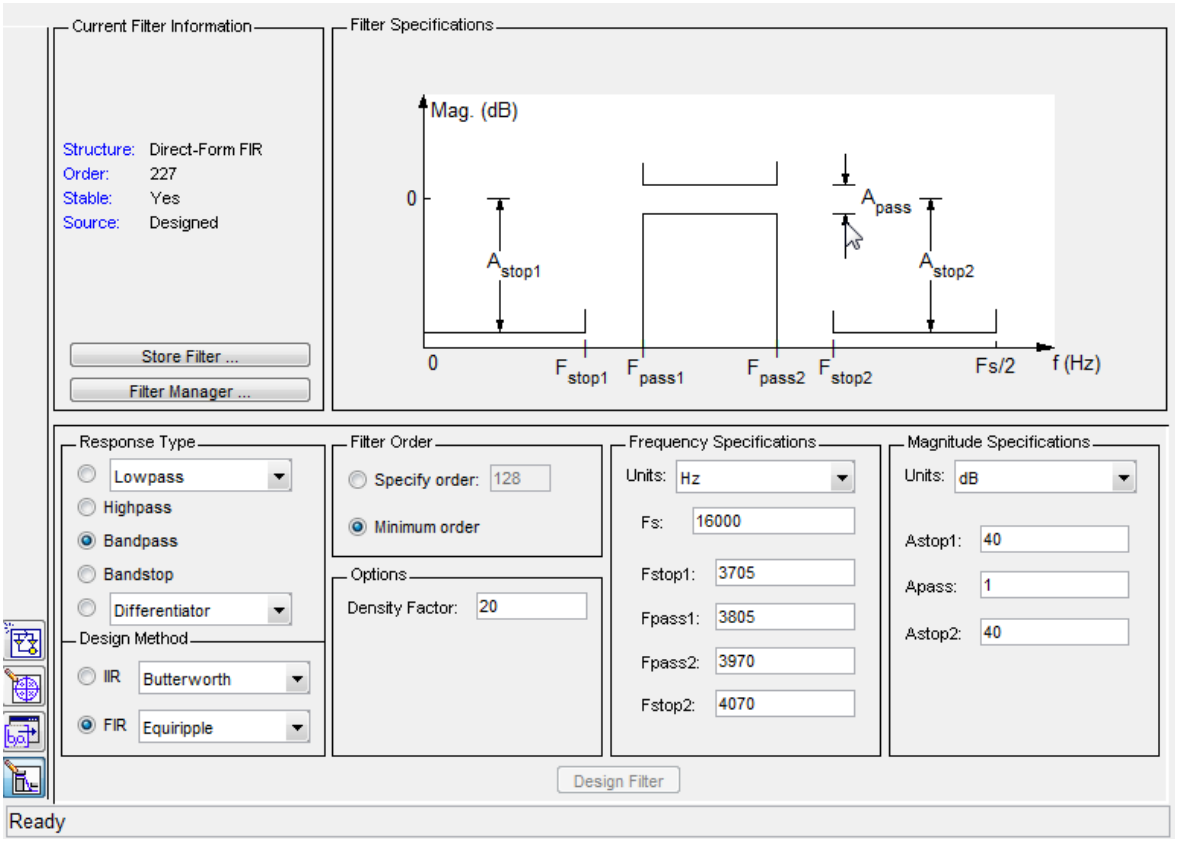


Figure 5.14: FIR filter specifications

Table 5.2: Selected band widths of FIR band pass filter

Equipment	Var. Code	FStop1 (Hz)	FPass1 (Hz)	FPass2 (Hz)	FStop2 (Hz)
Jigsaw	J1	1470	1570	1752	1852
	J2	2890	2990	3375	3475
	J3	3586	3686	3690	3790
	J4	4515	4615	5015	5115
Staple	S1	3225	3325	3425	3525
	S2	3705	3805	3970	4070
	S3	4135	4235	4470	4570
	S4	4590	4690	4717	4817
	S5	5040	5140	5325	5425
Grinder	G1	130	230	700	800
	G2	3225	3325	3425	3525
	G3	3675	3775	3885	3985
	G4	3900	4000	4450	4550
	G5	4460	4560	4965	5065
Hammer	H1	161	261	321	421
	H2	392	492	532	632
	H3	540	640	700	800
	H4	720	820	900	1000

5.7.3 Filter coefficients

According to the design indexes, the filter coefficients are designed and performed by using the FDA tool in MATLAB. We chose the band-pass filter and Equiripple method. As for the abovementioned design requirement, the FIR band pass filter for each tool has been designed using the minimum order. As an example, a magnitude response of 227^{th} order FIR filter for the 3^{rd} band width of staple gun is illustrated in the Figure 5.15.

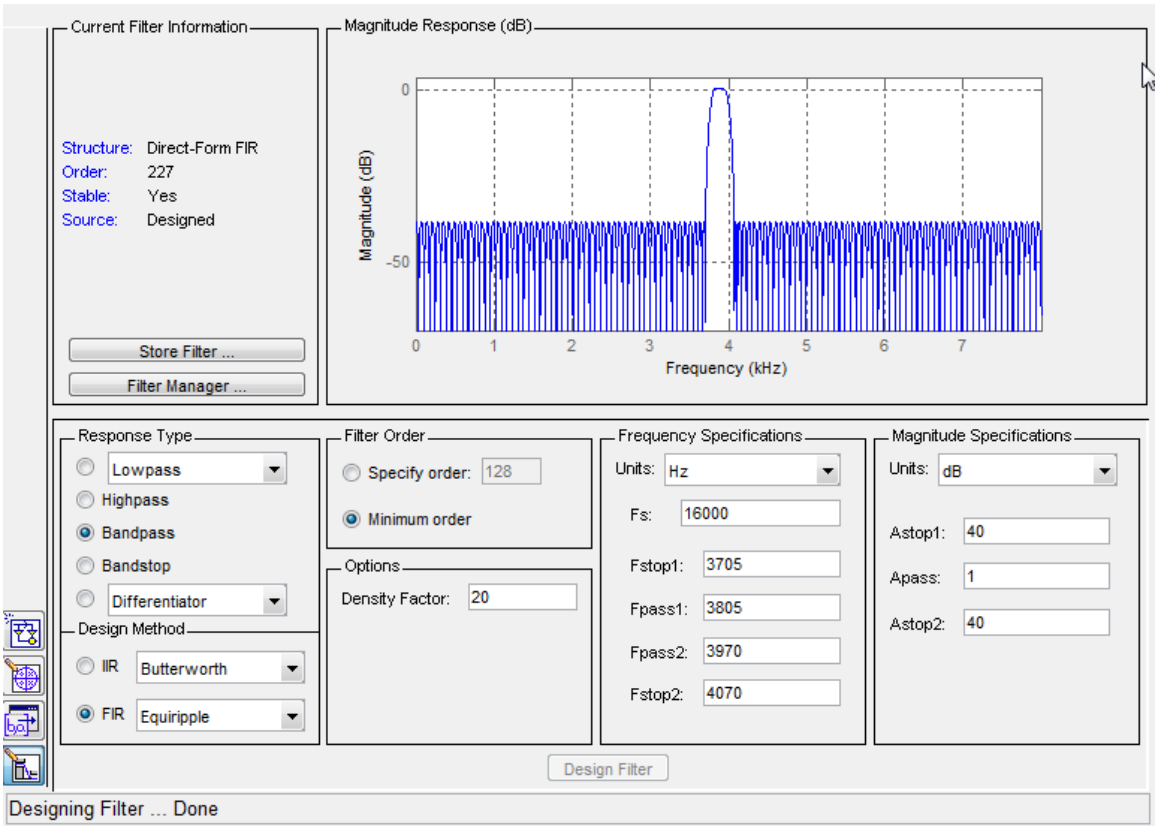


Figure 5.15: FIR Filter of 227th order for 3rd band width of staple gun

The application of designed FIR filter in a real audio signal is illustrated in following figures. The top image in the Figure 5.16 shows the spectrum distribution of a raw staple sound audio signal, while the bottom image depicts the effectiveness of the filter by separating peaks of specified frequency band width (3805-3970Hz).

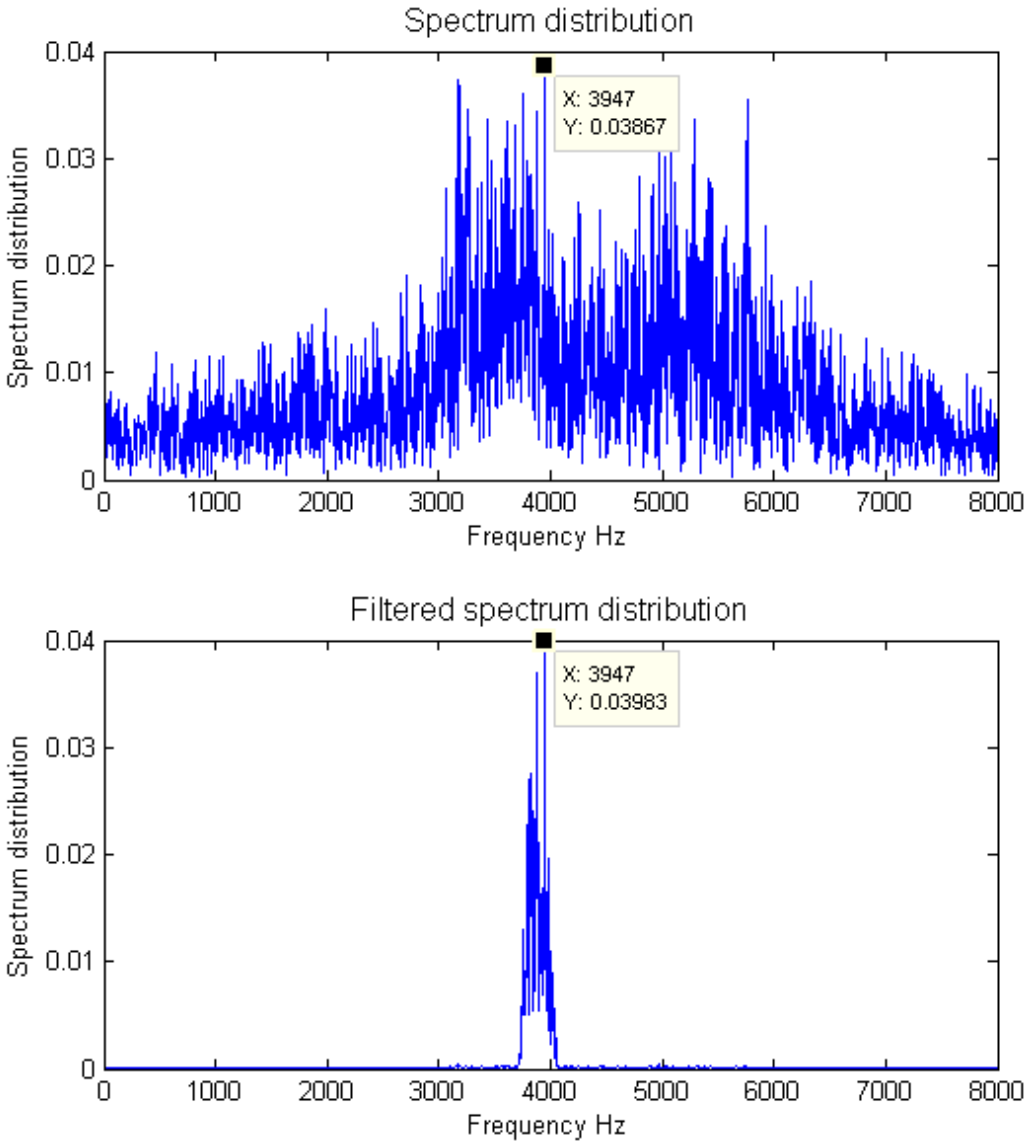


Figure 5.16: Filtered spectrum distribution

5.8 Audio Feature Selection

Table 5.3 lists the entire feature set composition. All the features are extracted from 256ms analysis frames at a sampling frequency of 16000 Hz with 32ms overlap. The feature vector dimension is then reduced by logistic regression procedure. We retain the

most significant audio features to create the more accurate classification model and start the analysis from the full set of 11 features. It is desirable to keep a small number of feature variables in a model in order to reduce the computational complexity of the feature extraction process and to limit the over fitting produced by the increasing number of parameters associated with features in the classification model.

Table 5.3: List of selected audio features

Feature Type	Features	Number of Features
Temporal	ZCR	1
Energy	STE	1
Spectral	Centroid/mean	1
	Variance/spread	1
	Skewness	1
	Kurtosis	1
FIR Band pass filter	Up to 5 frequency band widths	5
Total number of features		11

5.9 Multivariate Modelling

Multivariate statistical models were developed for the prediction of tool sound occurrence in the context of construction activity type assessment. The models take into account the factors captured by different tool operations in a noisy environment. Logistic regression was the preferred statistical procedure for this study because the technique is suited to models with a dichotomous outcome (tool or non-tool) with multiple predictor variables that include a mixture of continuous and categorical parameters. In addition, logistic regression is also especially appropriate for case-control studies because it allows the use of samples with different sampling fractions depending on the outcome variable without giving biased results. In this study, it allows the sampling fractions of accident

flights and that of normal flights to be different. This property is not shared by most other types of regression analysis (Nagelkerke et al., 2005).

We used backward stepwise logistic regression as the initial step to calibrate the tool sound models because of the predictive nature of the research. The selected technique is able to identify relationships missed by forward stepwise logistic regression (Hosmer & Lemeshow, 2000). However, there are some special cases in which a manual variable selection approach has to be applied. The predictor variables were entered by blocks, each consisting of related factors, as shown in Table 5.4, such that the change in the model's substantive significance could be observed as the variables were included.

Table 5.4: Blocks of variables – Audio classifier

Block variables	**Variable**	**Description**	**Type**
Block-Jigsaw	J1, J2, J3, J4	FIR band pass filter for jigsaw	Continuous
Block-Staple	S1, S2, S3, S4, S5	FIR band pass filter for staple	Continuous
Block-Grinder	G1, G2, G3, G4, G5	FIR band pass filter for angle grinder	Continuous
Block-Hammer	H1, H2, H3, H4	FIR band pass filter for hammer	Continuous
Block-Spectral	C	Centroid/mean (kHz)	Continuous
	VAR	Variance/spread (kHz2)	Continuous
	SK	Skewness	Continuous
	KUR	Kurtosis	Continuous
	CJ	C<3.878kHz & C>3.950kHz	Categorical
	CG	C<4.056kHz & C>4.115kHz	Categorical
	EH	E<0.874kJ & E>1.364kJ	Categorical
Temporal	ZCR	Zero crossing rate	Continuous
Energy	E	STE (kJ)	Continuous

The previous section discussed the application of a FIR band pass filter for each tool sound. All the frequency band widths in each filter were used as predictor variables in

this logistic regression model. Apart from the general continuous spectral moments, we used three categorical variables: CJ, CG, and EH. In order to emphasize the difference of the spectral first moment between tools, we proposed frequency ranges for jigsaw and grinder as indicated in Table 5.4. Likewise, we proposed another categorical variable for energy, EH, in order to enhance the characteristics of the hammer sound.

We used the IBM SPSS Statistic software (version 20) to analyze the logistic regression of this study. Statistical software SPSS begins by conducting backward stepwise logistic regression, removing non-significant variables of that block before conducting backward stepwise logistic regression on the remaining variables.

With the model coefficients, the probability formula for tool sound occurrence could be obtained. For each tool sound model:

$$p(ToolsoundOccurance) = \frac{1}{1+e^{-z}} \qquad (25)$$

where, $z = b_0 + b_1(valriable_1) + b_2(valriable_2) + \cdots + b_n(valriable_n)$, b_0 is the constant and b_1 to b_n are the corresponding parameter coefficients.

5.9.1 Tool sound classifier

This section provides a detailed analysis of each tool. Variable selection, model comparison based on different criterion, independent variable correlation, and predicted model accuracy will be also reviewed.

5.9.1.1 Mastercraft jigsaw

Table 5.5 shows the output resulting from most of the candidate predictor variable selection in the equation. We tested and rejected three spectral parameters: centroid, skewness, and kurtosis. Instead we added CJ and variance to represent the spectral shape. The significance of each variable is measured using a Wald statistic. $p = 0.05$ is used as the cut off criterion for not including variables in the equation. Hence, variable significance in all models listed in the table is accepted. As shown in Figure 5.17, case numbers three, five, and six can be considered competitive models in terms of the basic selection criterion as indicated in Table 5.5. Variable significance, Nagelkerke R squared, and independent variable correlations are considered equivalent. We selected case number 3 as the final logistic model for a jigsaw sound classifier since it consists of a strong discriminative variable CJ. Further, CJ is the most significant variable in this model (Wald 16.428). All coefficients and other properties of selected variables are listed in Table 5.6.

Table 5.5: Step wise analysed models: Mastercraft Jigsaw

No	Variables added	R2	Model Sig.	Max Var Sig.	Max Corr	Accuracy %
1	J1, CJ	0.895	0.555	0.000	0.518	95.2
2	J1, J2, CJ	0.943	0.610	0.001	0.518	98.0
3	J1, J2, E, CJ	0.952	0.995	0.029	0.597	98.0
4	J1, J2, E, CJ, VAR	0.959	0.425	0.039	0.597	98.0
5	J1, J2, E, VAR	0.915	0.898	0.000	0.597	96.4
6	J1, J2, J3, E, VAR	0.929	0.996	0.008	0.597	95.6
7	J1, J2, J4, CJ	0.950	1.000	0.054	0.783	97.6

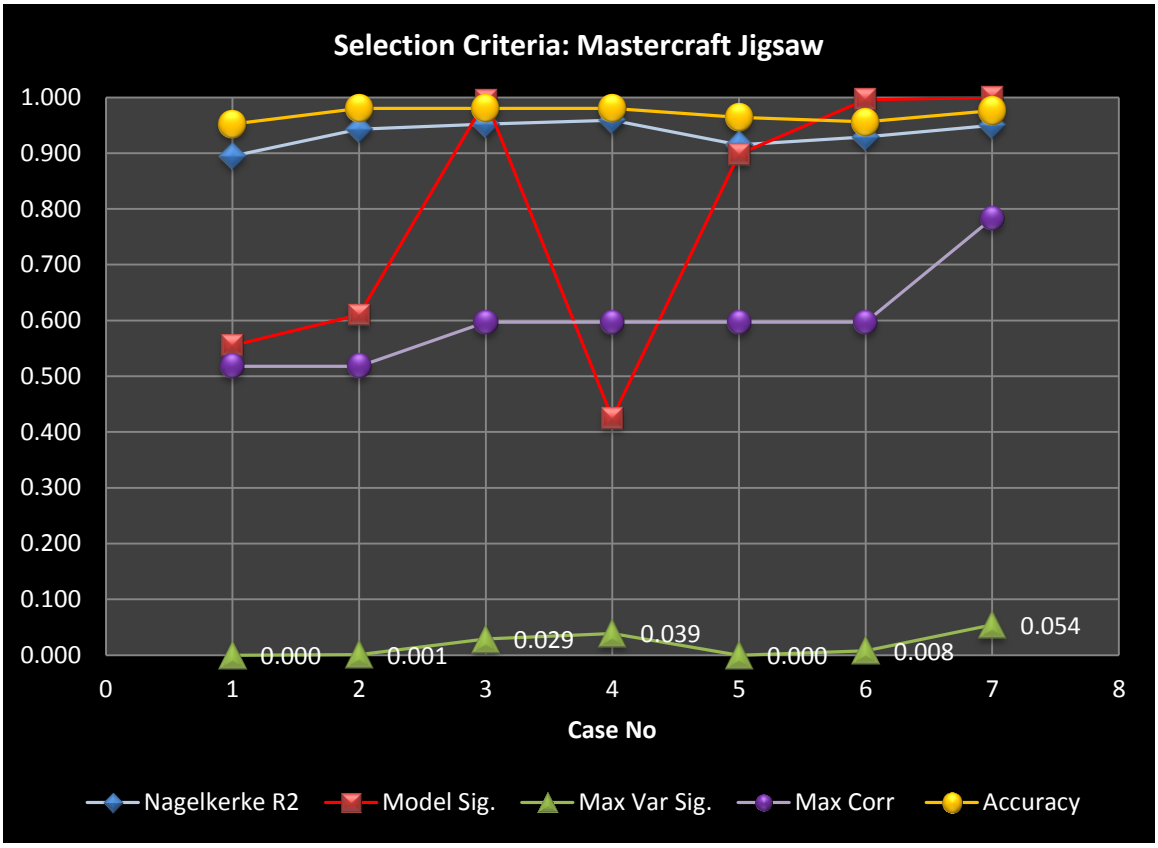


Figure 5.17: Model summary: Mastercraft Jigsaw

Table 5.6: Variables in the equation: Mastercraft Jigsaw

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Jigsaw1	12.5987	3.488	13.043	1	0.0003	296178.76
Jigsaw2	10.5082	3.484	9.097	1	0.0026	36614.43
CJ	-18.8865	4.660	16.428	1	0.0001	0.00
E	1.3478	0.619	4.738	1	0.0295	3.85

One way to make use of the information in the model is to use the results to predict the probability of a jigsaw sound that comes from the construction job site. To calculate this value, use the prediction equation shown below.

The z for the Mastercraft jigsaw tool sound probability formula is;

$$z_{Jigsaw} = 0 + 12.5987(Jigsaw1) + 10.5082(Jigsaw2) - 18.8865(CJ) + 1.3478(E)$$

In statistical terms, a correlation is a mathematical measure of the strength of association between two quantitative variables. Correlation between each variable in a model is an important factor when deciding variables into a model. We used Pearson's correlation coefficient. Following table shows the correlation matrix of the selected variables and the maximum correlation is reported between J2 and J4: 0.783.

Table 5.7: Correlation matrix: Mastercraft jigsaw

	J1	J2	J3	J4	E	CJ	VAR
J1	1.000	0.410	0.092	0.310	0.333	0.518	0.141
J2	0.410	1.000	0.552	0.783	0.597	0.419	0.501
J3	0.092	0.552	1.000	0.477	0.390	0.044	0.365
J4	0.310	0.783	0.477	1.000	0.772	0.318	0.509
E	0.333	0.597	0.390	0.772	1.000	0.351	0.405
CJ	0.518	0.419	0.044	0.318	0.351	1.000	0.140
VAR	0.141	0.501	0.365	0.509	0.405	0.140	1.000

In order to measure the accuracy of the model we tested predicted probabilities of success using the constructed model with coefficients.

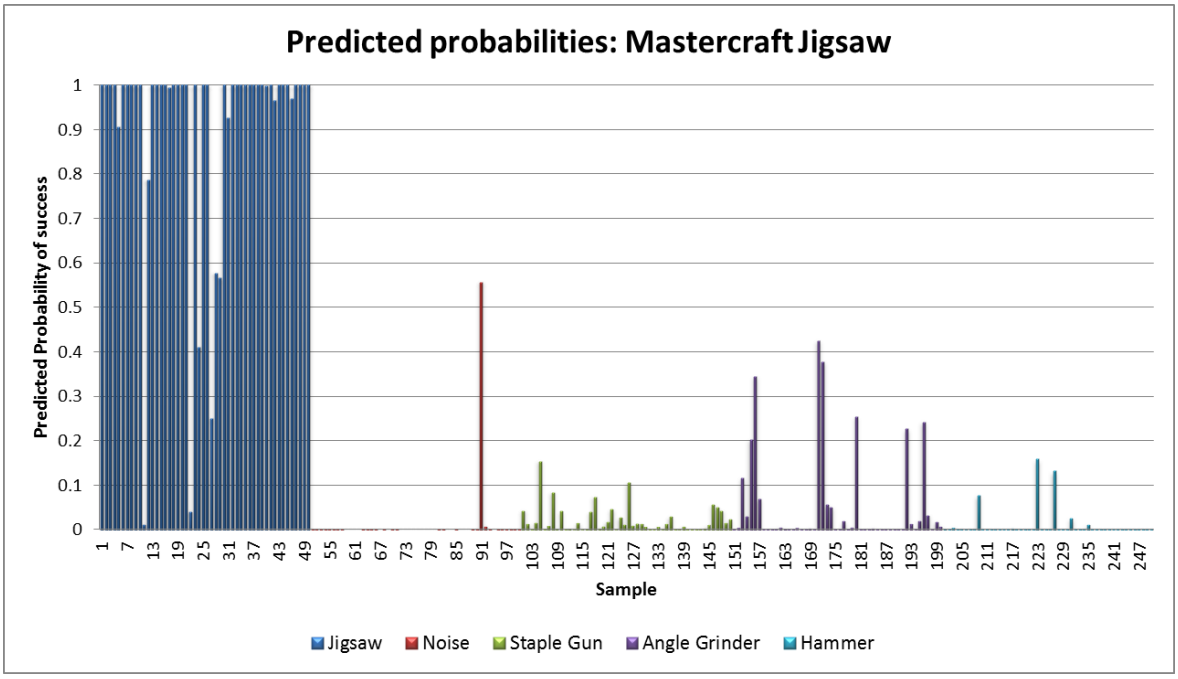


Figure 5.18: Predicted probabilities of Jigsaw model

With this selected model, the Mastercraft staple gun and angle grinder sounds have some potential of being tracked as a Mastercraft jigsaw tool sound. From Figure 5.18 we see that on two occasions the jigsaw sound completely missed being tracked from the model. According to the receiver operating characteristic (ROC) curve illustration, 0.25 is set as the probability cut off level and the following classification table demonstrates the detailed accuracy level of the model. Overall accuracy of the model gained 97.2% while achieving 97.5% of the sensitivity or true positive rate of jigsaw sounds.

Table 5.8: Classification table: Mastercraft Jigsaw

Observed		Predicted		
		Observed Jigsaw		Percentage Correct
		.00	1.00	
Observed Jigsaw	.00	195	5	97.5
	1.00	2	48	96.0
Overall Percentage				97.2
a. The cut value is 0.25				

5.9.1.2 Mastercraft staple

Table 5.9 shows the different models we tested to find the best suitable model for recognizing the staple sound. p = 0.05 is used as the cut off criterion for not including variables in the equation and variable significance in all models except the last one can be considered a statistically acceptable amount. Case numbers 2, 3, and 4 can be identified as potential models to be the staple sound classifier. According to the summary of models illustrated in Figure 5.19, variable selection is the key deciding factor of these three models as no major differences are flagged in the criterion. Since different wave energies can be observed on job sites, strong FIR filter magnitudes improve the robustness of the final model. Hence case number 4 is selected and Table 5.10 depicts the properties of selected variables in the equation.

Table 5.9: Step wise analysed models: Staple

No	Variables added	R2	Model Sig.	Max Var Sig.	Max Corr	Accuracy %
1	S5,E,VAR	0.845	0.416	0.000	0.563	92.8
2	S2,S5,E,VAR	0.928	0.945	0.000	0.763	96.8
3	S1,S2,S5,E,VAR	0.937	0.978	0.026	0.765	96.4
4	S1,S2,S4,S5,E,VAR	0.953	0.999	0.009	0.765	96.4
5	S2,S5,E,VAR,C	0.948	0.852	0.016	0.837	97.2
6	S2,S4,S5,E,VAR,C	0.966	1.000	0.019	0.837	97.6
7	S1,S2,S4,S5,E,VAR,C	0.972	1.000	0.086	0.837	98.4

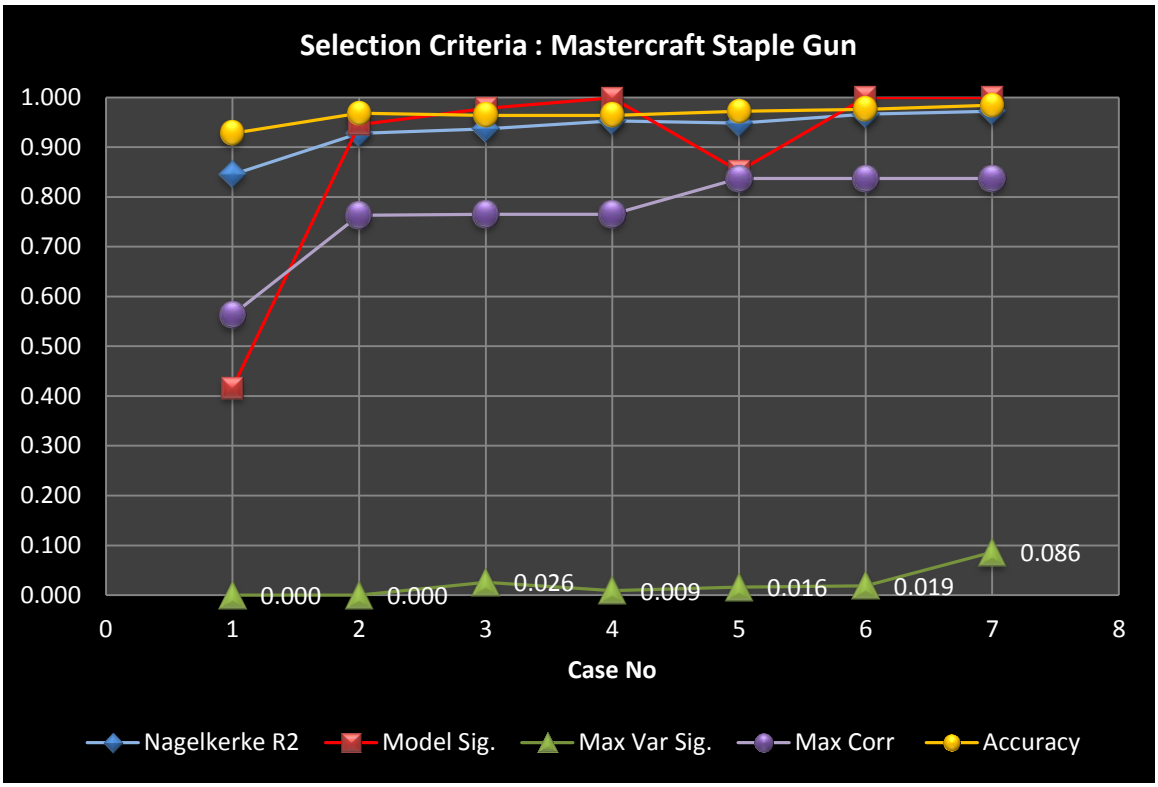


Figure 5.19: Selection criteria comparison: Mastercraft staple gun

Table 5.10: Variables in the equation: Mastercraft staple

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Staple1	7.1649	2.598	7.607	1	0.0058	1293.24
Staple2	13.3360	3.884	11.789	1	0.0006	619093.91
Staple4	-10.9256	4.209	6.738	1	0.0094	0.00
Staple5	11.3485	2.855	15.796	1	0.0001	84835.89
E	-2.2514	0.630	12.753	1	0.0004	0.11
VAR	-4.0123	1.036	15.011	1	0.0001	0.02

According to Wald statistics, Staple5 variable is the most significant variable. The wave energy, Staple4, and variance have the inverse relationship with the staple sound. Moreover, energy (E) has been used to differentiate the staple sound from more powerful

tools such as the grinder and jigsaw. On the other hand, three selected FIR filters increase the robustness of the model. A correlation matrix and predicted probabilities are shown in and Figure 5.20 respectively. Correlation coefficients between FIR filter variables are significantly higher than any other models.

The z for the Mastercraft staple tool sound probability formula is:

$$z_{Staple} = 0 + 7.1649(Staple1) + 13.3360(Staple2) - 10.9256(Staple4) + 11.3485(Staple5) - 2.2514(E) - 4.0123(Variance)$$

Table 5.11: Correlation matrix: Mastercraft staple

	S1	S2	S4	S5	C	VAR	E
S1	1.000	0.765	0.663	0.765	0.804	0.499	0.577
S2	0.765	1.000	0.640	0.763	0.837	0.534	0.599
S4	0.663	0.640	1.000	0.682	0.743	0.489	0.695
S5	0.765	0.763	0.682	1.000	0.829	0.563	0.510
C	0.804	0.837	0.743	0.829	1.000	0.473	0.774
VAR	0.499	0.534	0.489	0.563	0.473	1.000	0.405
E	0.577	0.599	0.695	0.510	0.774	0.405	1.000

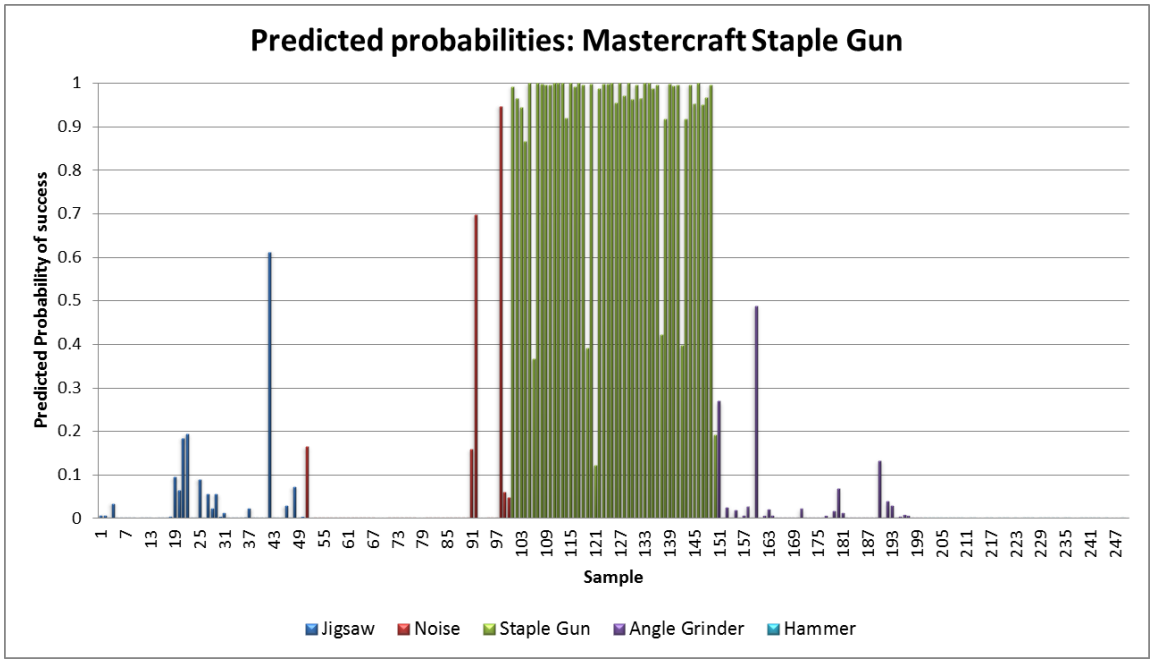


Figure 5.20: Predicted probabilities of Mastercraft staple gun model

The jigsaw sound has been correlated with the staple sound on a few occasions, as depicted in the graph. We assumed the reason for this correlation would be that the selected FIR band widths are slightly overlapped (S4 and J4). Selecting a cut off level plays a major role when it comes to the discussion of the accuracy level of the model. Considering the ROC curve discussed in the summary section, 0.19 was set as the cut off value to this model. As a result overall model accuracy pointed to 97.2%.

Table 5.12: Classification table: Mastercraft staple

Observed		Predicted		
		Observed Staple		Percentage Correct
		.00	1.00	
Observed Staple	.00	194	6	97.0
	1.00	1	49	98.0
Overall Percentage				97.2
a. The cut value is 0.19				

5.9.1.3 Mastercraft angle grinder

This is the most powerful and loudest tool used in this study. Hence it is important to highlight this audio feature when selecting variables to the model. Energy (E) represents the power of the tool. Thus adding E dramatically increases the model accuracy. Table 5.13 below shows the variable groups tested in each case and Figure 5.21 illustrates the corresponding goodness of fit, significance, and accuracy of the model. In this variable selection we added Grinder1 to create a model against hammer and noise sound. Lower magnitudes in lower frequencies differentiate these sounds. This negative relationship is proven with the negative variable coefficient indicated in Table 5.14. Case numbers 1, 3, and 6 are in the competitive level. Since the grinder is the loudest tested tool, more strong audio features should be reinforced with the energy feature. Thus, case number 6 is selected with three FIR filters.

Table 5.13: Step wise analysed models: Angle grinder

No	Variables added	R2	Model Sig.	Max Var Sig.	Max Corr	Accuracy %
1	G2, CG, E	0.942	0.960	0.000	0.580	96.4
2	G1,G2,E	0.840	0.904	0.005	0.795	89.2
3	G1,G2,G4,E	0.858	0.973	0.005	0.830	92.0
4	G2,G3,E,Const	0.905	0.713	0.022	0.690	98.4
5	G2,G3,CG,E	0.956	0.738	0.011	0.694	99.2
6	G1,G2,G3,CG,E	0.963	1.000	0.074	0.795	99.2
7	G1,G2,G3,G5,CG,E	0.966	1.000	0.216	0.807	98.8

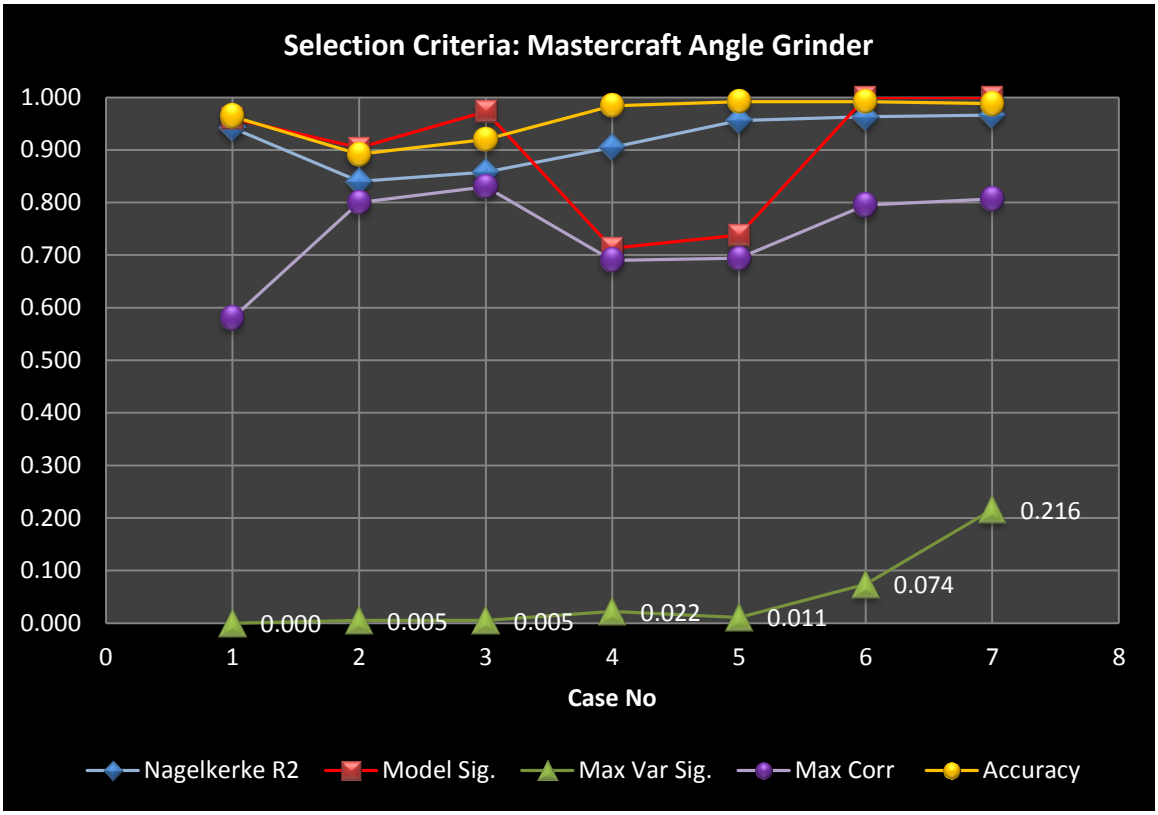


Figure 5.21: Selection criteria comparison: Mastercraft angle grinder

Table 5.14 demonstrates the properties of each variable in the equation. Wald statistics prove that energy (E) is the most significant variable in this model.

Table 5.14: Variables in the equation: Angle grinder

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
G1	-22.0940	12.3530	3.1990	1	0.0740	0.00
G2	-9.8700	4.5700	4.6650	1	0.0310	0.00
G3	7.8800	3.2340	5.9360	1	0.0150	2643.73
E	6.6060	2.0160	10.7350	1	0.0010	739.81
CG	-15.6180	4.9820	9.8280	1	0.0020	0.00

The z for the Mastercraft angle grinder sound probability formula is:

$$z_{Grinder} = 0 - 22.0940(Grinder1) - 9.8700(Grinder2) + 7.8800(Grinder3) + 6.6060(E) - 15.6180(CG)$$

Absolute values of Pearson correlation coefficients between related independent variables are listed in Table 5.15.

Table 5.15: Correlation matrix: Angle grinder

	G1	G2	G3	G4	G5	E	CG
G1	1.000	0.744	0.750	0.775	0.807	0.795	0.319
G2	0.744	1.000	0.694	0.828	0.731	0.577	0.120
G3	0.750	0.694	1.000	0.754	0.682	0.622	0.306
G4	0.775	0.828	0.754	1.000	0.822	0.602	0.111
G5	0.807	0.731	0.682	0.822	1.000	0.764	0.320
E	0.795	0.577	0.622	0.602	0.764	1.000	0.468
CG	0.319	0.120	0.306	0.111	0.320	0.468	1.000

The following figure displays the predicted probabilities for the model after the above equation is applied.

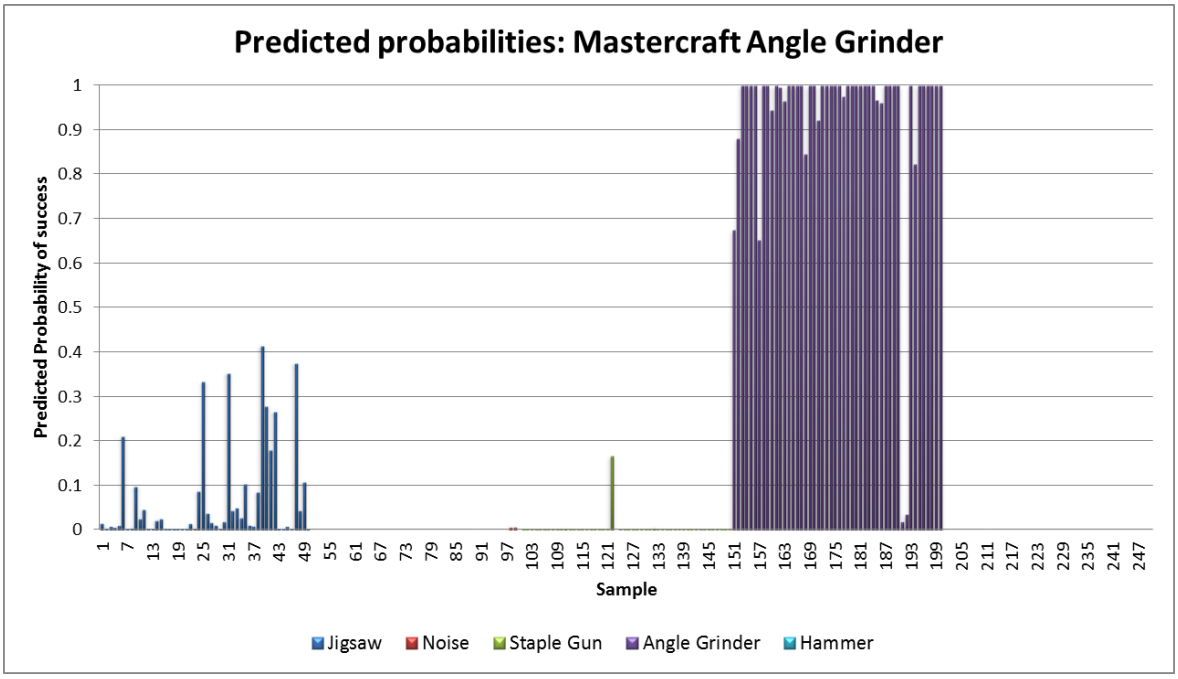


Figure 5.22: Predicted probabilities of Angle Grinder model

Figure 5.22 clearly indicates that the selected model precisely differentiates the grinder sound from most of the other tested sounds and the noisy construction environment. Moreover, this indicates that the predicted probability for the grinder sound exceeds 0.6 in most cases. In addition, the jigsaw sound has some relationship with this constructed model. We collected sound samples from the grinder at a mixture of frequencies, though the highest machine frequency is 11000rpm. Similarity between the jigsaw and the grinder is increased because of operations under lower frequencies. This clarifies the reason for detecting the jigsaw in the grinder model. Considering the above chart and ROC curve, 0.535 was suggested as the cut off level to this model. Table 5.16 demonstrates the detailed accuracy level of the model with the selected cut off level. The overall prediction rate has a 99.2% success while achieving 100% model precision.

Table 5.16: Classification table: Angle grinder

Observed		Predicted		
		Observed Grinder		Percentage Correct
		.00	1.00	
Observed Grinder	.00	200	0	100.0
	1.00	2	48	96.0
Overall Percentage				99.2
a. The cut value is 0.535				

5.9.1.4 Hammer

As shown in Figure 5.5 the spectrum frequency components are highly concentrated in the low frequency band. As a negative finding, the spectrum shape of the hammer sound is more similar to the background noise pattern. According to the skewness and centroid distribution, hammer sound is significantly different than other selected tool spectrum shapes. However, Pearson correlation coefficient reveals that both skewness and centroid variables are highly correlated. Hence we tested a few models using only the centroid variable as it is the most significant among them. In order to avoid the false detection over background noise we tested models with added energy variable, which is a powerful discriminative variable for noise. EH is a categorical variable that represents energy of the wave excluding the range of 0.874kJ < E>1.364kJ. Results of the forward and backward stepwise analysis are shown below. All cases listed in the tables met the selection criteria. Case numbers 3 and 6 depict similar performance. We selected case number 3 as it consisted of 2 discriminative variables (i.e. EH, H3) to the noise sound.

Table 5.17: Step wise analysed models: Hammer

No	Variables added	R2	Model Sig.	Max Var Sig.	Max Corr	Accuracy %
1	H1,EH	0.928	0.847	0.000	0.530	96.0
2	H1,H4,EH	0.944	0.948	0.005	0.690	98.8
3	H1,H3,H4,EH	0.988	1.000	0.017	0.690	99.2
4	H1,H4,C,	0.930	0.931	0.000	0.580	95.6
5	H1,H3,H4,C	0.945	1.000	0.011	0.580	96.8
6	H1,H4,C, EH	0.984	0.998	0.040	0.690	98.8

Figure 5.23: Selection criteria comparison: Hammer

Table 5.18 shows the properties of selected variables in the equation. EH is the most significant variable in the model. H1 and H3 got approximately similar significance levels. The lower the H3 values, the higher the chances of it being a hammer sound.

Table 5.18: Variables in the equation: Hammer

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Hammer1	19.5660	8.114	5.814	1	0.0160	314218578.27
Hammer3	-61.9380	25.622	5.843	1	0.0160	0.00
Hammer4	28.4400	11.925	5.688	1	0.0170	2246314911564.47
EH	-15.5210	6.371	5.936	1	0.0150	0.00

p = 0.05 is used as a cut off criterion for including variables in the equation. This indicates that all the variables are statistically significant and similar predictors of the probability of being a hammer sound.

The z for the hammer sound probability formula is:

$$z_{Hammer} = 0 + 19.566(Hammer1) - 61.9380(Hammer3) + 28.4400(Hammer4) - 15.5210(EH)$$

The following table illustrates the correlation matrix of the variables used in the selected model.

Table 5.19: Correlation matrix: Hammer

	H1	H2	H3	H4	C	EH
H1	1.000	0.427	0.278	0.583	0.567	0.526
H2	0.427	1.000	0.440	0.494	0.663	0.443
H3	0.278	0.440	1.000	0.449	0.310	0.256
H4	0.583	0.494	0.449	1.000	0.388	0.690
C	0.567	0.663	0.310	0.388	1.000	0.313
EH	0.526	0.443	0.256	0.690	0.313	1.000

The predicted probabilities for the selected binomial logistic regression model are illustrated in Figure 5.24. Analyzing the ROC curve, 0.115 is set as the cut off level for the hammer sound recognition. Further, hammer sound could be differentiated noticeably

from all the other tested three tools, while background noise has some potential of false detection as a hammer sound.

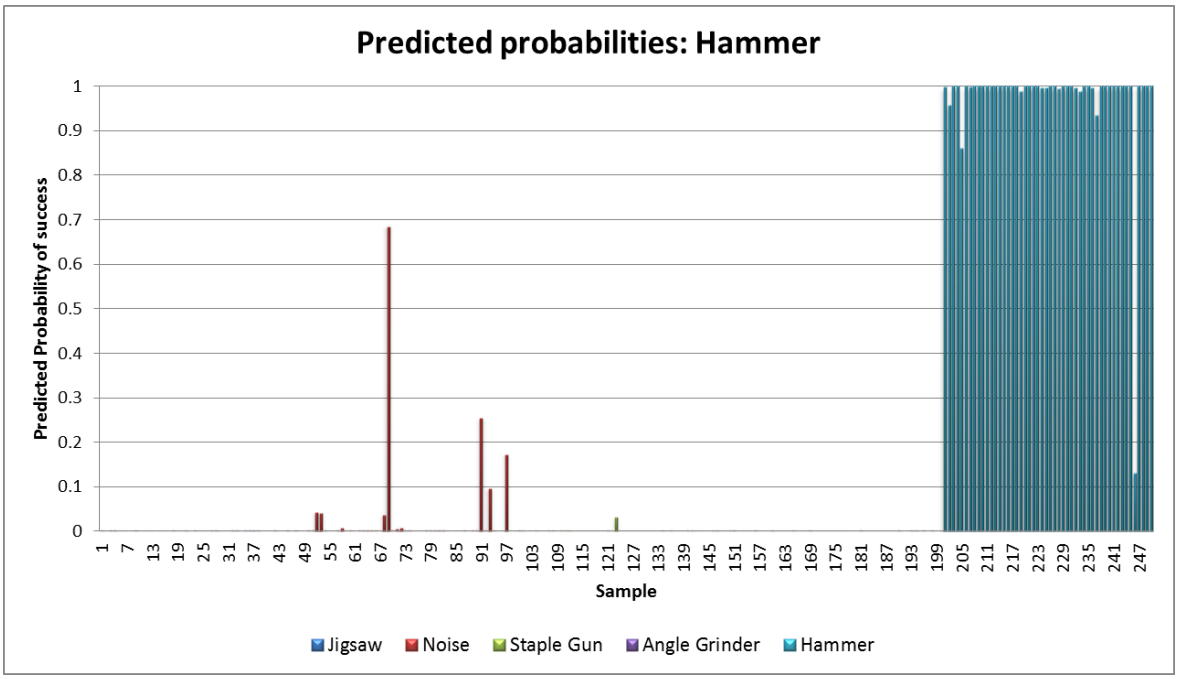


Figure 5.24: Predicted probabilities of Hammer model

Table 5.20 demonstrates the contingency level of the model. Overall model accuracy is 98.8% while 100% true positive rate in the noisy construction environment.

Table 5.20: Classification table: Hammer

Observed		Predicted		
		Observed Hammer		Percentage Correct
		.00	1.00	
Observed Hammer	.00	197	3	98.5
	1.00	0	50	100.0
Overall Percentage				98.8
a. The cut value is 0.115				

5.9.2 Classifier performance

Our four classification models produce a continuous output to which different thresholds may be applied to predict correct tool sounds (class membership). To distinguish between the actual class and the predicted class, we define P positive instances and N negative instances for the class predictions produced by a model. Given a classifier and an instance, there are four possible outcomes formulated in a 2×2 contingency table or confusion matrix, as follows:

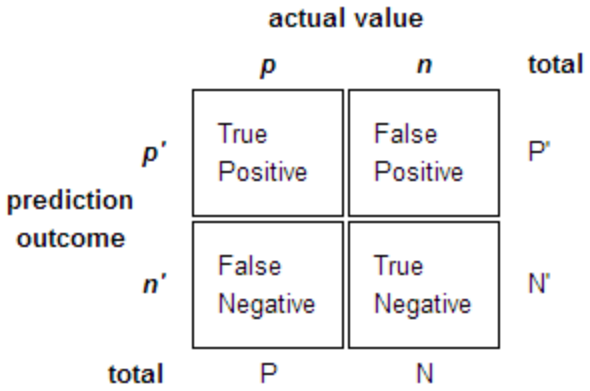


Figure 5.25: 2×2 contingency table

The following table illustrates the terminologies and derivations from the above table. The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

Table 5.21: Terminology and derivations from the contingency table

Sensitivity or true positive rate (TPR)	$TPR = TP / P = TP / (TP + FN)$
False positive rate (FPR) /false alarm	$FPR = FP / N = FP / (FP + TN)$
Accuracy (ACC)	$ACC = (TP + TN) / (P + N)$
Specificity (SPC) or true negative rate (TNR)	$SPC = TN / N = TN / (FP + TN) = 1 - FPR$
Positive predictive value (PPV) or precision	$PPV = TP / (TP + FP)$
Negative predictive value (NPV)	$NPV = TN / (TN + FN)$
False discovery rate (FDR)	$FDR = FP / (FP + TP)$

5.9.2.1 Receiver operating characteristic (ROC) space

A ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. The cut-off point is the critical probability above which the model will class an event as a pre-determined tool sound. The ROC curve plots all potential cut-off points according to their respective true positive rates (TPR) (percentage of tool sounds correctly classed as the target tool) and false positive rates (FPR) (percentage of background noise or other tool sounds incorrectly classed as the target tool).

Several points in ROC space are important to note. The lower left point (0, 0) represents the strategy of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1). The point (0, 1) represents perfect classification. Informally, one point in ROC space is better than another if it is to the northwest of the first.

The best cut-off point would have an optimally high TPR and low FPR. Figure 5.26 displays the ROC curves of all constructed tool sound models. The ROC curve graphically presents the trade-off between TPR and FPR for all possible cut-off points (0

to 1 with 0.01 interval), the best of which is likely to be the point closest to the top-left corner of the graph.

An ROC curve is a two-dimensional depiction of classifier performance. The total area under the ROC curve has been used as the common method to compare classifiers, abbreviated as AUC (Bradley, 1997; Hanley and McNeil, 1982).

The larger the area under the curve, the better the model is at identifying target tool sounds in the noisy environment. Figure 5.26 clearly depicts that the hammer recognition model produced better results than any other models.

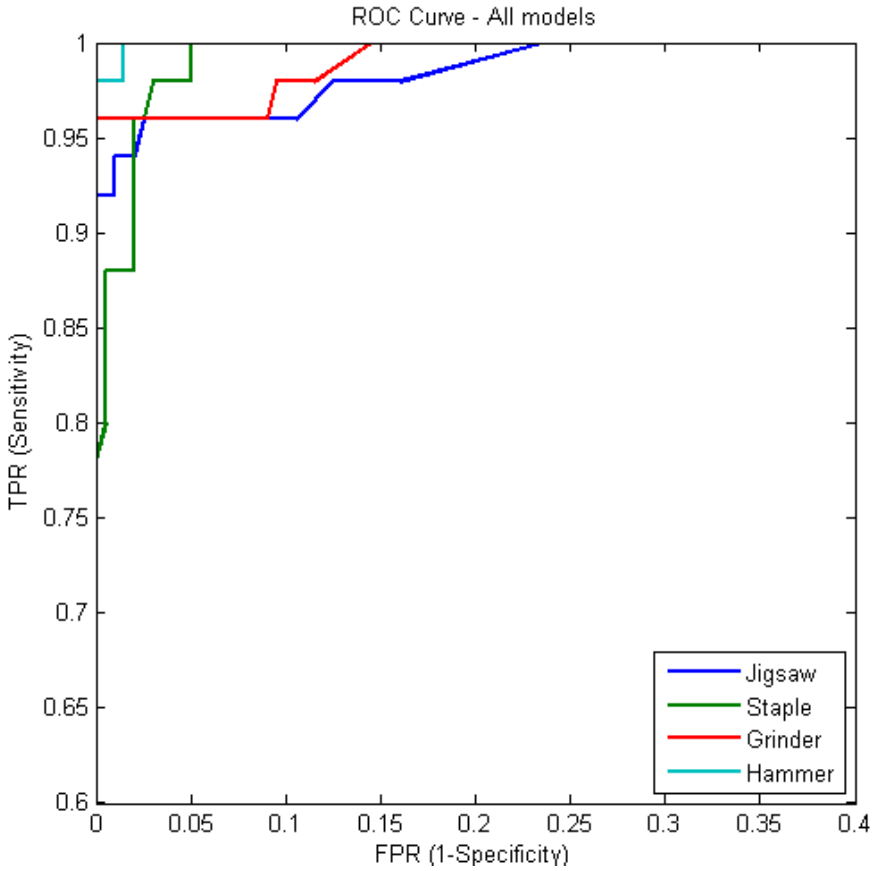


Figure 5.26: ROC curve: Audio classifier (4 models)

Analyzing the closest point to the northwest corner of the ROC curve, cut off points for all four models have been determined. If it is a range, at that point we take the mean as the cut off level. Table 5.22 and Figure 5.27 show the calculated cut off points and corresponding TPR and FPR values.

Table 5.22: True positive rate and false positive rate comparison

Model	Cut-off point	TPR	FPR
Jigsaw model	0.250	0.960	0.025
Staple model	0.190	0.980	0.030
Grinder model	0.535	0.960	0.000
Hammer model	0.115	1.000	0.015

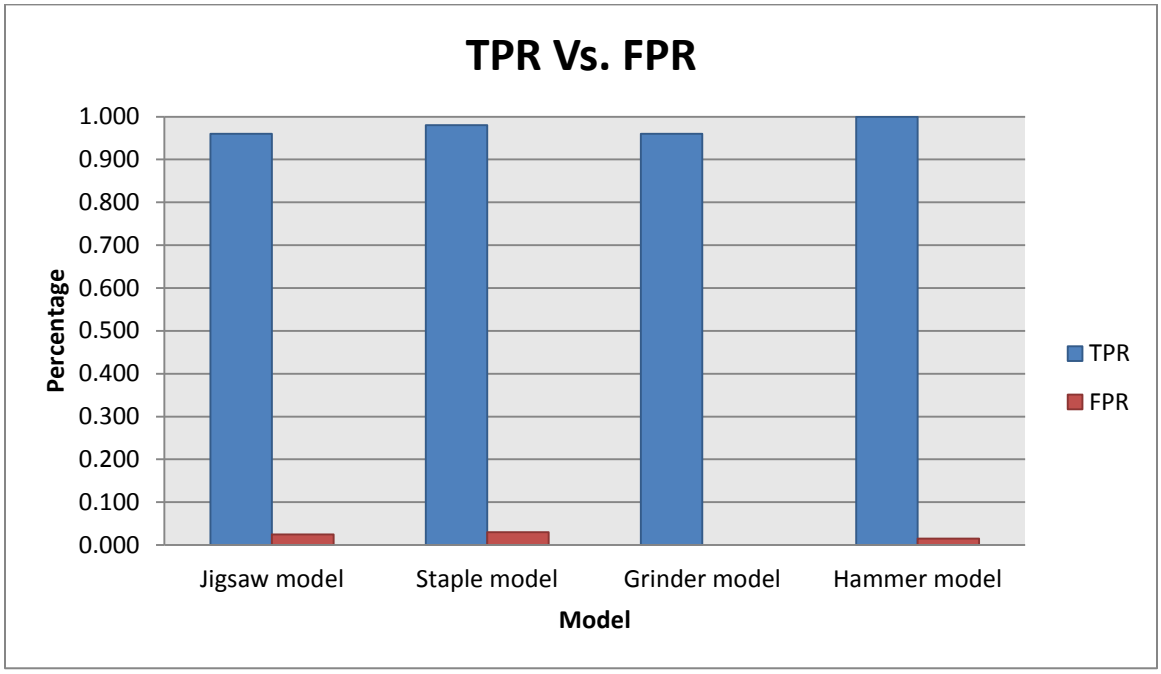


Figure 5.27: True positive rate vs. false positive rate: Audio classifier

5.9.3 Summary of model analysis

For each model construction, we began by examining the significance of each variable in a fully populated model. In this fully populated model we included maximum power values of up to five frequency band widths of each tool, four spectrum statistical values, energy, and zero crossing rate of the wave. Five factors were considered selecting the best suitable model for each tool: Nagelkerke R^2, model significance value from Hosmer and Lemeshow test, significance value (*p-value)* of variables, correlation of variables, and accuracy level of the model. In order to assess the models' goodness-of-fit, the Nagelkerke R squared measures of the respective models were calculated and are shown in Table 5.23. The model for recognizing the hammer sound is the most potent out of the four constructed models.

Table 5.23: Selection criteria parameter comparison of final models

Criterion	Jigsaw model	Staple model	Grinder model	Hammer model
Nagelkerke R Square	0.929	0.953	0.963	0.988
Significance	0.996	0.999	1.000	1.000
Max. var. significance	0.008	0.009	0.074	0.017
Correlation	0.597	0.765	0.795	0.690
Final overall accuracy	95.6	96.4	99.2	99.2

In order to assess the effectiveness of these models in classifying sounds on a construction job site correctly as per pre-determined tool sounds, and to find the appropriate cut-off points for the logistic regression models, ROC curves were used.

Initially we considered a total of 24 variables (i.e.18 FIR filters, 4 spectrum stat, E, ZCC) for model construction.

Table 5.24: Required variables and frequency components in the model

Feature Type	Variable	Var. Code	FStop1 (Hz)	FPass1 (Hz)	FPass2 (Hz)	FStop2 (Hz)
FIR Band pass filter	Hammer1	H1	161	261	321	421
	Hammer4	H4	720	820	900	1000
	Jigsaw1	J1	1470	1570	1752	1852
	Jigsaw2	J2	2890	2990	3375	3475
	Staple1	S1	3225	3325	3425	3525
	Grinder2	G2	3225	3325	3425	3525
	Jigsaw3	J3	3586	3686	3690	3790
	Grinder3	G3	3675	3775	3885	3985
	Staple2	S2	3705	3805	3970	4070
	Staple5	S5	5040	5140	5325	5425
Spectral	Centroid	C				
	Variance	VAR				
	Skew	SK				
Energy	E	E				

Finally we managed to make the classifier using only 14 variables as shown in Table 5.24. As discussed in the previous section we calculated all the accuracy parameters from the contingency table as listed below. This reveals the staple model has the lowest precision rate: 0.891. Accuracy levels of all models exceed 97%. Considering all accuracy parameters, the hammer sound classifier can be considered the best constructed mathematical model.

Table 5.25: Detailed accuracy of Audio classifier (4 models)

Derivation from contingency table	Jigsaw model	Staple model	Grinder model	Hammer model
Sensitivity or true positive rate (TPR)	0.960	0.980	0.960	1.000
False positive rate (FPR) /false alarm	0.025	0.030	0.000	0.015
Accuracy (ACC)	0.972	0.972	0.992	0.988
Specificity (SPC) or true negative rate (TNR)	0.975	0.970	1.000	0.985
Positive predictive value (PPV) or precision	0.906	0.891	1.000	0.943
Negative predictive value (NPV)	0.990	0.995	0.990	1.000
False discovery rate (FDR)	0.040	0.020	0.040	0.000

5.10 Sound source localization

The sound source localization is considered one of the most important functions of the auditory system. The sound source localization can be categorized based on the output type produced by an auditory system: azimuth, elevation, and distance to the sound source. The goal of this research is to identify the tool type of a worker on a construction job site. Finding the direction of arrival (azimuth) of a sound source can be effectively utilized to differentiate the worker tool type. Further, a direction of arrival (DOA) algorithm can be implemented using a small microphone array system.

5.11 Direction of Arrival (DOA)

5.11.1 Fundamental principles of DOA

The fundamental principle behind direction of arrival (DOA) estimation using microphone arrays is to use the phase information present in signals picked up by microphones that are spatially separated. When the microphones are spatially separated, the acoustic signals arrive at them at different times. These time delays are correlated

with the DOA of the signal and for known array geometry DOA can be estimated. There are three main categories of methods that process this information to estimate the DOA(Krishnaraj Varma, 2002).

The first category is the steered beamformer based methods. The delay and sum beamformer (DSB) is the simplest kind of beamformer that can be implemented. In a DSB, the signals are combined so that the theoretical delays computed for a particular look direction are compensated and the signals get added constructively. The minimum-variance beamformer (MVB) is an improvement over simple DSB. In an MVB, we minimize the power of the array output subject to the constraint such that the gain in the look-direction is unity.

The second category is high-resolution subspace based methods. This category of methods divides the cross-correlation matrix of the array signals into signal and noise subspaces using Eigen-value decomposition (EVD) to perform DOA estimation. These methods are also used extensively in the context of spectral estimation. Multiple signal classification (MUSIC) is an example of one such method. The algorithm again involves an exhaustive search over the set of possible source locations.

The third and final type of DOA estimation method is time delay estimation (TDE) method which consists of first computing the TDE between all pairs of microphones and then combining them, with the knowledge of the array geometry, to obtain the DOA estimate. In terms of computational requirements, the TDE based methods are the most efficient because they do not involve an exhaustive search over all

possible angles. Because of the simplicity of the algorithm and the fact that a closed form solution can be obtained, the TDE based DOA method is applied for the research study.

5.12 Microphone Array Structure and Conventions

Figure 5.28 depicts a 4-element, non-linear Kinect microphone array and a sound source in the far field of the array. We will be using the non-linear array to develop the principles of these conventional methods. The array consists of 4 microphones placed in a straight line with distance, d1, d2, d3, between adjacent microphones (see Figure 5.29). The coordinates of the microphone array elements are tabulated as L= [-113, 36, 76, 113] mm. The sound source is assumed to be in the far field of the array. This means that the distance of the source, S from the array is much greater than the distance between the microphones. Under this assumption, we can approximate the spherical wave front that emanates from the source as a plane wave front, as shown in the figure. Thus the sound waves reaching each of the microphones can be assumed to be parallel to each other. The direction perpendicular to the array is called the broadside direction or simply the look direction. All DOA's will be measured with respect to this direction.

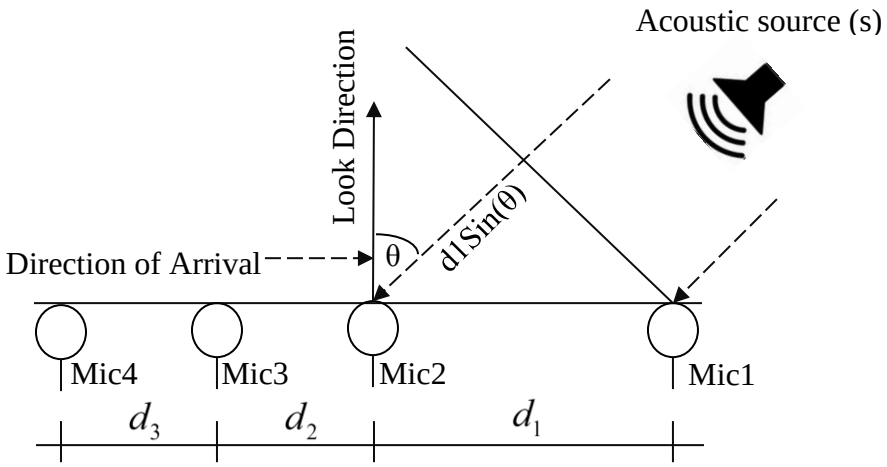


Figure 5.28: Non-linear Kinect microphone array with far field source

Angles in the clockwise direction from the broadside are considered to be positive angles and angles in the counter-clockwise direction from the broadside are taken as negative angles. The signal from the source reaches the microphones at different times. This is because each sound wave has to travel a different distance to reach different microphones. For example, the signal incident on microphone 2 has to travel an extra distance of $d_1 Sin(\theta)$ as compared to the signal incident on microphone 1. The received signal at the two microphones can be modelled by:

$$\begin{aligned} r_1(t) &= s(t) + n_1(t) \\ r_2(t) &= s(t - \tau_{i,j}) + n_2(t) \end{aligned} \qquad (26)$$

where, $r_1(t)$ and $r_2(t)$ are the outputs of two spatially separated microphones, $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ represent the additive noises, and $\tau_{i,j}$ yields the time delay

between the two received signals. The signal and noises are assumed to be uncorrelated, having zero-mean and Gaussian distribution.

$$\tau_{i,j} = \frac{d_{i,j} Sin(\theta)}{c} \qquad (27)$$

where, $c = 331.3\sqrt{1+\frac{Temp^0C}{273.15^0C}} ms^{-1}$ is the velocity of sound, $d_{i,j}$ is the distance between two microphones, and θ is the direction of arrival of the sound signal. Generally sound speed is taken as $343.2\, ms^{-1}$ in a 20C environment. Thus positive values of θ give positive delays and negative values of θ give negative delays.

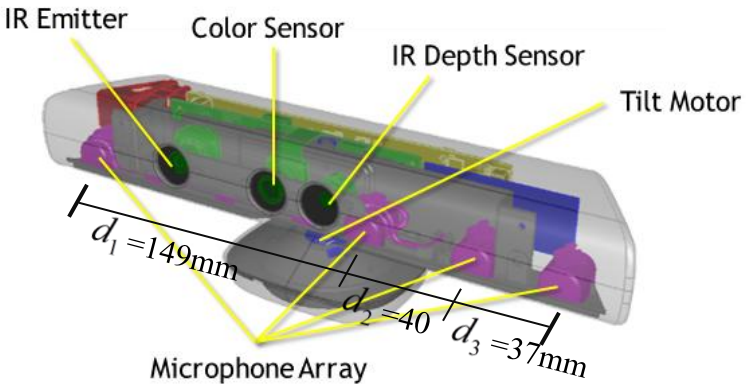


Figure 5.29: Non-linear microphone geometry

5.13 Time Delay Estimation (TDE) Method

There are many algorithms to estimate the time delay of two microphones (Dhull, Arya, & Sahu, 2010; Elkamchouchi & Mofeed, 2005; Zhang & Abdulla, 2005). The cross-

correlation (CC) method is one of the basic solutions of the TDE problem. Many other TDE methods are developed based on this algorithm. The CC method cross-correlates the microphone outputs and considers the time argument that corresponds to the maximum peak in the output as the estimated time delay. To improve the peak detection and time delay estimation, various filters or weighting functions have been suggested for use after the cross correlation (Knapp & Carter, 1976). Knapp and Carter (1976) proposed a technique called generalized cross-correlation, which is the most popular technique for TDE due to its accuracy and moderate computational complexity. The role of the filter or weighting function in the GCC method is to ensure a large sharp peak in the obtained cross-correlation, thus ensuring a high time delay resolution. There are many techniques used to select the weighting function, such as the Roth Processor, the Smoothed Coherence Transform (SCOT), the Phase Transform (PHAT), the Eckart Filter, and the Maximum Likelihood (ML) estimator (Knapp & Carter, 1976). They are based on maximizing some performance criteria.

As mentioned above, there are several TDE algorithms that have advantages in reducing computation complexity, easing hardware implementation, and increasing precision. Three of these commonly used TDE methods are adopted here: cross correlation, PHAT method, and ML method. We used all six pairs of microphones in Kinect's four channel arrays to compute TDE, followed by pair-wise time delay estimates which are usually determined in the least squares sense by solving a set of linear equations to minimize the least squared error. The maximum time delays can be expected when the acoustic sound source is ± 90 degrees from the look direction. Table 5.26

describes the maximum delays that can be expected in each pair of microphones when the sampling frequency is set to 16000Hz.

Table 5.26: Expected delays of each Kinect microphone pair

Mic (i)	Mic (j)	Dist (mm)	Max delay ms	Max sample delay	Allowable sample delay
1	2	149	434.15	6.95	7
1	3	189	550.70	8.81	9
1	4	226	658.51	10.54	11
2	3	40	116.55	1.86	2
2	4	77	224.36	3.59	4
3	4	37	107.81	1.72	2

5.13.1 Cross correlation (CC) method

One common method to estimate the time delay is to compute the cross correlation function between the received signals at two microphones. Then locate the maximum peak in the output, which represents the estimated time delay. The CC can be modelled by:

$$R_{i,j}(\tau) = E[x_i(n)x_j(n-\tau)] \qquad (28)$$

where, E denotes the expectation value $E[f(n)]$ of $f(n)$ and for an observation window of N samples of $f(n)$. An estimate of the expected value can be written as:

$$E[f(n)] = \frac{1}{N}\sum_{i=1}^{N} f(i) \qquad (29)$$

5.13.2 Phase transform (PHAT) method

The phase transform (PHAT) is the most commonly used pre-filter for the GCC. In order to improve the accuracy of the delay estimate, Knapp and Carter (1976) proposed pre-filtering the signals prior to the integration. The estimated time delay for a pair of microphones is assumed to be the delay that maximizes the GCC-PHAT function for that pair. The PHAT is a GCC procedure that has received considerable attention due to its ability to avoid causing spreading of the peak of the correlation function. This can be mathematically expressed by:

$$R_{r_1 r_2}^{PHAT}(\tau) = \int_{-\infty}^{\infty} \psi_p(f) G_{r_1 r_2}(f) e^{j2\pi f \tau} df \qquad (30)$$

where, $\psi_p(f) = \dfrac{1}{|G_{r_1 r_2}(f)|}$, $G_{r_1 r_2}(f)$ is the cross-power spectrum of the received signal and $\psi_p(f)$ is the PHAT weighting function. According to the above expression, only the phase information is preserved after the cross-spectrum is divided by its magnitude. This processor approaches a delta function centered at the correct delay.

5.13.3 Maximum likelihood (ML) method

The ML is another important method within the GCC family since it gives the maximum likelihood solution for the TDE problem. The ML weighting function $\psi_{ML}(f)$ is chosen to improve the accuracy of the estimated delay by attenuating the signals fed into the correlator in the spectral region where the SNR is the lowest. The ML method can be represented by:

$$R_{r_1r_2}^{ML}(\tau) = \int_{-\infty}^{\infty} \psi_{ML}(f) G_{r_1r_2}(f) e^{j2\pi fr} df$$
$$\psi_{ML}(f) = \frac{1}{|G_{r_1r_2(f)}|} \times \frac{|\gamma_{r_1r_2}(f)|^2}{1-|\gamma_{r_1r_2}(f)|^2} \quad (31)$$

where, $|\gamma_{r_1r_2}(f)|^2 = \frac{|G_{r_1r_2}(f)|^2}{G_{r_1r_1}(f) G_{r_2r_2}(f)}$ is the magnitude coherency squared, and $\psi_{ML}(f)$ is the ML weighting function. The ML processor weights the cross-spectral phase according to the estimated cross-spectral phase when the variance of the estimated phase error is the least.

5.13.4 Error analysis

Generally, room reverberation is considered the main problem for TDE. Moreover, acoustic background noise may further decrease the performance of time-delay estimators. The performance of TDE is always affected by the reverberation in a room (Bedard, Champagne, & Stephenne, 1994; Jingdong, Yiteng, & Benesty, 2005; Ming, Kot, & Er, 1998). The problem becomes more challenging once room reverberations rise. In a highly reverberant room, all the known TDE methods become unreliable and may even fail. In particular, the quantitative behaviour of the estimator variance for reverberation can be explained naturally in terms of an equivalent signal-to-noise ratio (SNR), which treats the reverberant energy at the microphone output as undesirable noise. Research findings from Bedard et al. (1994) has proven that the high level of reverberation causes the low value of SNR. Figure 5.30 illustrates the SNR values for the collected sound samples of four construction tools. The highest SNR value is recorded

from the most powerful tool we used in the test: the angle grinder. Meanwhile, hammer sound has a SNR range from 7-12, which is the lowest among this tool set.

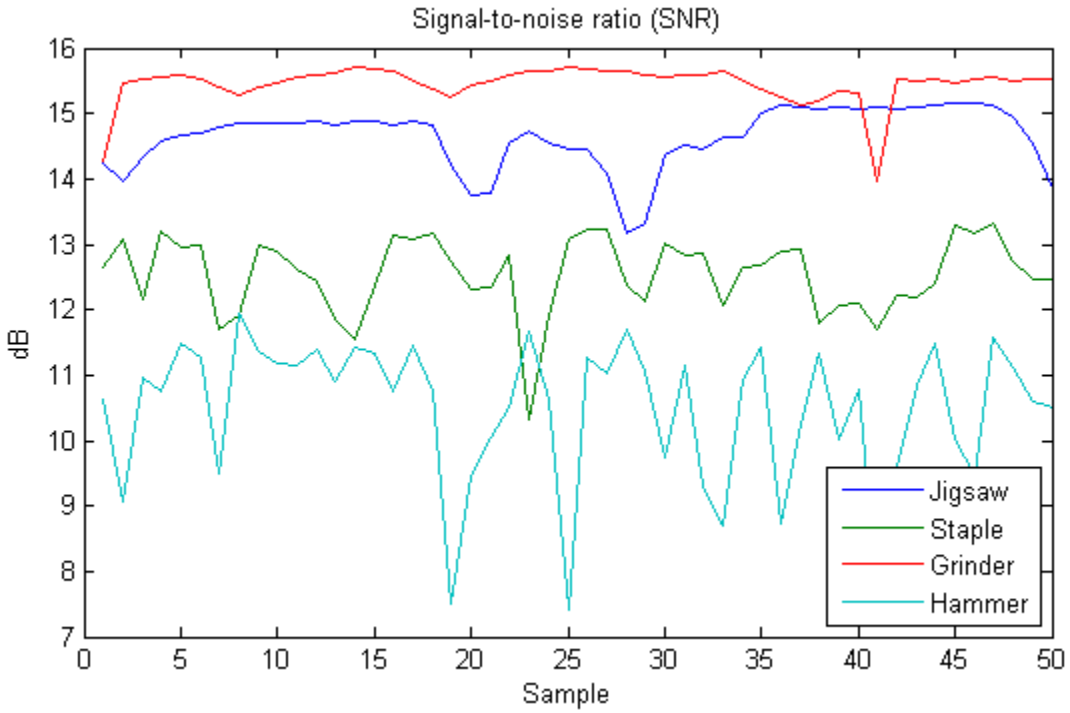


Figure 5.30: SNR values for collected sound samples of construction tools

Another problem in TDE is the observation interval. Finite time measurement causes the estimated cross-power spectrum variance, which may affect the accuracy of TDE (Ianniello, 1982). It can lead to a large error when the actual observation interval is very short. In many cases of practical interest, however, the assumption of a long observation interval is inconsistent with other prevailing conditions, such as that assumption of stationary processes and constant delay will only be satisfied over a limited time interval. However, there is actually a trade-off between observation time and SNR in the TDE problem. In the case where SNR is low, long observation time is required to ensure

accuracy. Likewise the higher the SNR, the shorter observation interval is needed. Therefore, the two main problems in TDE can be combined into a situation where the SNR is low. There is degradation of performance at low SNR. At high SNR, this is the ambiguity-free mode of operation where differential delay estimation is subjected only to local errors. For very low SNR values, observations are dominated by noise and are essentially unhelpful for TDE.

5.14 DOA Model Construction

The direction of arrival (DOA) estimation method consists of first computing the time delay estimates (TDE) between all six pairs of microphones and then combining them, with the array geometry, to obtain the DOA estimate. The selected TDE methods are modeled in the MATLAB platform. The validation is carried out in actual noisy environments. Comparison of accuracy level of the DOA from all three TDE methods is further reviewed in the Chapter Six: Integrated application and model validation.

For performance comparison of three TDE methods, a simulation was carried out in practical environment. An actual noise was recorded from the real environment and adopted as an additive noise to a jigsaw sound sample. Two noisy jigsaw signals were created with a 10 frame delay and performance of TDE (CC, GCC-PHAT and ML) is illustrated in the Figure 5.31. It is evident that the peak position corresponds to the actual time delay. The x-coordinate denotes the time-lag, and the y-coordinate denotes the resulting cross-correlations.

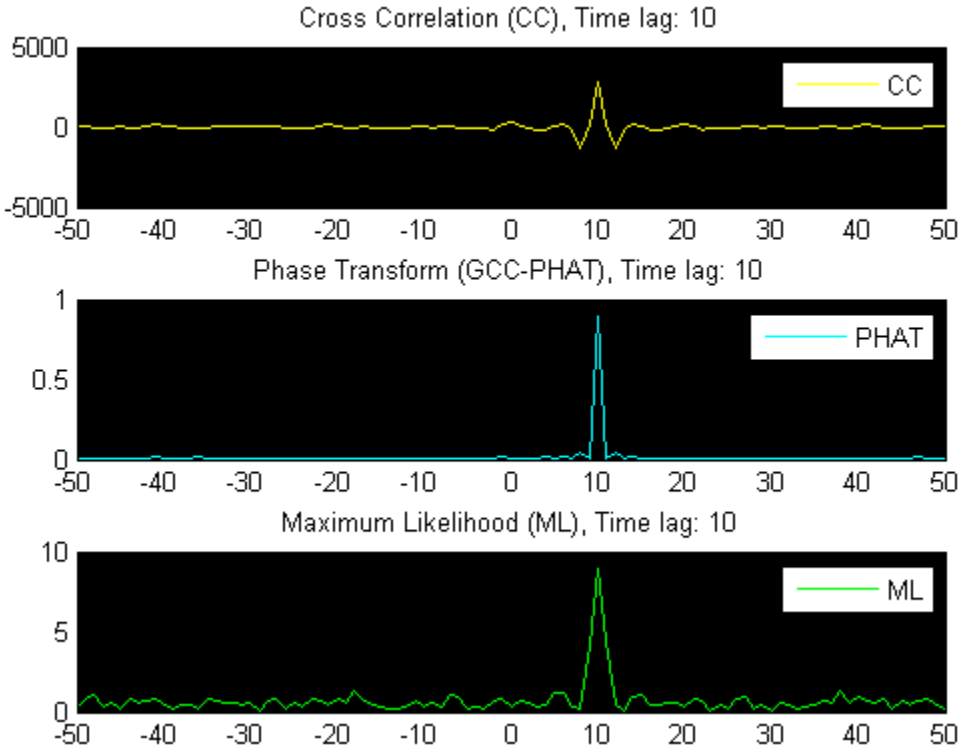


Figure 5.31: Cross correlation results using: CC, PHAT, ML

TDE for all 6 pairs have been measured from all three methods. Errorless TDE have been filtered using the maximum allowable sample delays for each pair as Table 5.26. Then DOA is determined using least squares estimation with remaining correct delay values using the following mathematical model

$$Time\_lag(samples)_{i,j} = \left[\frac{d_{i,j} \times Fs}{c} \right] Sin\theta \qquad (32)$$

where, Fs is the sampling frequency (16000Hz), c is the speed of sound (343.2m/s), $d_{i,j}$ is the distance between microphone pair and θ is the DOA angle.

The Gauss-Markov Model (GMM) is used to solve the DOA. The general form of the GMM is illustrated in Equation 14. A higher number of available pairs of microphones increase the redundancy and key to optimized levels of accuracy and robustness in the DOA process. Since θ is the only unknown parameter of the model, redundancy is 5 for all 6 pairs of microphones. The least squares equation set is used to determine sin θ and the error value vector.

5.15 Factors Affecting DOA Model Accuracy

Various factors affect the accuracy of the DOA estimates obtained using the TDE based algorithm. Accuracy of the hardware used to capture the array signals, sampling frequency, number of microphones used, and reverberation and noise present in the signals are some of these factors. The hardware that is used should introduce minimum phase errors between signals in different channels. This is a requirement no matter what method is used for DOA estimation. Also, the more microphones we use in the array, the better the resulting estimates. The sampling frequency becomes an important factor for TDE based methods especially when the array is small in terms of distance between the microphones. This is because small distances mean smaller time delays and this requires higher sampling frequencies to increase the resolution of the delay estimates.

5.16 DOA Model Validation

The developed DOA model is validated in the structural lab at the University of Calgary, which is considered an actual noisy environment similar to an indoor construction job

site. We recorded a video and audio data file while operating all four tools in this work space. Then we analysed the DOA and all these results are comprehensively reviewed in Chapter Six: Integrated application and model validation.

Chapter Six: **Integrated application and model validation**

6.1 Introduction

This chapter explains integration of the MATLAB application with different subcomponents designed for various purposes, and describes different analysis methods used for verification and validation of the proposed research concepts. Model validation consists of 4 major areas: hardhat detection, tool sound detection, acoustic sound direction (i.e. DOA), and tool-time and performance measurements.

6.2 Application Development

This application is developed using graphical user interface (GUI) controls in MATLAB GUIDE and Simulink models to provide a user friendly interface while allowing for structured programming underneath it. Different user interface windows will be described in the following sections.

6.2.1 Main window

The main window is the central unit, which combines all 6 sub component applications. This unit allows the user to define colour ranges as the local hardhat colour code system, calibrate the camera in terms of determining exterior orientation parameters, view 3D point clouds of the job site, define geo zones of the job site, execute a worker tracking module, and view tool time and performance information. Each task has a separate button as shown in Figure 6.1, and detailed functions of each button will be discussed in the following section.

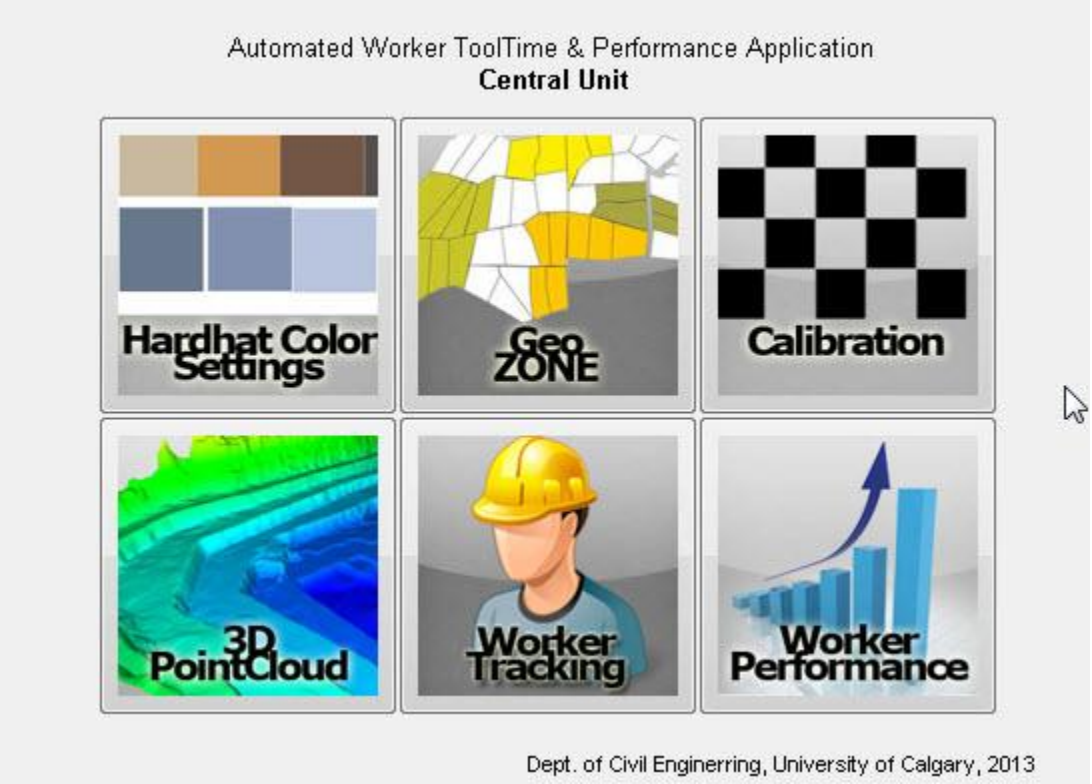


Figure 6.1: GUI of Main application

The following sequential steps are recommended prior to the execution of the construction worker tracking component for an ideal process flow.

1. Camera calibration and SPR: The purpose of camera calibration is to determine numerical estimates of the interior orientation parameters (IOP) and image coordinate corrections that compensate for various deviations from the assumed perspective geometry of the implemented Kinect camera. Then a single-photo resection (SPR) process is followed to determine the exterior orientation parameters.

2. Hardhat colour settings: The default range for pre-identified, colour-coded hardhat is set in the developed application. However, it further allows users to change the range for different colour specification requirements.
3. Geo zone declaration: Define construction job site block arrangement.

6.2.2 Hardhat colour settings

It is recommended that this process be done prior to the execution of the worker tracking process so that the colour segmentation and filtering process will be more precise to the target. YCbCr colour space is selected for the colour segmentation in the research. Default values for the pre-defined colours (i.e. red, blue, yellow, and white) are set in the application start-up as mentioned in Chapter Four:. Further, as shown in Figure 6.2, the system allows the user to customize colour ranges by using sliders as required, and the estimated colour range is visualized in the colour sample panel.

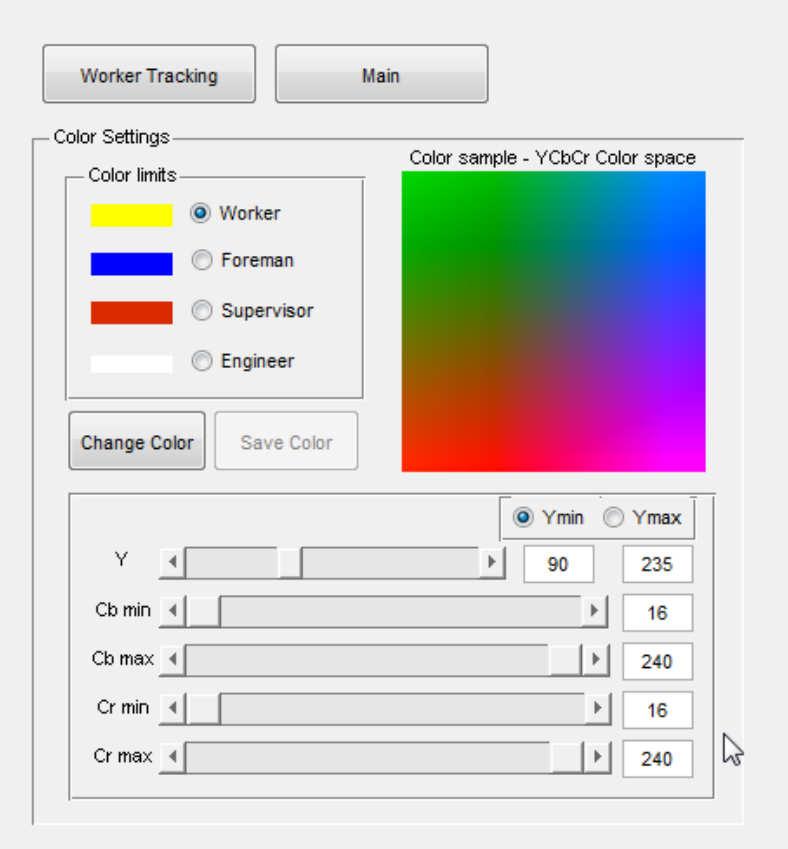


Figure 6.2: GUI of Hardhat colour settings form

6.2.3 Point cloud

Figure 6.3 shows the application for viewing the 3D point cloud of the indoor job site environment with supporting 360 degree panoramic sight (i.e. consisting of pan, zoom, and orbit tools). This point cloud has been created using 3D depth data and RGB data and consists of 307200 data points. This point cloud has been further proposed for future research, to measure the site progress by analyzing a 4D CAD model and the real as-built model.

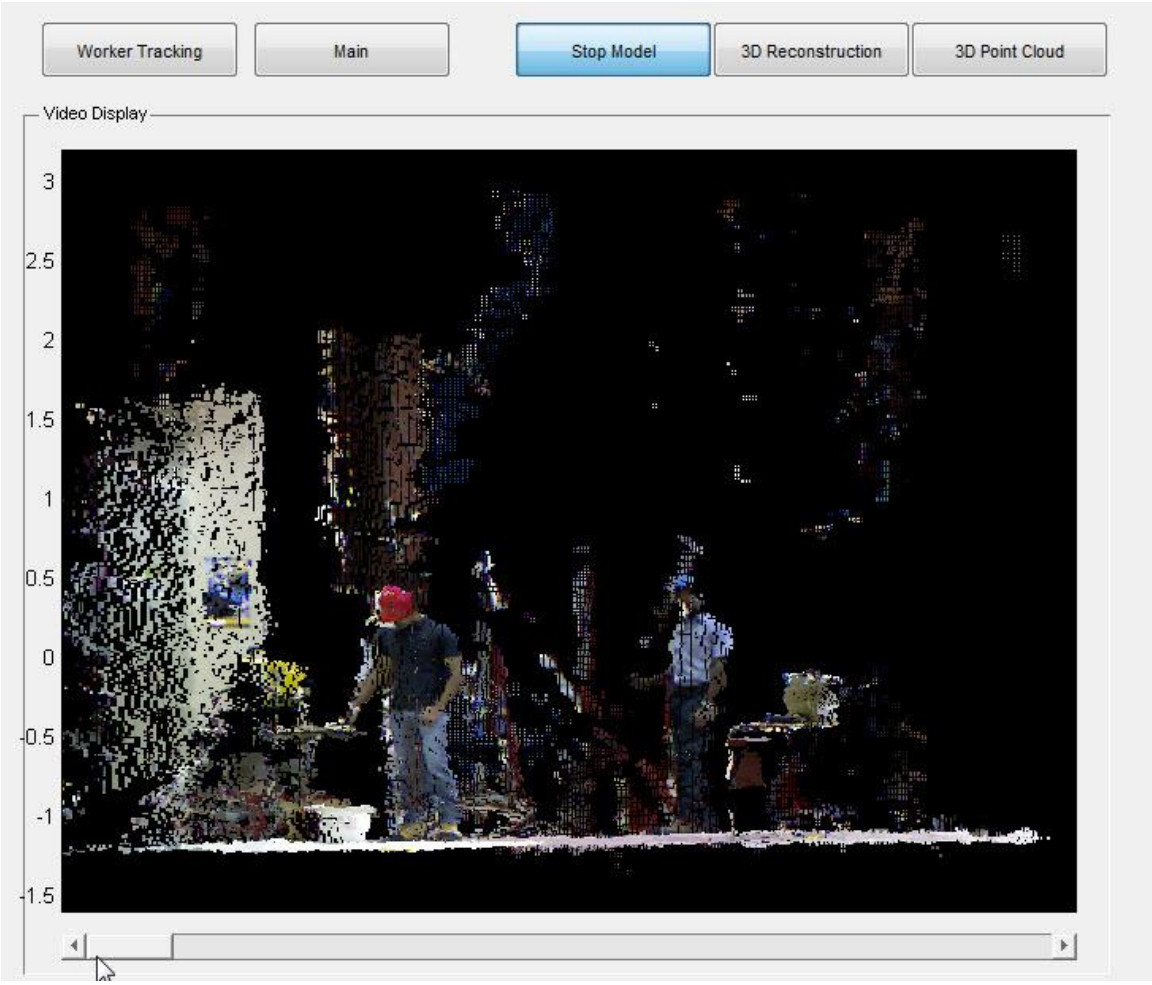


Figure 6.3: GUI of 3D point cloud view form

6.2.4 Geo zone settings

Figure 6.4: GUI of Geo zone settings, illustrates the application for entering, reviewing, and modifying the geo zones of the construction site. In order to organize the construction flow of a job site, labour allocation, and scheduling, the work area is structured both into a higher level and a micro level. In this approach, the work phase is divided into several blocks such as material storage, tools and equipment storage, workplace area 1, 2, etc. We adopted this general practice into our proposed system in

terms of finding non-tool time actions, assuming workers engage in a location related activity such as material handling or tool handling. A user interactive tool is developed for defining, reviewing, and editing geo zone blocks with the support of 3D point cloud as illustrated in Figure 6.4. The coloured boxes in Figure 6.4 indicate geo zones defined by the user. The end user shall define a related construction activity of the job site when defining a block. For instance, a worker moves into a material storage area and the duration spent at that block is taken as the material handling time over the period. This period directly contributes to the non-tool time activity group of a worker. Further, time spent in other geo zones can also be considered as engaging in declared, related, non-tool time activities (i.e. tool handling).

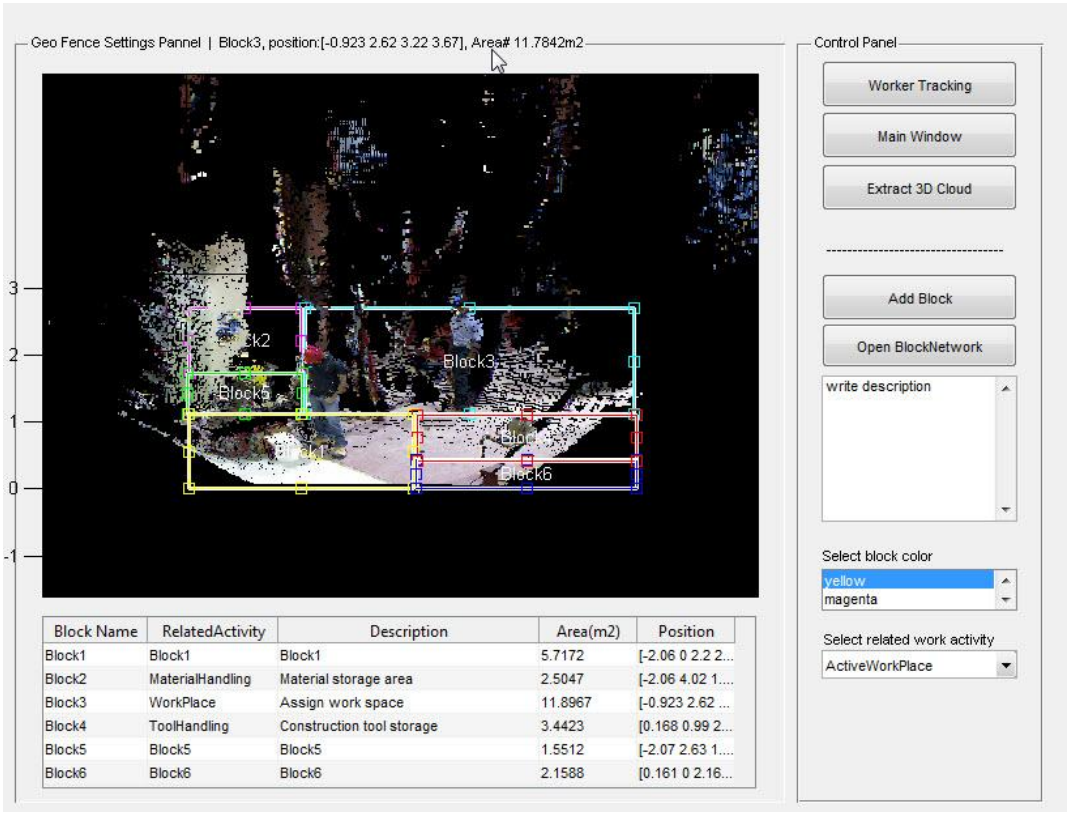


Figure 6.4: GUI of Geo zone settings form

6.2.5 Camera Calibration and SPR

This application is developed to measure the exterior orientation parameters using the single photo resection (SPR) method described in Chapter Four:. This application becomes a platform to connect with the Simulink model and enables video extraction to provide snap shots of the scene that consist of several ground control points. Semi-automated target extraction is introduced to the system, which tracks checkerboard grid lines, circular shapes, and manually selected arbitrary target points (see Figure 6.5). Analyzed results are tabulated and displayed in the application and archived for future use in the integrated worker tracking system. A detailed description of the camera calibration procedure can be reviewed in Chapter Four: and results of the SPR procedure are attached in Appendix A.

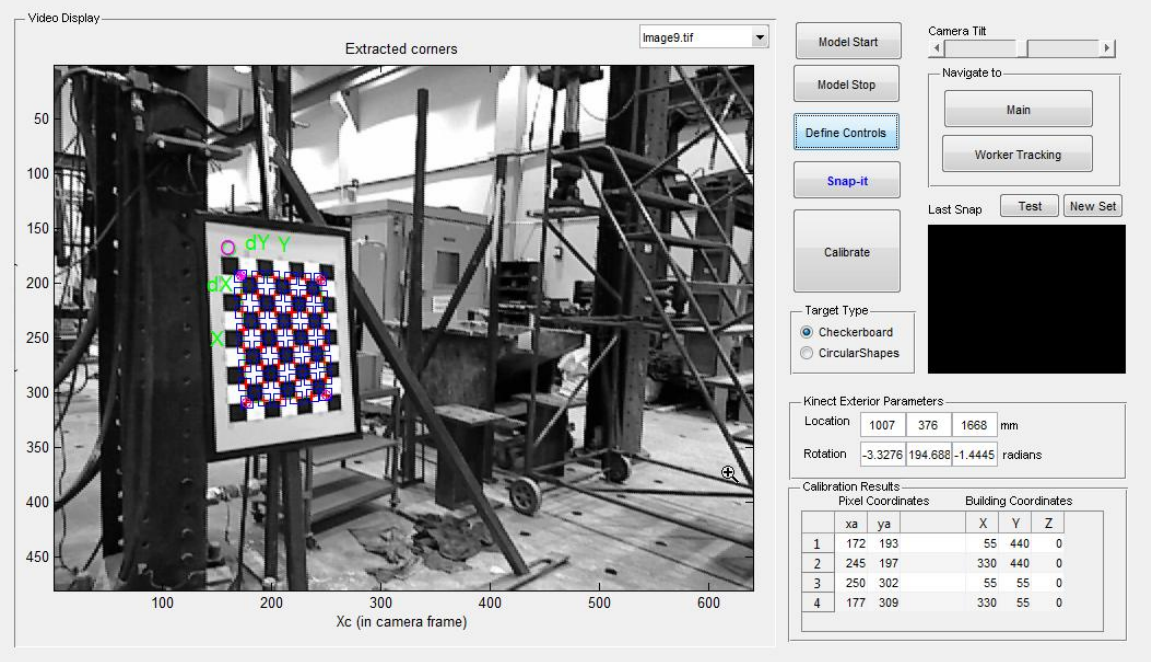


Figure 6.5: GUI of Single Photo Resection (SPR) form

6.2.6 Worker tracking

This is the key sub component in the application, which simultaneously tracks multiple workers, detects construction activities, and categorizes worker movements over the monitoring period. The graphical user interface (GUI) maximizes information provided on screen for each time frame of extracted data (see Figure 6.7). This includes raw data (i.e. RGB, signal time series), intermediate processed data (i.e. spectral distribution, confidence levels of hardhat and construction tools) and fully processed data (i.e. worker category, tool type, DOA) in visual and numerical form. Further, DOA of the tool sound is visually represented by coloured lines as an overlay on the RGB image to better understand the preciseness of the system. Figure 6.6 shows two different instances of worker tracking system. More detailed snapshots of GUI for different instances are given in Appendix B.

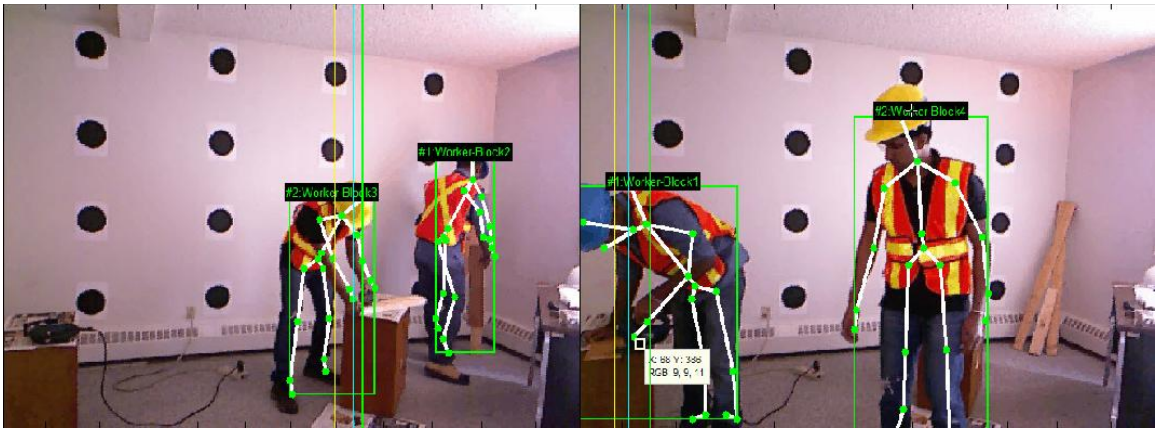


Figure 6.6: Snapshots of worker tracking instances

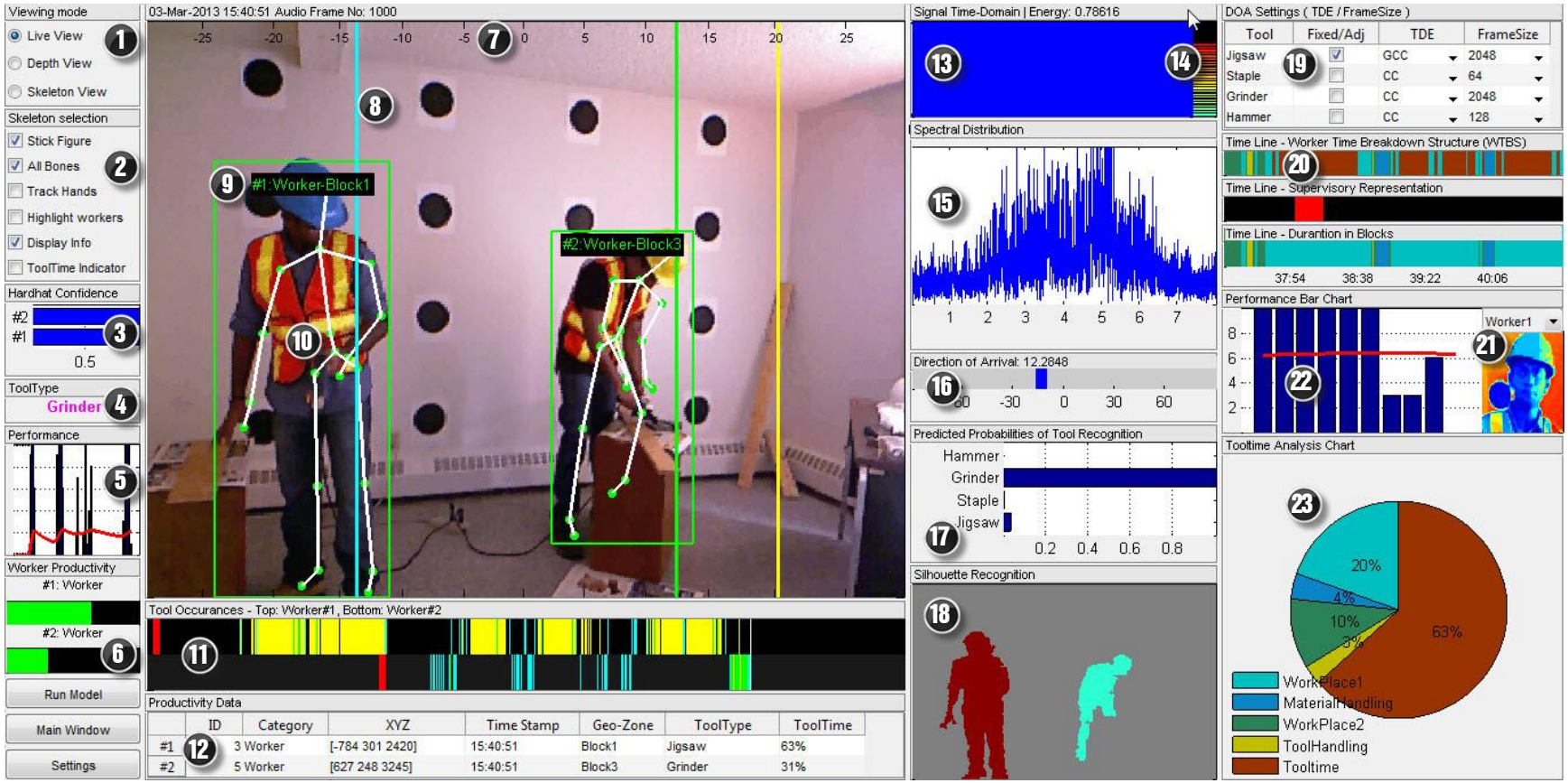


Figure 6.7: GUI of worker tracking form

Apart from the current frame information, the system further provides historical information of tool usage by displaying the performance rate and tool occurrences in past frames. This enables the user to identify tool time information about the worker.

In the upper left corner of the application, the viewing panel consists of three viewing modes: live RGB, depth, and skeleton image, which enables the user to view extensive information of the scene or separated target figures of the scene.

In summary, the worker tracking application extracts, analyzes, and presents extensive information of worker movements, tool time, and performance while providing historical and current frame information to the end user. In the end, productivity related information is tabulated and recorded for detailed analysis of worker tool time and worker behavioral patterns.

The features and components of the developed worker tracking application are numbered on the Figure 6.7 and briefly describe as follows:

1. Viewing mode selection panel: supports RGB, Depth and skeleton view
2. Skeleton display settings panel:
    a. Display supports stick figure, 20 skeleton joints, worker information
    b. Highlight worker: display background subtraction (Left image Figure 6.8)
    c. Tool-time indicator: Indicate the tool-time percentage by filling the silhouette figure of each tracked person (Right image Figure 6.8)
3. Confidence level of hardhat recognition of two people (0-1 scale)
4. Recognized tool type in the current audio frame (frame size 0.25s)
5. Performance of the worker

6. Tool-time percentage of two workers are indicated in the green bar chart
7. Angle measurement in degrees for DOA
8. DOA for tool sound is indicated from three different colors:
    a. Yellow: Jigsaw tool sound
    b. Cyan: Staple gun sound
    c. Green: Angel grinder sound
9. Silhouette bounding box with worker information: worker ID, category and geo-zone
10. Skeleton figure
11. Tool sound occurrence is indicated for two workers in this two line charts. Red line indicates the initial recognized time, and other coloured lines indicate the tool sound occurrences as: yellow – jigsaw, cyan – staple gun, green – angle grinder.
12. Data summary of current tracked workers
13. Time domain of the audio signal
14. Energy of the audio signal is indicated from a color bar
15. Spectral distribution of the audio signal
16. DOA of the current audio signal is indicated in the graph
17. Predicted probabilities in a 0 to 1 scale for the tool recognition are indicated in this figure. If the maximum probability is higher than the cut off level of the tool, then that tool is considered as the current operating tool in the jobsite.
18. Silhouette map of the recognized people on site

19. DOA settings: this table allows user to define or change settings of the TDE and audio frame size in order to change the DOA results.
20. Time line of the selected worker
    a. Worker time breakdown structure: the time line with activities (i.e. tool time, material handling time, supervisory instructions time, geo-zone related activities) of the selected worker (see 21) are indicated in this bar chart.
    b. Supervisory representation: Supervisor's representation is indicated by red color bar chart.
    c. Duration in blocks: this bar chart indicates the duration spent in each block of the site.
21. Worker selection drop down menu and photo of the selected worker
22. Performance of the selected worker
23. Tool time distribution chart of the selected worker

Figure 6.8: Background subtraction and tool-time indicator

6.2.7 Activity based worker tool-time and performance

Figure 6.9 depicts the GUI of the tool time and performance analysis application. The major function of this application is determining the tool time information for the entire workforce in the company by using a series of recognized construction activities, duration spent in geo-zones, and supervisory representation analysis. The time series of movements in geo zones, tool occurrences, and supervisory representation should be aligned as shown in four parallel graphs, and the final activity analysis graph is developed based on time segments.

The video playback feature provides an opportunity to review the archived video for a given time duration if there was any further information required. The application allows the end user to view tool time and performance results of all workers, and using manual face recognition, duplicates may be merged. Another important aspect of the application is the work sampling measurement comparison. This allows the user to compare different work sampling measurement with their worker profiles and find the best among the crew.

Supervisory effect on the worker performance can be obtained by analysing the supervisory representation bar chart. This provides whether the performance is increased due to a supervisor onsite. This information combined with worker movement results generates an analysis of worker behavioral patterns that assist labour allocation and supervision. Tool time information of individual workers can be effectively used for project scheduling with the reference of typical performance values. Further workers having higher tool time can be used for critical activities in the project life cycle.

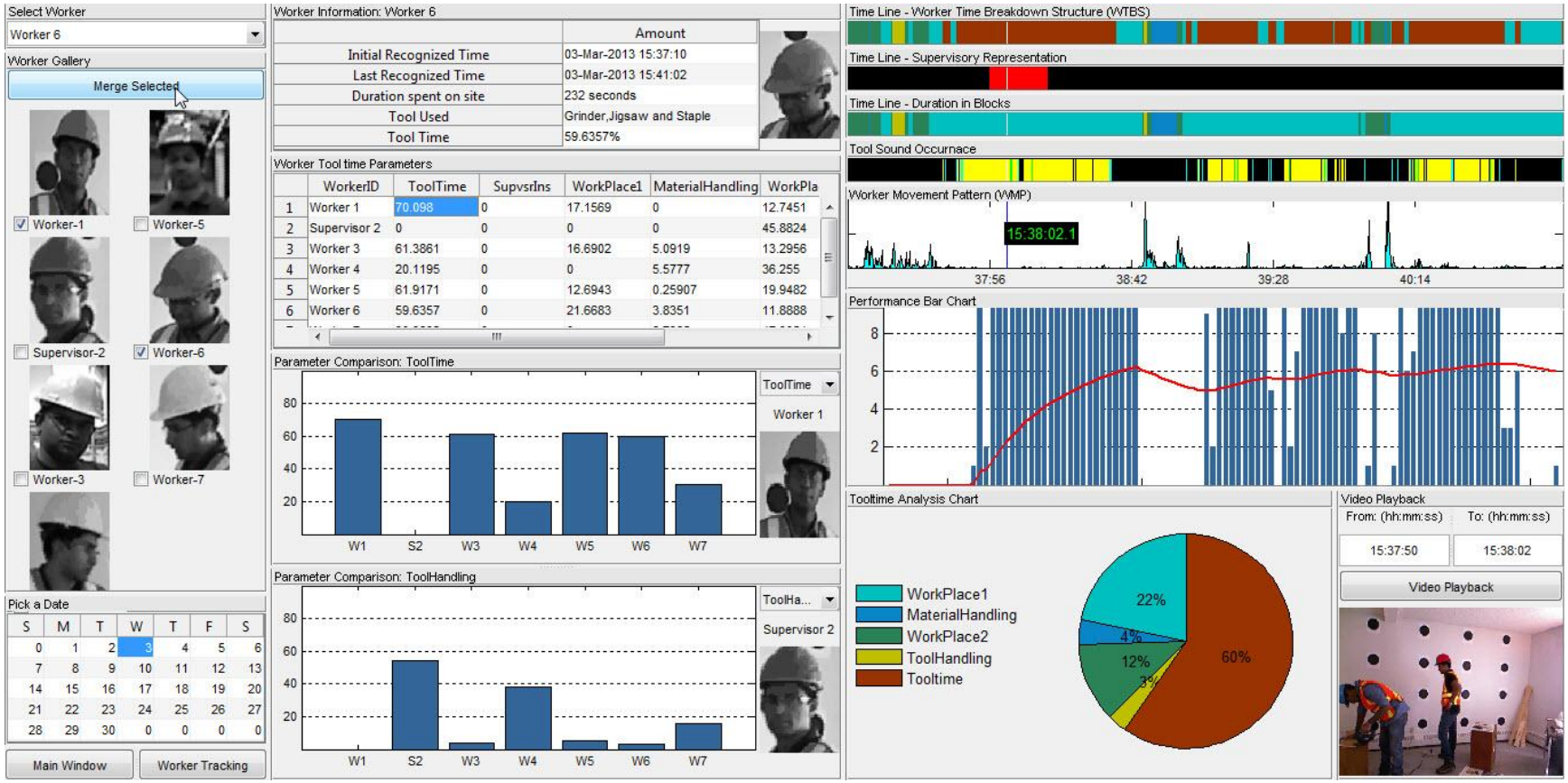


Figure 6.9: GUI of Tool-time and performance analysis form

6.3 Summary of Application Development

A comprehensive, structured, user-friendly application package (bundle of six different sub components) was developed to address the automation flow of worker tool time and performance evaluation. A graphical representation of results (i.e. bar charts, pie charts, and line charts) and user interactive tools attract the end user and rapidly communicate the essence of the output. In brief, this application minimizes the human involvement in tool time monitoring and completely eliminates human error and data limitation issues from the conventional method.

6.4 Field Testing

As the primary concept of the research was conceived through the experience of real construction operations, the proposed research concepts and the developments needed to be tested at the site to verify and validate whether the proposed system really addresses the identified issues or not. The testing process in ideal situations needs to be implemented by the industry itself without having any intervention from the researcher or any external parties. But realistically, it was not possible to get construction companies to implement the idea by themselves due to the initial commitments (financial and non-financial) required from the companies.

6.5 Site Description and Model Validation Process

The testing and validation of the proposed concepts was conducted in an indoor laboratory environment and in the civil engineering structural laboratory at the University

of Calgary. The following sections describe the validation procedure, and the validated results of recorded, extensive data samples for two statistical models (i.e. hardhat recognition model and tool sound recognition models), acoustic sound direction of arrival (DOA), and total integration system.

6.6 Hardhat Recognition Model

The hardhat recognition model was validated with a recorded video against the manual observation. The video was recorded with a person wearing a coloured hardhat at a specific time (i.e. blue, red, white, and yellow), then changing to the different coloured hardhats over the recorded period. Movements of the collected data set cover the entire visual spectrum and full scale of physical depth range (0.8m-4.0m).

As described in Chapter Four:, the hardhat recognition model is constructed using binomial logistic regression. The hardhat probability formula is illustrated as:

$$p(Hardhat) = \frac{1}{1+e^{-z}} \qquad (33)$$

where, $z = 0 + 5.549(ECC) - 0.338(DHC) + 1.102(AREADFG2)$, ECC is eccentricity of the blob, DHC is pixel distance between head and blob centroid, and AREADFG2 is blob area * (distance between Kinect and human figure) ^{2}. The following figure illustrates sample image frames of the recorded video file taken for the hardhat classifier validation.

Figure 6.10: Sample image frames of hardhat classifier validation

Figure 6.11 graphically demonstrates a comparison of hardhat classifier output against the observed data for 900 image frames (continuous 30seconds with 30fps). In Figure 6.11, the first row demonstrates the observed successive image frames of the person wearing a hardhat. The second row shows the predicted successive image frames.

It is apparent that there has been a precise colour code distinguished in most of the frames. The bottom image shows the predicted probabilities for the hardhat detection calculated from the logistic regression. The cut off level was set to 0.43 as we described in the previous section in model construction. The blue hardhat was precisely tracked with 100% probability over the time period. The classification Table 6.1 numerically demonstrates results to a further extent. The highest TPR has been shown for the yellow hardhat. White shows the highest TNR while having the worst TPR. To sum up, all four coloured hardhats have displayed more than 94% of the accuracy of detection and less than 4% of FPR.

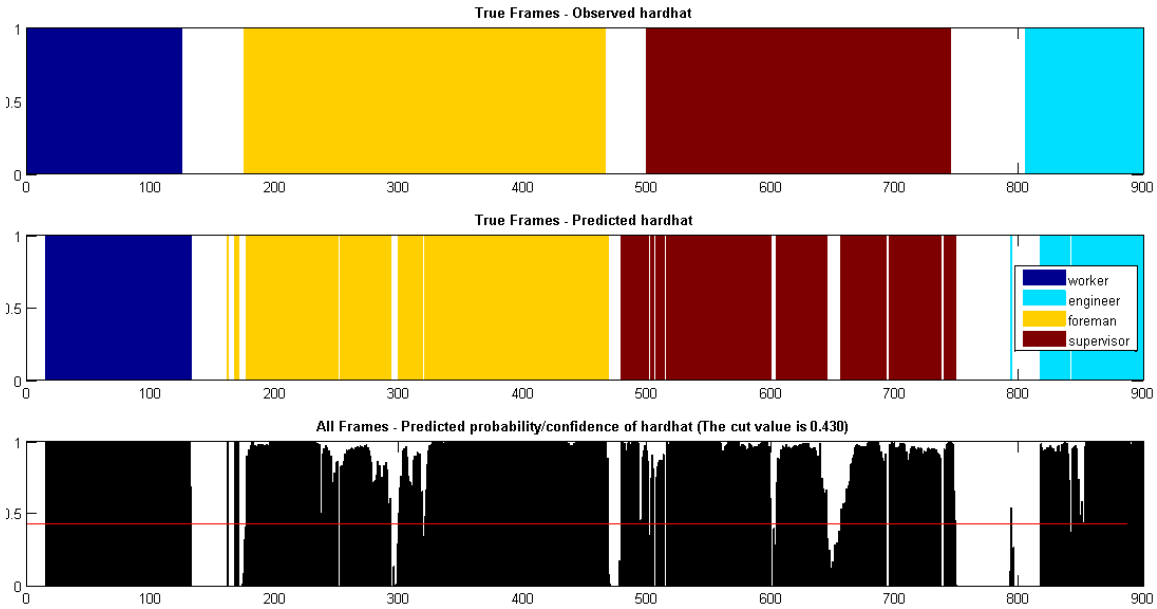


Figure 6.11: Visualization of observed and predicted hardhat

Table 6.1: Classification table: Hardhat classifier

		Blue			**Yellow**		
Observed		**Predicted**			**Predicted**		
		Observed Hardhat		Percentage Correct	Observed Hardhat		Percentage Correct
		0	1		0	1	
Observed Hardhat	0	766	7	99.1	601	7	98.8
	1	15	111	88.1	12	279	95.9
Overall Percentage				97.6			97.9
		Red			**White**		
Observed		**Predicted**			**Predicted**		
		Observed Hardhat		Percentage Correct	Observed Hardhat		Percentage Correct
		0	1		0	1	
Observed Hardhat	0	630	24	96.3	803	1	99.9
	1	24	221	90.2	14	81	85.3
Overall Percentage				94.7			98.3

6.6.1 Classifier failures

Colour segmentation, skeleton joint and image recognition, and coefficients of the logistic regression are the major attributes that really drive the results for the hardhat classifier. If one of these fails, the results cannot be precise.

The graph proves the accuracy of the implemented colour segmentation by detecting the correct colour order. A few predicted time frames are slightly varied for numerous reasons. The initial time gap occurred in the prediction graph because of the delay in initial skeleton recognition. Extra positive predictions are witnessed on either side of the observed period because wearing and removing a hardhat is also detected as potential being a hardhat (1, 2 from left in Figure 6.12).

Further, the classifier obviously fails when the hardhat is physically covered by objects. This usually occurs right after wearing the hardhat (3 from left in Figure 6.12). The right hand side picture in Figure 6.12 shows another rare occasion of failure. The skeleton image has been divided and as a result the skeleton joints are erroneously positioned. Consequently, the distance between head and blob centroid (a significant parameter in logistic function) increases, and results in reducing the probability of being a hardhat.

Figure 6.12: Sample image frames of hardhat classifier failures

6.7 Tool Sound Recognition Model

This construction activity recognition model consisted of 4 tool sound models: jigsaw, staple gun, angle grinder, and hammer. All four models were validated with recorded audio files (i.e. wav format) against the manual observation. The audio files were recorded in 16000 Hz, the frames are of 4096 samples (256ms) each, with 12.5% (512 samples or 32ms) overlap in each two adjacent frames.

Four data sets were recorded for each tool sound and results for the combined data set (783 frames) have been presented.

At the same time, video footage was also recorded for the validation of direction of arrival of sound source. Acoustic sound source location of the collected data set covers 57 degree angle (i.e. the entire visual spectrum) and depth ranges from 0.8m-4.0m (i.e. full scale of physical depth range of Kinect skeletal vision). All models were constructed based on binomial logistic regression as in the following log odds ratio calculation.

$$p(ToolSound) = \frac{1}{1+e^{-z}} \qquad (34)$$

where, $z = b_0 + b_1(valriable_1) + b_2(valriable_2) + \cdots + b_n(valriable_n)$, b_0 is the constant and b_1 to b_n are the corresponding parameter coefficients.

Frame-to-frame analysis produces confidence values for each tool, and the system detects the correct tool based on the cut off values. If two possible tools have been detected, the tool with the maximum confidence is selected as the final tool for the frame. Further, discrete sound events such as the stapler and the hammer are possibly tracked in two adjacent frames because of the sound content in each frame. These extra repetitive

detections were removed by applying a peak detector algorithm to its detected energy distribution.

6.7.1 Jigsaw

A data set of 179 frames long, including 144 observed jigsaw sounds, was recorded for the jigsaw tool.

The z for the Mastercraft jigsaw tool sound probability formula is:

$$z = 0 + 12.5987(Jigsaw1) + 10.5082(Jigsaw2) - 18.8865(CJ) + 1.3478(E)$$

Table 6.2: Classification table: Jigsaw

Observed		Predicted		
		Observed Jigsaw		Percentage Correct
		.00	1.00	
Observed Jigsaw	.00	639	0	100.0
	1.00	3	141	97.9
Overall Percentage				99.6

6.7.2 Staple gun

A data set of 175 frames long, including 59 observed staple sounds, was recorded for the staple gun tool.

The z for the Mastercraft staple gun sound probability formula is:

$$z = 0 + 7.1649(Staple1) + 13.3360(Staple2) - 10.9256(Staple4) + 11.3485(Staple5) - 2.2514(E) - 4.0123(Variance)$$

Table 6.3: Classification table: Staple Gun

Observed		Predicted		
		Observed Staple		Percentage Correct
		.00	1.00	
Observed Staple	.00	716	8	98.9
	1.00	4	55	93.2
Overall Percentage				98.5

6.7.3 Grinder

A data set of 163 frames long, including 124 observed grinder sounds, was recorded for the angle grinder.

The z for the Mastercraft angle grinder tool sound probability formula is:

$$z = 0 - 22.0940(Grinder1) - 9.8700(Grinder2) + 7.8800(Grinder3) + 6.6060(E) - 15.6180(CG)$$

Table 6.4: Classification table: Angle Grinder

Observed		Predicted		
		Observed Grinder		Percentage Correct
		.00	1.00	
Observed Grinder	.00	659	0	100.0
	1.00	1	121	99.2
Overall Percentage				99.9

6.7.4 Hammer

A data set of 266 frames long, including 65 observed hammer sounds, was recorded for the hammer tool.

The z for the hammer tool sound probability formula is:

$$z = 0 + 19.566(Hammer1) - 61.9380(Hammer3) + 28.4400(Hammer4) - 15.5210(EH)$$

Table 6.5: Classification table: Hammer

Observed		Predicted		
		Observed Hammer		Percentage Correct
		.00	1.00	
Observed Hammer	.00	715	3	99.6
	1.00	6	59	90.8
Overall Percentage				98.9

6.7.5 Summary of tool sound classifier

Table 6.6 lists different aspects of accuracy of all four models over 783 sample frames. Similar numbers can be observed from the jigsaw and grinder model. It is apparent that there has been a higher overall NPV and accuracy reported for all models, which are over 98% and 99% respectively. Further, zero FDR can be observed from both jigsaw and grinder models, while staple reported the worst at 12.7%. Most of these false detections (6/8) were observed in the midrange of energy frames while tapering the jigsaw and grinder sounds.

To sum up, jigsaw and grinder models displayed comparatively remarkable performance while the other two tools also have an acceptable level.

Table 6.6: Summary of accuracy: Tool sound classifier

Derivation from contingency table	Jigsaw model	Staple model	Grinder model	Hammer model
Sensitivity or true positive rate (TPR)	97.9	93.2	99.2	90.8
False positive rate (FPR) /false alarm	0.0	1.1	0.0	0.4
Accuracy (ACC)	99.6	98.5	99.9	98.9
Specificity (SPC) or true negative rate (TNR)	100.0	98.9	100.0	99.6
Positive predictive value (PPV) or precision	100.0	87.3	100.0	95.2
Negative predictive value (NPV)	99.5	99.4	99.8	99.2
False discovery rate (FDR)	0.0	12.7	0.0	4.8

6.8 Validation of Acoustic Sound Direction of Arrival

Direction of arrival (DOA) of an acoustic sound is analyzed based on time delay estimation (TDE) between all pairs of microphones and then combining them with the knowledge of the array geometry. As described in the 0, three commonly used TDE methods are adopted: cross correlation, PHAT method, and ML method. We used 4096 audio frame size in the audio sound detection classifier. However, we proposed much smaller sized audio frames (i.e. 32, 64, 128, 256, 512, 1024, and 2048) for the DOA analysis in order to improve the performance of the model. DOA of each detected sound is calculated using all methods that are considered as predicted DOA, and analysis suggests the best method for each tool. Table 6.7 depicts the model codes for all 21 methods. Figure 6.13 shows an image frame of a jigsaw tool operating worker. The actual x pixel coordinate (e.g. x=234) is manually picked from the image and in this validation process, the observed DOA is measured by transforming the pixel coordinate of the actual tool in the RGB image that is extracted in the same time frame. The Kinect covers 57.5 degrees in RGB image, which has a resolution of 640*480. Thus observed azimuth can be determined using the camera's internal geometry for a given pixel coordinate. The yellow, cyan, and green lines indicate the DOA from 2048 frame sized CC, GCC, and ML models.

Table 6.7: Model codes for DOA

Cross correlation (CC)	32CC	64CC	128CC	256CC	512CC	1024CC	2048CC
Phase Transform (GCC-PHAT)	32GCC	64GCC	128GCC	256GCC	512GCC	1024GCC	2048GCC
Maximum Likelihood (ML)	32ML	64ML	128ML	256ML	512ML	1024ML	2048ML

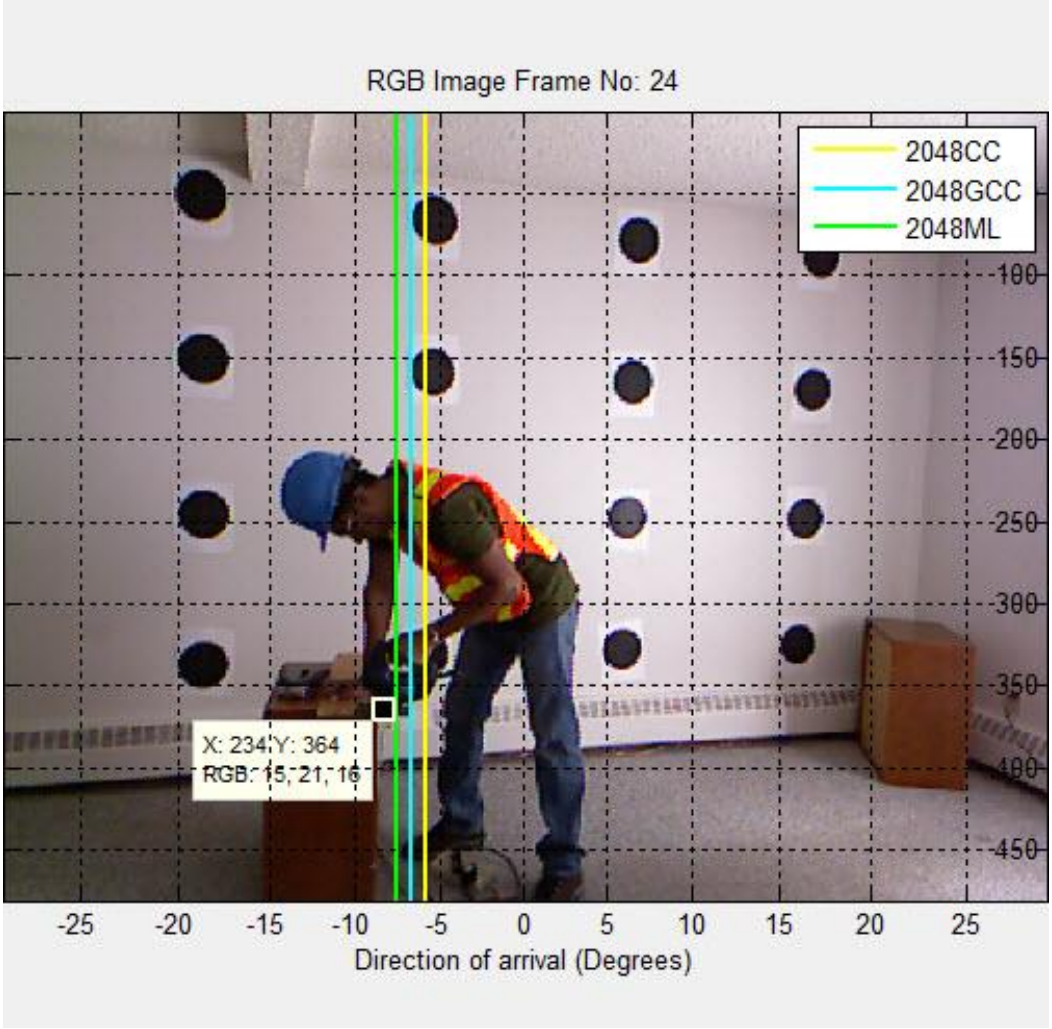


Figure 6.13: Visualization of observed and predicted DOA

The error values of DOA are then calculated against the observed direction. Figure 6.14, Figure 6.17, Figure 6.20, and Figure 6.23 illustrate the Box-and-Whisker plot of error values of DOA of tool sound against different proposed models (i.e. CC, GCC-PHAT and ML). The maximum whisker length w is set to 1.5. Points are drawn as outliers if they are larger than $q_3 + w(q_3 - q_1)$ or smaller than $q_1 - w(q_3 - q_1)$ where q_1 and q_3 are the 25th and 75th percentiles respectively. The plotted Whisker extends to the adjacent value, which is

the most extreme data value that is not an outlier. Pearson correlation coefficient, Q1, Q3, Whisker extended adjustment (upper and lower), and finite outliers are considered when selecting the optimum DOA model for a tool sound. The selected model is further validated by analyzing the positions of predicted DOA, worker silhouette bounding box, and worker skeleton hand location. Then the maximum allowable pixel range (horizontal) between worker and DOA is determined.

6.8.1 Jigsaw

Figure 6.14 illustrates the Box-and-Whisker plot for the error values of DOA in various jigsaw models. It can clearly be seen that DOA models of higher frame size increase the quality of results by reducing Whisker and Q3-Q1 range. Lesser range interprets a smaller variance around the zero value. However, CC and GCC models show better results than ML models. Error results from smaller sized frames show larger Whisker lengths and higher variances around the zero value. Out of these 21 models, 1024CC, 2048CC, and 2048GCC models are considered competitive models for the jigsaw model.

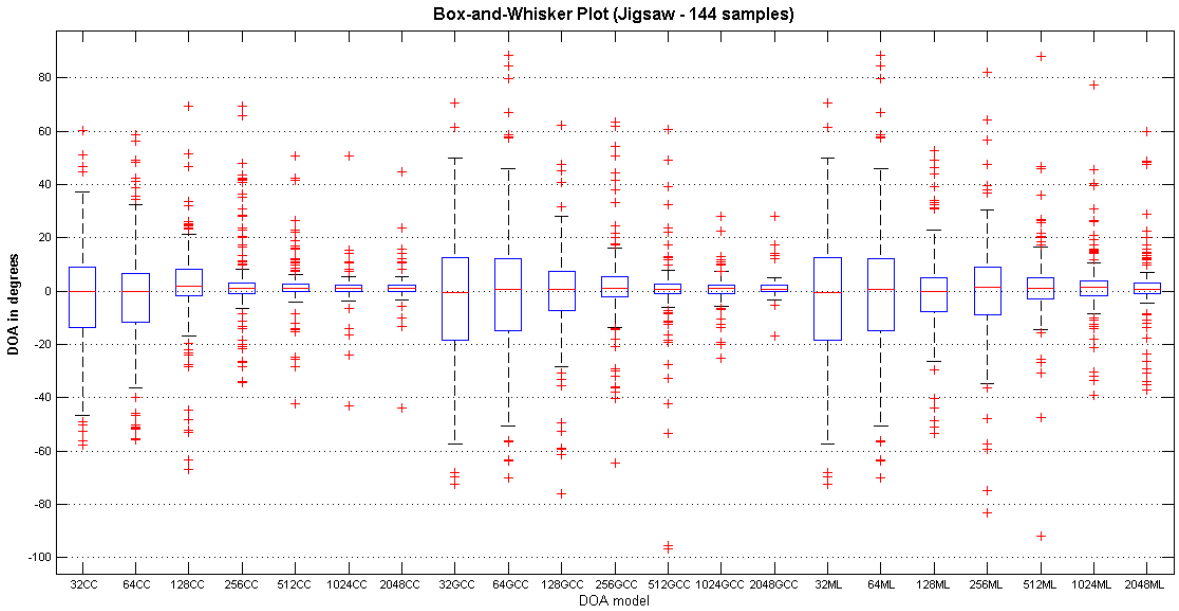


Figure 6.14: Box-and-Whisker Plot (Jigsaw)

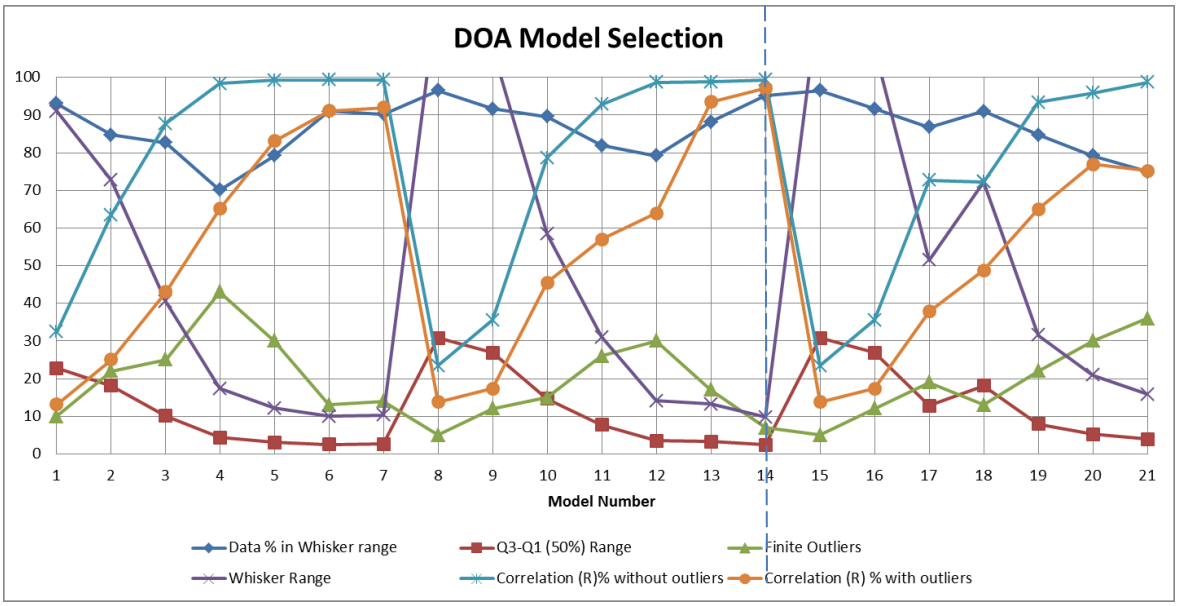


Figure 6.15: DOA parameters (Jigsaw)

Table 6.8: DOA model parameters: Jigsaw

Parameter	1024CC	2048CC	2048GCC
Total data samples	144	144	144
Upper	5.94	6.19	5.86
Q3	2.21	2.33	2.19
Q2	1.08	0.90	0.80
Q1	-0.28	-0.25	-0.26
Lower	-4.01	-4.11	-3.93
Finite Outliers	13	14	7
Data % in Whisker Range	90.97	90.28	95.14
Q3-Q1 (50%) range	2.49	2.58	2.45
Whisker Range	9.95	10.30	9.80
Pixel Range (50%)	12.66	13.11	12.47
Pixel Range (Whisker)	50.77	52.59	49.98
Pearson Correlation % with outliers	91.06	92.00	97.06
Pearson Correlation % without outliers	99.29	99.29	99.38

As shown in Figure 6.15 and Table 6.8, 2048GCC (model number 14) has the highest correlation coefficient value (97.06) even with the outliers in the model, and 2048GCC model has the lowest finite outliers (i.e. 7) compared to the other 2 considered models. Further, more than 95% of the data has a less than 10 degree error margin. Hence, 2048GCC model is selected for predicting DOA of jigsaw tool sound. The following figure depicts the error of DOA distribution of 2048GCC and 2048CC models.

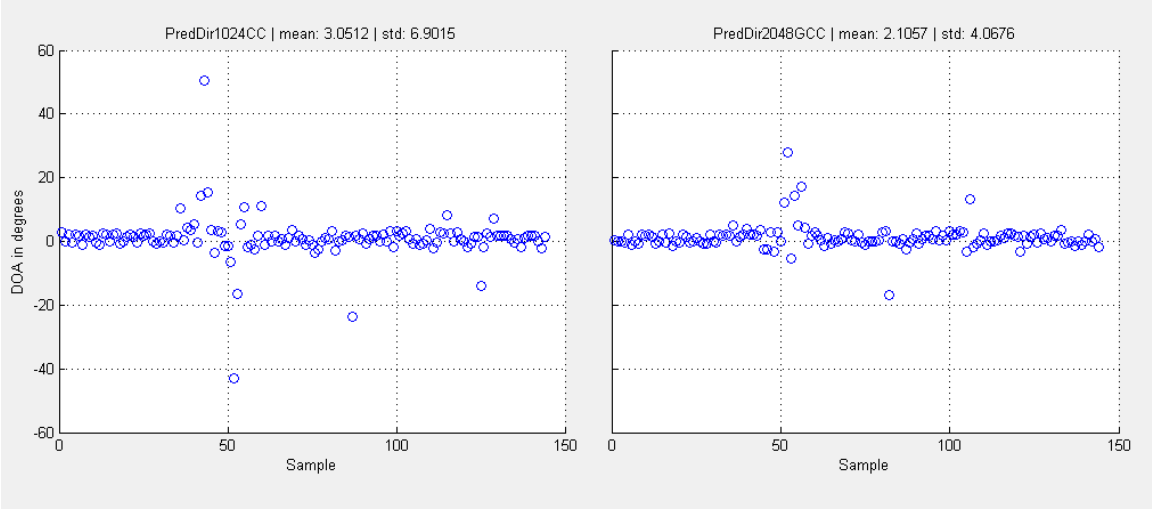


Figure 6.16: Error of DOA – Jigsaw (1024CC & 2048GCC models)

6.8.2 Staple Gun

Figure 6.17 shows an inverse trend compared to the jigsaw DOA models. The graph shows the models of smaller frame size increase the quality of results by reducing Whisker and Q3-Q1 range. However, this argument fails in the 32 frame size but the trend continues to other sizes.

Hence, 64CC, 128CC, and 64GCC are considered competitive models for the staple DOA model. As the parameter comparison illustrates in Figure 6.18, the cross correlation based TDE model (64CC) is selected (model number 2 in Figure 6.18).

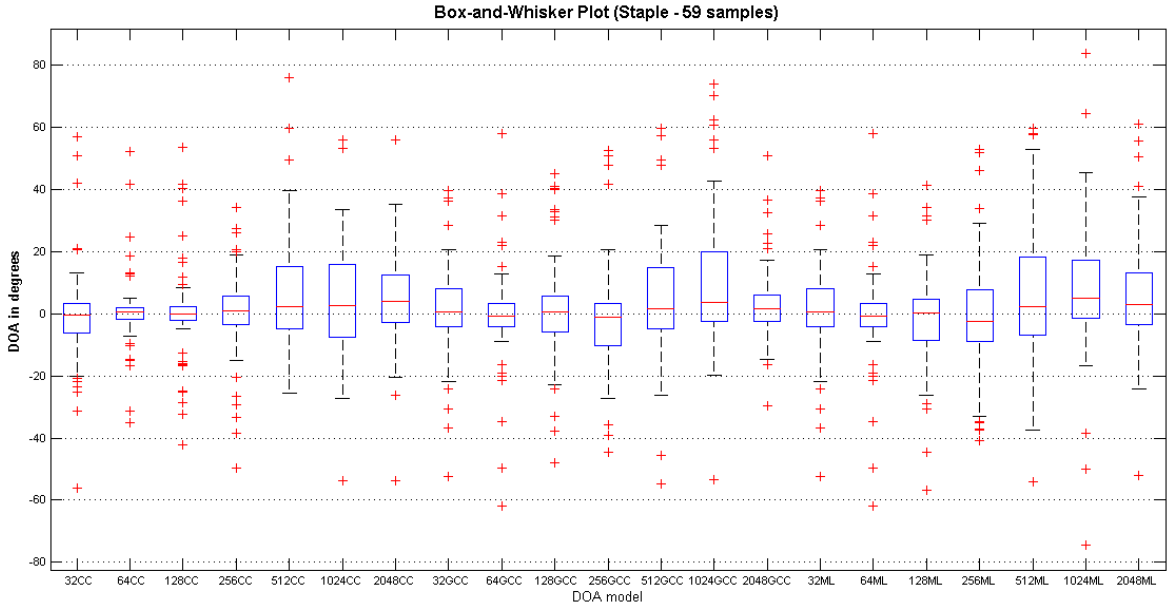


Figure 6.17: Box-and-Whisker Plot (Staple)

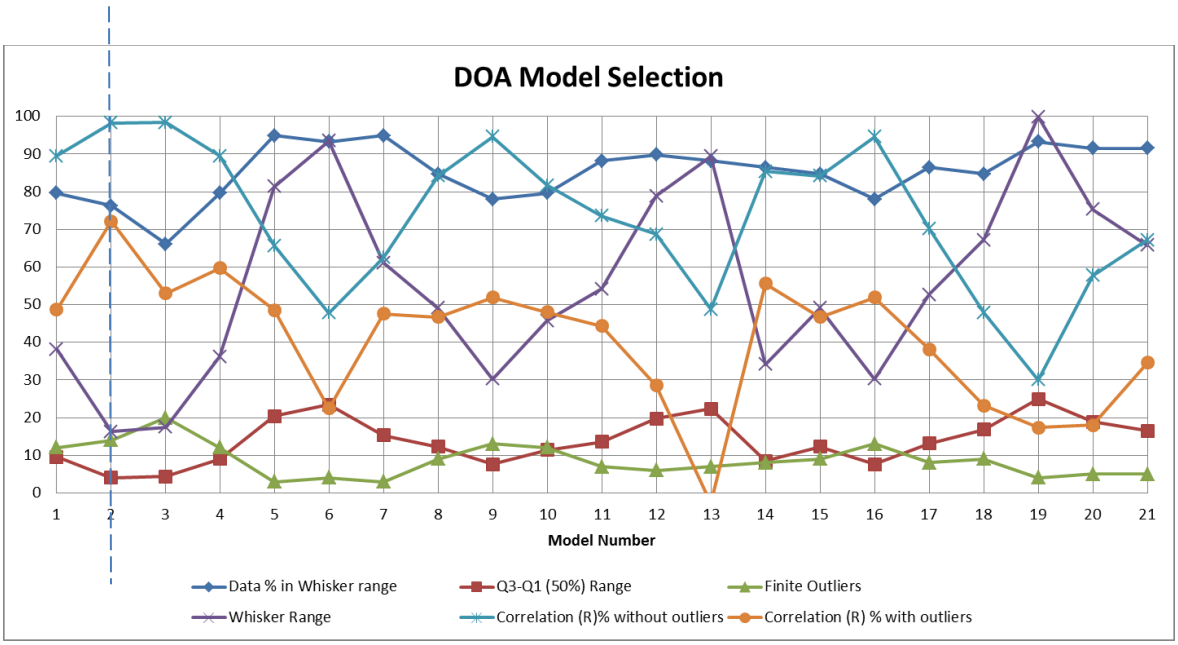


Figure 6.18: DOA parameters (Staple)

Table 6.9: DOA model parameters: Staple

Parameter	64CC	128CC	64GCC
Upper	8.24	8.74	14.81
Q3	2.15	2.20	3.45
Q2	0.69	-0.02	-0.60
Q1	-1.91	-2.16	-4.12
Lower	-7.99	-8.70	-15.48
Finite Outliers	14	20	13
Data % in Whisker Range	76.27	66.10	77.97
Q3-Q1 (50%) range	4.06	4.36	7.57
Whisker Range	16.23	17.44	30.29
Pixel Range (50%)	20.66	22.21	38.59
Pixel Range (Whisker)	83.16	89.48	157.85
Pearson Correlation % with outliers	72.13	53.03	51.86
Pearson Correlation % without outliers	98.25	98.35	94.60

Fifty percent (50%) of predicted DOA values have approximately two degrees of error margin with regards to the actual acoustic sound direction when considering the figures for Q1 and Q3 in 64CC model. Further, more than 75% of the collected data are in the Whisker range (-7.99, 8.24). One way to make use of the DOA information in the model is to use the results to match the worker and the tool sound direction. Hence, after the DOA-to-pixel transformation and pixel error analysis, the model demonstrates approximately 20 pixel precision of predicting DOA for the dataset (Q1 to Q3).

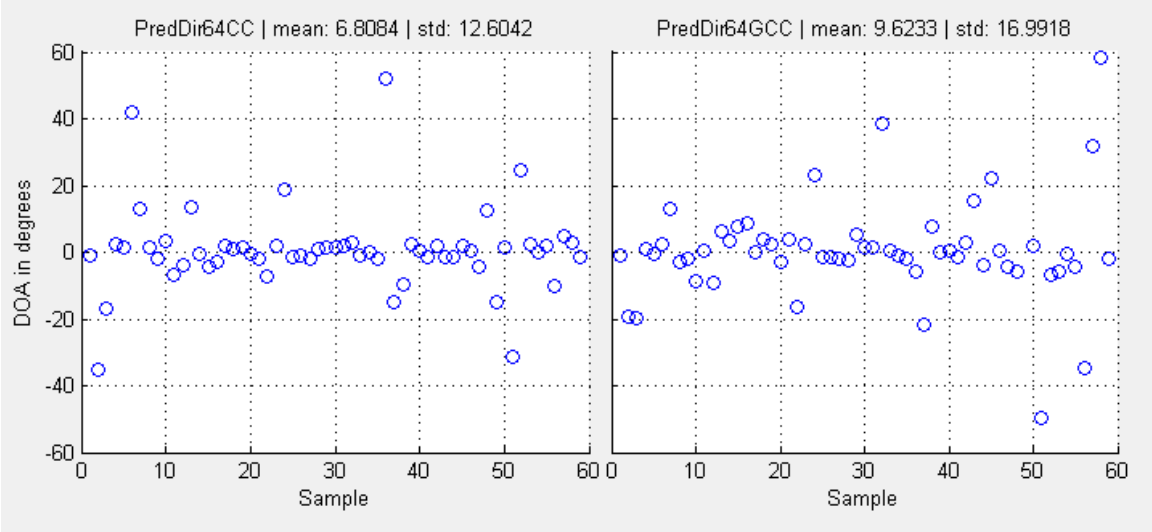


Figure 6.19: Error of DOA – Staple Gun (64CC model)

6.8.3 Grinder

Reviewing the Box-and-Whisker Plot reveals that when the frame size is higher, the quality of DOA increases in all three models. Considering Figure 6.20 and Figure 6.21, 2048CC and 2048GCC models are shortlisted as the better DOA models. Apart from slight differences, both models show similar figures in most of the parameters. However, 2048CC model is selected because of its low computational cost, lesser finite outliers, and higher correlation coefficient value (including outliers). Nearly 90% of the data has an error margin between +5 and -6.5 degrees. Further, 50% of the data are plotted within the 15pixel range to the actual location.

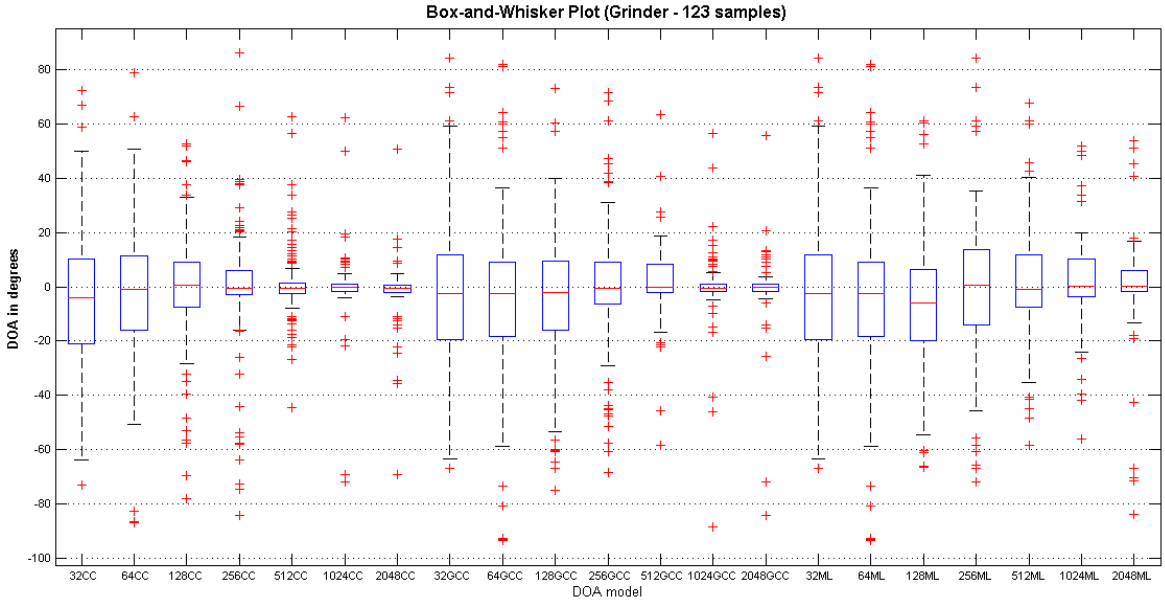


Figure 6.20: Box-and-Whisker Plot (Grinder)

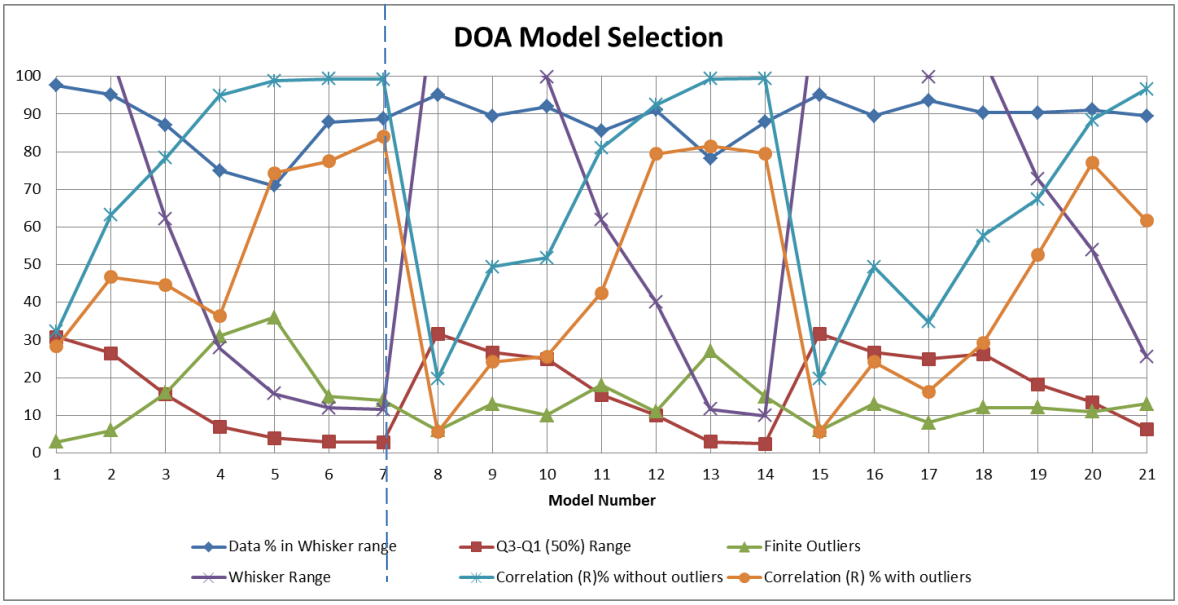


Figure 6.21: DOA parameters (Grinder)

Table 6.10: DOA model parameters: Grinder

Parameter	2048CC	2048GCC
Total data samples	122	122
Upper	4.96	4.50
Q3	0.65	0.81
Q2	-0.50	-0.22
Q1	-2.23	-1.65
Lower	-6.54	-5.34
Finite Outliers	14	15
Data % in Whisker Range	88.71	87.90
Q3-Q1 (50%) range	2.87	2.46
Whisker Range	11.50	9.84
Pixel Range (50%)	14.64	12.52
Pixel Range (Whisker)	58.73	50.20
Pearson Correlation % with outliers	83.98	79.49
Pearson Correlation % without outliers	99.20	99.41

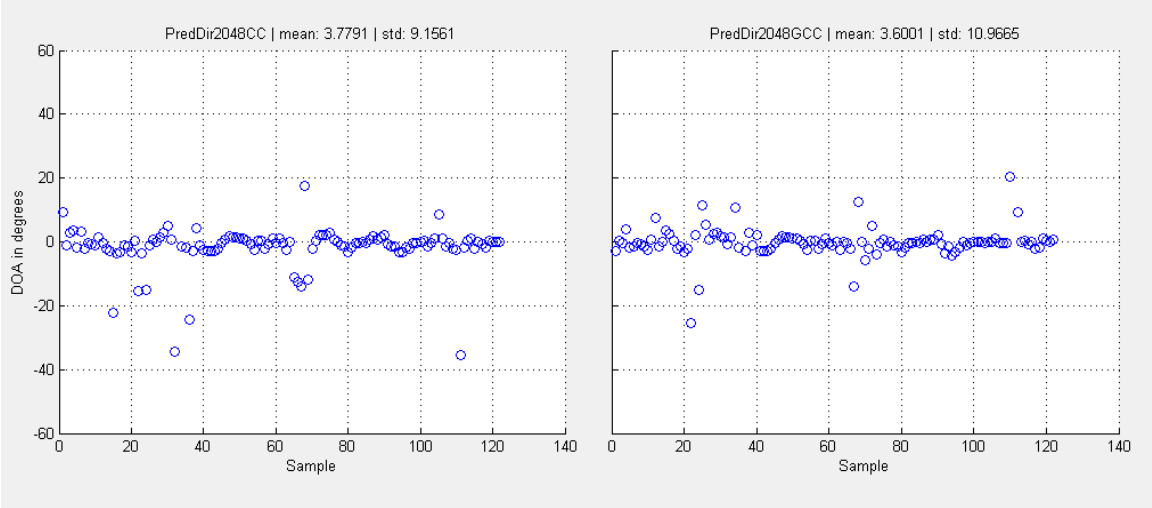


Figure 6.22: Error of DOA - Grinder (2048CC & 2048GCC model)

6.8.4 Hammer

Box-and-Whisker plot (Figure 6.23) of the hammer sound proves that most of the models do not provide an accurate or significant DOA, but figures from only the 2048GCC model express somewhat of a correlation for the hammer sound. A total of 68 hammer sound samples were observed and fed into the DOA model. This shows nearly 60% of correlation with the actual direction excluding 14 finite outliers. The main reason for distorted DOA is this hammer sound does not originate from a single location but rather from a vibrated plane. These reverberations highly distort the TDE and consequently it generates false directions.

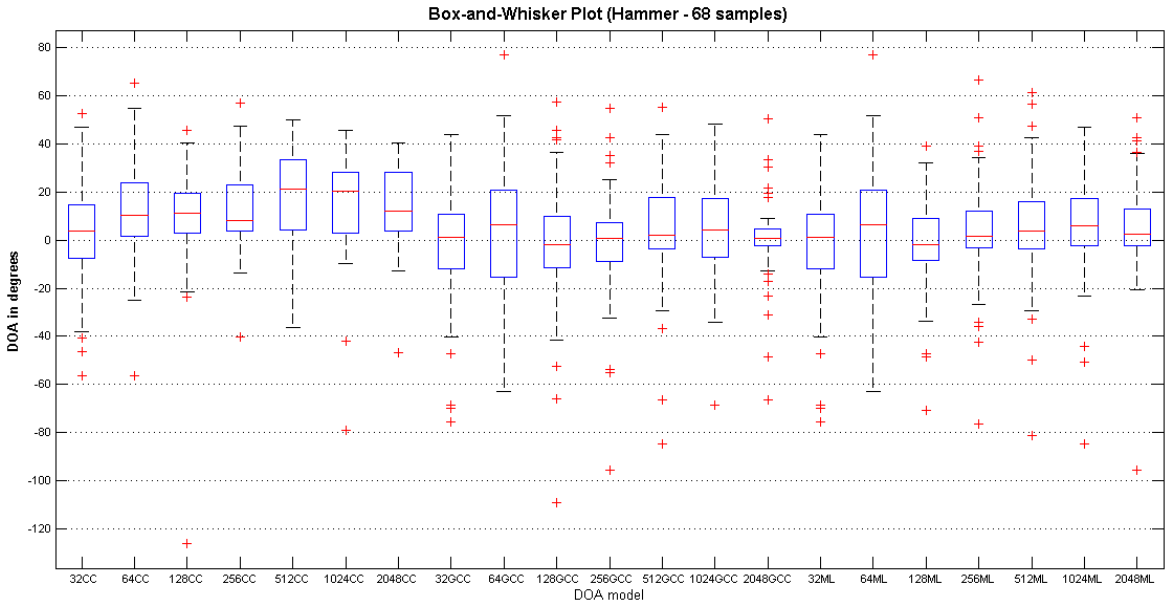


Figure 6.23: Box-and-Whisker Plot (Hammer)

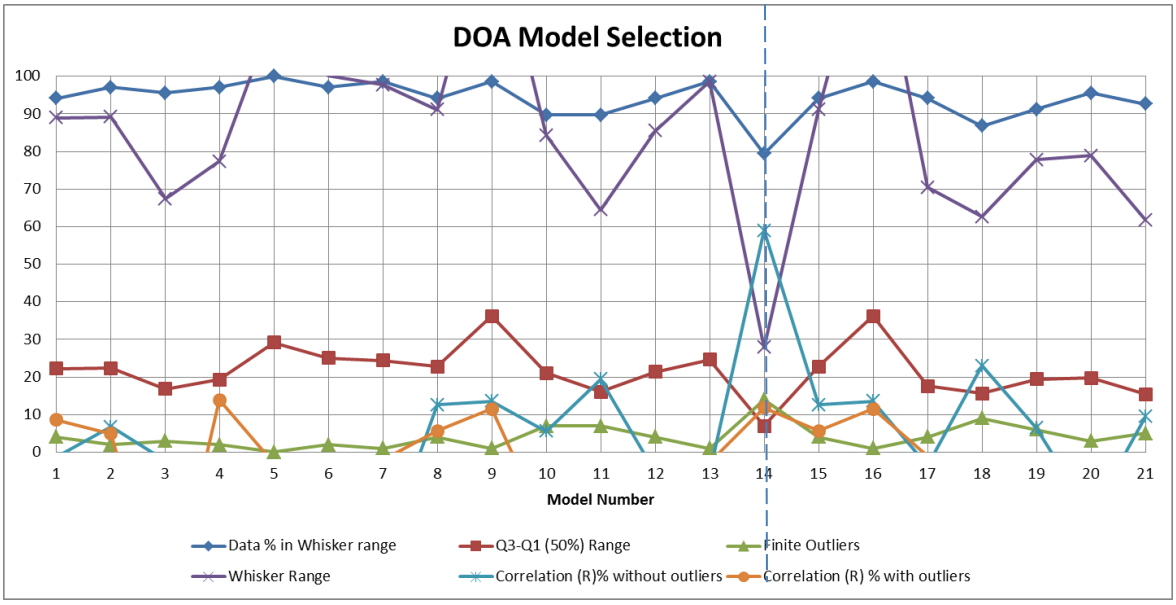


Figure 6.24: DOA parameters (Hammer)

Table 6.11: DOA model parameters: Hammer

Parameter	2048GCC
Total data samples	68
Upper	15.04
Q3	4.60
Q2	0.69
Q1	-2.36
Lower	-12.79
Finite Outliers	14
Data % in Whisker Range	79.41
Q3-Q1 (50%) range	6.96
Whisker Range	27.84
Pixel Range (50%)	35.47
Pixel Range (Whisker)	144.54
Pearson Correlation % with outliers	11.80
Pearson Correlation % without outliers	58.92

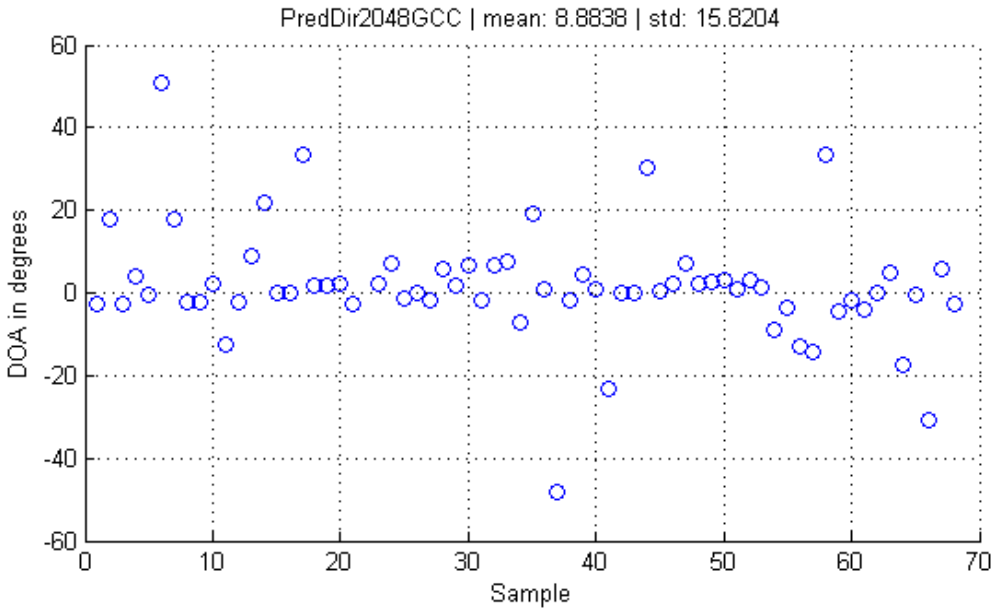


Figure 6.25: Error of DOA - Hammer (2048GCC model)

6.9 Summary of DOA

The three TDE methods described in this study performed well in the case of moderate SNR in the actual environment. Table 6.12 summarizes the parameters of selected DOA models.

It is evident that there has been a similar pattern of performance displayed in jigsaw, staple, and grinder, while figures for hammer reveal considerably lower performance values. Moreover, correlation coefficients of the hammer model do not show sufficient statistical relationship between predicted and actual values. Hence, the hammer model is omitted from the discussion.

The selected models and frame sizes demonstrate that discrete events (staple sounds) perform better in shorter frames, while continuous events (jigsaw and grinder sounds) perform well in larger frames. The figures of finite outliers, correlation

coefficient, and data percentage in Whisker range prove the jigsaw model is the best from all constructed DOA models.

In conclusion, jigsaw, staple, and grinder models demonstrated the potential to be in the automated system, and Correlation and GCC-PHAT were adopted due to their performance, simple computation, and easy detection.

Table 6.12: DOA parameter comparison

Parameter	Jigsaw	Staple	Grinder	Hammer
Selected model	2048GCC	64CC	2048CC	2048GCC
Total data samples	144	59	122	68
Upper	5.86	8.24	4.96	15.04
Q3	2.19	2.15	0.65	4.60
Q2	0.80	0.69	-0.50	0.69
Q1	-0.26	-1.91	-2.23	-2.36
Lower	-3.93	-7.99	-6.54	-12.79
Finite Outliers	7	14	14	14
Data % in Whisker Range	95.14	76.27	88.71	79.41
Q3-Q1 (50%) range	2.45	4.06	2.87	6.96
Whisker Range	9.80	16.23	11.50	27.84
Pixel Range (50%)	12.47	20.66	14.64	35.47
Pixel Range (Whisker)	49.98	83.16	58.73	144.54
Pearson Correlation % with outliers	97.06	72.13	83.98	11.80
Pearson Correlation % without outliers	99.38	98.25	99.20	58.92

6.10 SNR Threshold Analysis

The correlation between SNR level and accuracy of TDE has been analyzed using actual noise samples adopted as the additive noise to the collected tool sound sample. In order to study the performance of the TDE, the following experiments have been set.

As mentioned in 0, the allowable time delays in Kinect fall in a range between -11 and +11 audio frames, provided that the sampling frequency is 16000Hz. Hence, a second signal was created with an actual time delay of 10 frames. Then TDE was calculated using selected DOA models (2048GCC, 64CC, and 2048CC) in various SNR levels obtained by altering the noise power.

Figure 6.26 integrates the results of the estimated time delay for three tools. The x-coordinate presents the various SNR values, while the y-coordinate presents the estimated time delay.

It can be clearly seen that TDE is distorted when the SNR exceeds a certain threshold in lower bound. Different trends can be observed from three tools while figures for the jigsaw showed the greatest lower SNR threshold. It can be seen from the above analysis that only minor error values have been detected from the DOA model constructed for the staple gun sound.

To sum up, all three models showed a distorted TDE after certain SNR thresholds and these limits are extracted from the figure as -14dB, -6dB and -10dB for jigsaw, staple gun, and angle grinder respectively.

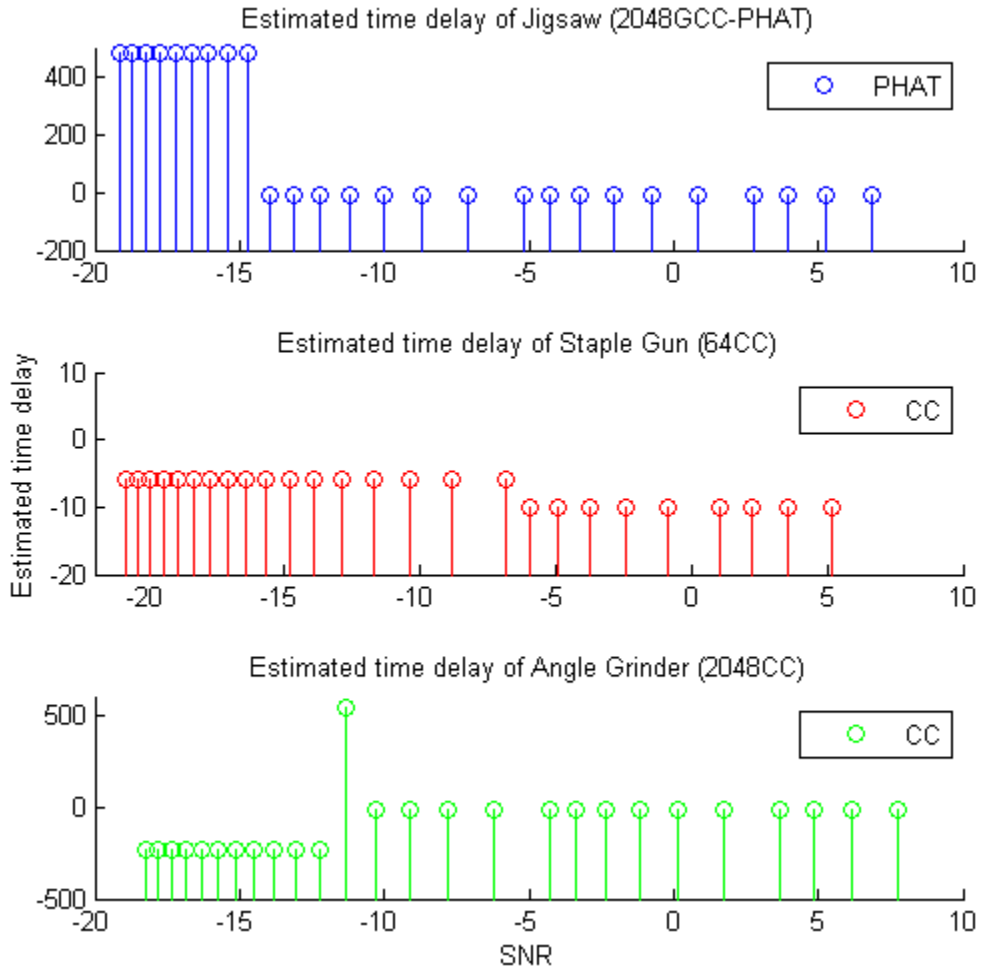


Figure 6.26: Accuracy variation of DOA models over different SNR levels

6.11 Pixel Threshold Analysis

Analysis of worker position and DOA prediction is efficiently used to amalgamate recognized workers and detected activities. Two different approaches have been studied for the selected models of jigsaw, staple, and grinder tools:

1. Proximity of DOA to worker silhouette bounding box
2. Proximity of DOA to worker wrist/hand positions

Figure 6.27 illustrates the input parameters for the analysis: DOA, bounding box positions, and skeleton joints.

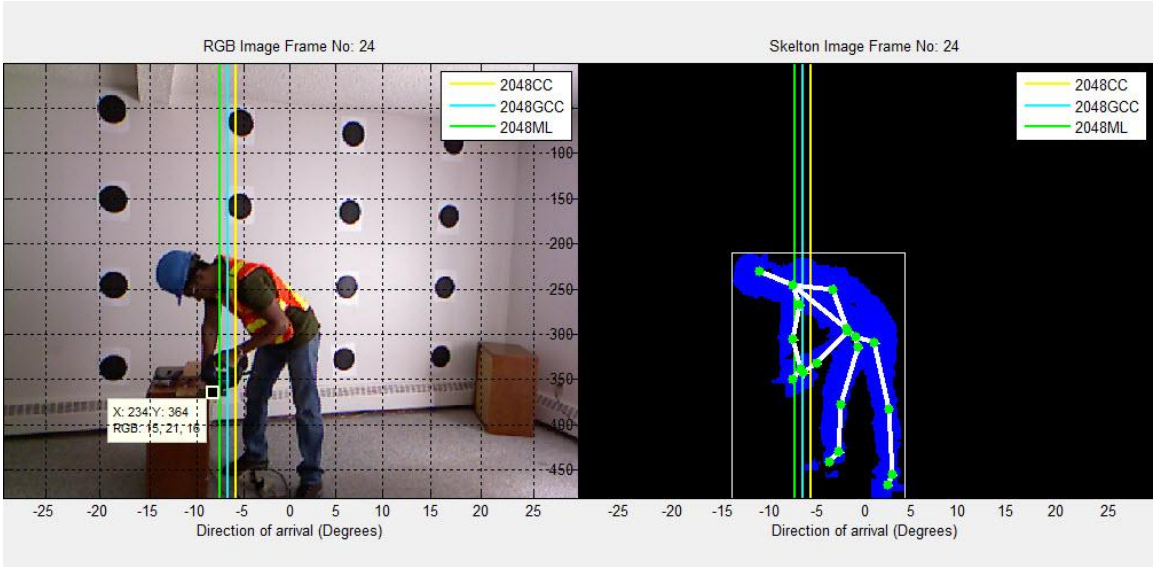


Figure 6.27: Visualization of DOA, silhouette bounding box, and skeleton joints

6.11.1 Proximity analysis of DOA to worker silhouette bounding box

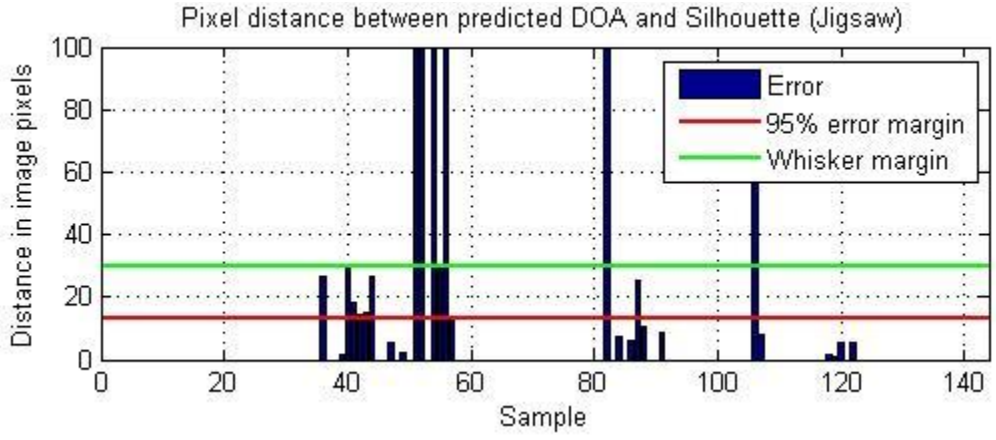
The analysis has been carried out based on the horizontal position of silhouette bounding box (BB) extracted from the skeleton image frame. In this approach, the horizontal (azimuth) pixel value of the DOA is analyzed as inside BB and outsize BB.

Then absolute pixel error, the distance between DOA and closest BB margin is calculated from samples categorized as outside BB. On the other hand, absolute pixel error has been taken as zero for the rest of the samples (inside BB). Figure 6.28 illustrates the calculated error distribution for observed sound samples for each tool and Table 6.13 depicts the outline results from the error distribution. It further indicates two threshold lines:

1. Whisker margin – threshold pixel distance that covers all data points in Whisker range
2. 95% error margin – threshold pixel distance that covers 95% of the data points in Whisker range.

It can be clearly seen that the grinder model covers the Whisker range and 95% of data from the lowest threshold levels, which are 21 and 7 pixels respectively, compared to the rest of the tools. At the same time the jigsaw model covers more than 95% of total data points within a perimeter of 30 pixels from the bounding box. Another significant factor in these diagrams is that more than 85% of the Whisker data fell inside the bounding box in both jigsaw and grinder models.

The figures from the staple DOA model demonstrate slightly lower performance compared to other models. However, approximately half of the total data points are covered from the BB while the 46 pixel threshold covers all Whisker data points.



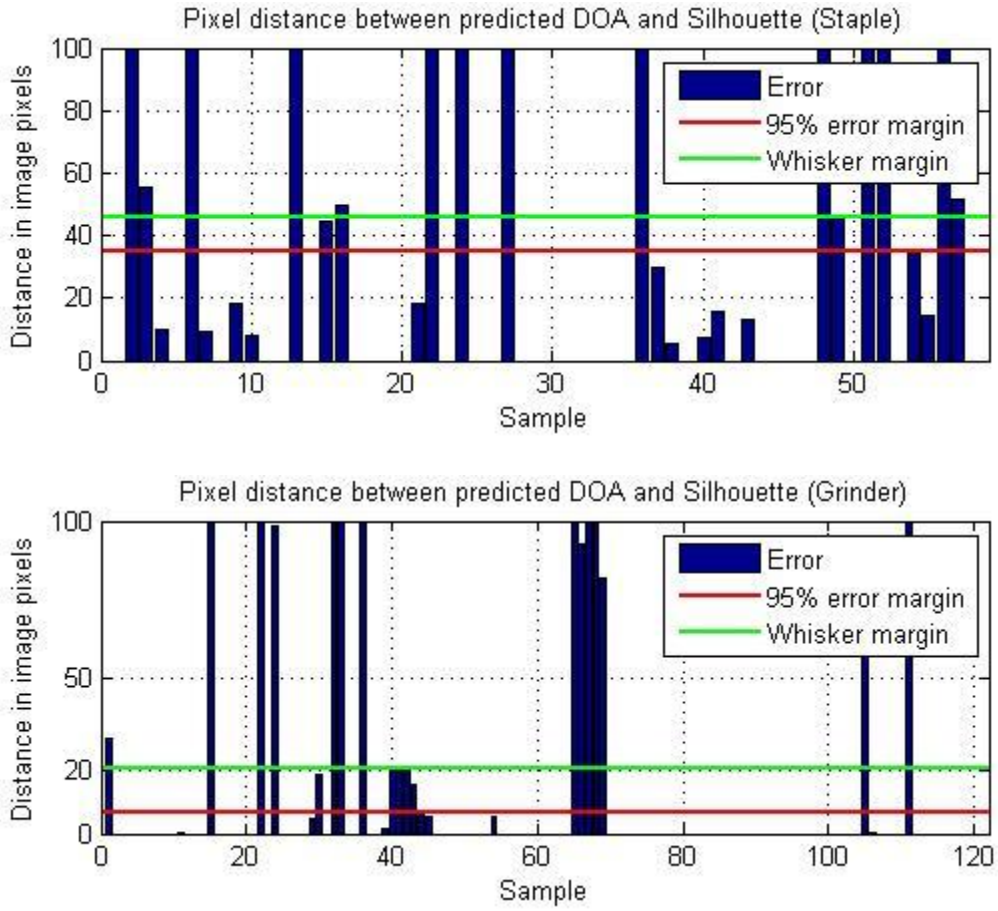


Figure 6.28: Pixel distance between predicted DOA and silhouette

Table 6.13: Pixel threshold (silhouette bounding box)

Parameter	Jigsaw	Staple	Grinder
Selected model	2048GCC	64CC	2048CC
Total data samples	144	59	122
Finite Outliers	7	14	14
Data % in Whisker Range	95.14	76.27	88.71
Pixel error for Whisker range	30	46	21
Pixel error for 95% Whisker range	14	35	7
Number of samples inside B. Box	117	28	96
Data % inside B. Box	85.40	62.22	88.9

6.11.2 Proximity analysis of DOA to worker wrist/hand positions

The proximity analysis of DOA to worker wrist position is another study carried out to find the pixel threshold to merge the worker and the detected construction activity. The pixel distance was measured from DOA to closest wrist of a worker and Figure 6.29 illustrates the pixel error distribution over the collected data points for jigsaw and grinder models. The pixel threshold level was determined based on the Whisker range and 95% confidence data coverage. Table 6.14 lists the results of the analysis.

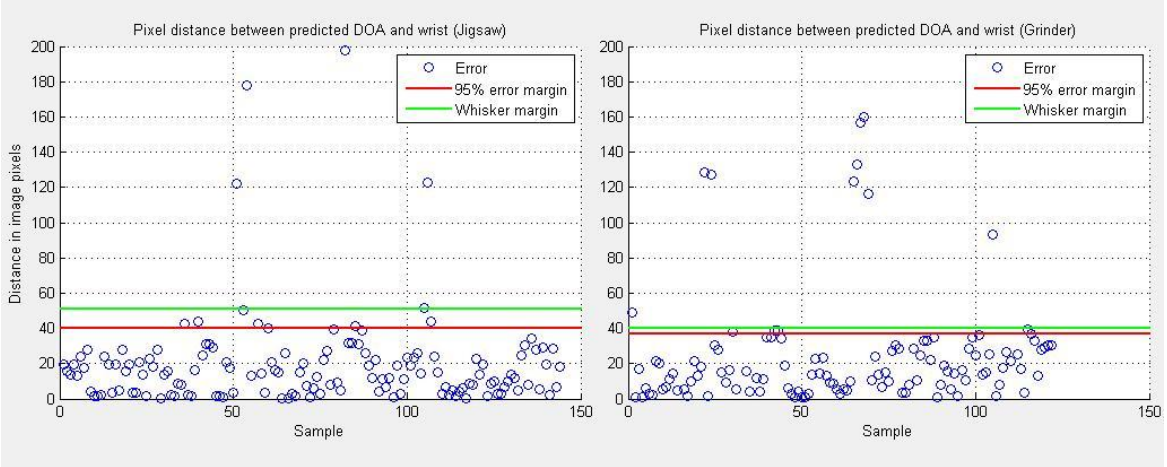


Figure 6.29: Pixel distance between predicted DOA and wrist

Table 6.14: Pixel threshold (wrist position)

Parameter	Jigsaw	Grinder
Pixel error for Whisker range	51	40
Pixel error for 95% Whisker range	40	37

6.12 Summary of Pixel Threshold Analysis

Proximity analysis of DOA to worker wrist is more sensible and theoretically the best approach to finding the pixel threshold to merge worker and detected construction

activity. However, the higher possibility of distorting positions of skeleton joints, especially the latter links of hands, may cause unreliable results. These distorted positions can often occur because of partial occlusions by other parts of the body.

On the other hand, the chance of getting a distorted BB is low in BB analysis and at the same time lower pixel threshold values have been discovered for each tool. Having considered BB analysis, it is also reasonable to look at the 95% error margin which shows a significant reduction of threshold from Whisker margin and this helps to reduce the chance of getting overlaps and multiple holders.

Hence, threshold values obtained by a 95% error in silhouette analysis are proposed for use in the model. These pixel thresholds are 14, 35, and 7 for jigsaw, staple gun, and grinder respectively.

Chapter Seven: **Conclusions and Recommendations**

7.1 Introduction

This chapter outlines the main research findings, reflecting the introductory thesis objectives synthesized with research problems, and integrated with various issues raised in the discussion sections. It further discusses research contributions, limitations of the research, and the barriers the researcher had to overcome over the period, and it proposes appropriate directions for future research.

7.2 Summary of the Research

The primary research idea was the probable reduction or prevention of manual observation in tool time and performance monitoring, in order to reduce human errors, labour cost, and data limitations. While the study was progressing, this primary idea grew into many other secondary areas, which made the study more comprehensive.

The primary objective of the research is to develop a sustainable, integrated, automated, and systematic mechanism to extract construction worker tool-time and performance information by using multiple modalities (audio and video) addressing the potential drawbacks of manual observation on construction sites.

A low-cost range sensor, the Microsoft Kinect device, was used in the research to capture audio and video data in an indoor work site. A MATLAB model, which integrated a user-friendly graphical user interface (GUI), real-time Kinect data, and Simulink, was developed to track construction workers and their activities. Consequently,

this location aware information is used to accurately estimate the tool-time and performance of workers on a job site.

The following hypothesis was assumed for the research:

A systematic examination of multiple modalities using a combination of signal processing techniques and statistical approaches will provide a direct way of finding location aware information of workers and activities.

This hypothesis was tested in series of steps in various chapters throughout the thesis and conclusions will be discussed in the next few sections.

7.3 Summary of Main Research Findings

7.3.1 Worker tracking system

This section provides answers to construction worker tracking related research problems stated in Chapter Three:. This includes techniques we used to detect and differentiate people on site, types of features used in the model, positioning methods, and accuracies of the developed model.

The research introduced an image processing based, efficient worker tracking technique by analyzing the unique shape and colour of the construction hardhat. The features of the key tracking object (i.e. colour difference) further provides a solution for differentiating people on site based on their work type (i.e. worker, foreman, supervisor and engineer).

A statistical model (i.e. logistic regression model containing image features as independent variables) has been introduced to track construction workers in indoor work

area and a dataset of 100 individual image frames consisting of different hardhats was collected to model the hardhat classifier. The constructed hardhat classifier demonstrates 99% accuracy and 98.5% of TPR for its model samples.

The hardhat classifier has been validated against the observed data of 900 image frames and all four coloured hardhats have displayed more than 94% of the accuracy in detection and exposed less than 4% of false prediction rate (FPR).

7.3.2 Activity recognition system

This section provides answers to construction activity recognition related research problems stated in Chapter Three:. This includes techniques we used to detect, types of activities we modeled, positioning methods, and accuracies of all constructed models.

An audio signal processing based system has been proposed for the activity recognition by analyzing unique sound patterns of construction tool sounds.

A logistic regression model containing distinctive audio features was constructed with a dataset of 250 audio frames. Four commonly used tool sound activities (i.e. jigsaw, angle grinder, staple gun, and hammer) were proposed to be identified by the constructed model.

Table 7.1 is reproduced from Table 5.25 to show the summary of parameters of the constructed tool sound classifier. It can be clearly seen that each model exceeds 95% accuracy; goodness-of-fit (Nagelkerke R squared value) exceeds 92%, and has approximately 100% of Hosmer and Lemeshow significance.

Table 7.1: Selection parameters of activity classifier – model construction

Parameters	Jigsaw model	Staple model	Grinder model	Hammer model
Nagelkerke R Square	0.929	0.953	0.963	0.988
Hosmer & Lemeshow model Significance	0.996	0.999	1.000	1.000
Maximum variable significance	0.008	0.009	0.074	0.017
Overall Accuracy %	95.6	96.4	99.2	99.2
Sensitivity or true positive rate (TPR)	0.960	0.980	0.960	1.000
False positive rate (FPR) /false alarm	0.025	0.030	0.000	0.015

Four audio data sets were recorded for each tool and total tool sound classifier has been validated for the combined data set of 783 frames. Parameters of the validated tool sound classifier model are taken from the Table 5.23 and relisted in Table 7.2 and figures show more than 98% accuracy for all models.

Table 7.2: Accuracy percentages of activity classifier – model validation

Parameter	Jigsaw model	Staple model	Grinder model	Hammer model
Sensitivity or true positive rate (TPR)	97.9	93.2	99.2	90.8
False positive rate (FPR) /false alarm	0.0	1.1	0.0	0.4
Accuracy (ACC)	99.6	98.5	99.9	98.9
Positive predictive value (PPV) or precision	100.0	87.3	100.0	95.2

In order to determine the position of the originated sound source, we proposed measuring the direction of arrival (DOA) by analyzing the time delay estimation (TDE) between all pairs of microphones and then combining them with the knowledge of the array geometry.

Three commonly used TDE methods were adopted: cross correlation, PHAT method, and ML method. Each model was validated using several data samples (i.e.

observed tool sound frames from the previously discussed 783 dataset) and shows some key parameters of each selected model. Except for the hammer model, the other three demonstrated higher performance figures, and the jigsaw model can be considered the best out of them. In brief, Correlation and GCC-PHAT methods were adopted due to their performance, simple computation, and easy detection. Table 6.12 is taken from the section 6.9, and illustrated below for further information.

Table 7.3: DOA parameter comparison

Parameter	Jigsaw	Staple	Grinder	Hammer
Selected model	2048GCC	64CC	2048CC	2048GCC
Total data samples	144	59	122	68
Upper	5.86	8.24	4.96	15.04
Q3	2.19	2.15	0.65	4.60
Q2	0.80	0.69	-0.50	0.69
Q1	-0.26	-1.91	-2.23	-2.36
Lower	-3.93	-7.99	-6.54	-12.79
Finite Outliers	7	14	14	14
Data % in Whisker Range	95.14	76.27	88.71	79.41
Q3-Q1 (50%) range	2.45	4.06	2.87	6.96
Whisker Range	9.80	16.23	11.50	27.84
Pixel Range (50%)	12.47	20.66	14.64	35.47
Pixel Range (Whisker)	49.98	83.16	58.73	144.54
Pearson Correlation % with outliers	97.06	72.13	83.98	11.80
Pearson Correlation % without outliers	99.38	98.25	99.20	58.92

Analysis of worker position and DOA prediction is efficiently used to amalgamate recognized workers and detected activities. Proximity analysis of DOA to silhouette bounding box is carried out and threshold values obtained by a 95% error in silhouette

analysis are proposed for use in the model. These pixel thresholds are 14, 35, and 7 for jigsaw, staple gun, and grinder respectively.

In addition, SNR thresholds were analyzed for each tool using actual noise samples and resulted -14dB, -6dB and -10dB for jigsaw, staple gun, and angle grinder respectively. This is evident that the system can be implemented in a typical indoor construction site with a moderate noise level.

In conclusion, it can be seen from each of these theoretical and practical implications that the null hypothesis is rejected with proven results, thus evidently the alternative hypothesis is accepted, "a systematic examination of multiple modalities (i.e. RGB, depth and multi-channel audio) using a combination of signal processing techniques and statistical approaches will provide a direct way of tracking workers, recognizing construction activities and finding location aware information of workers and activities.

7.4 Major Research Contributions

This research project provided several contributions to the body of knowledge in the areas of projects management, construction automation, worker productivity improvement, and information technology in construction. Some of the research findings reiterated the known facts previously investigated and unveiled by many other researchers. In addition there were new ideas, concepts, and techniques introduced by the research study. The following is a list of many different research contributions:

a) For the first time in the construction industry, this research project introduced an audio signal processing based construction activity recognition technique with proven results to justify the idea. The tool detection classifier has been constructed for four different commonly used tools: jigsaw, angle grinder, staple gun, and hammer.
b) Developed an efficient, simple, and accurate TDE based DOA system to recognize acoustic sound source direction of tool sounds in order to differentiate worker activities.
c) Introduced a low cost and compact Microsoft Kinect device to the construction industry as a multiple data acquiring device.
d) Developed an automated tool time and performance evaluation system based on activity recognition of construction workers. Tool time provides an estimation of the productivity on site and with the use of this framework, construction project managers and planners will be able to develop strategies for improving labour productivity and labour allocation, and can develop administrative schemes related to labour performance.
e) Introduced an efficient worker tracking technique by analyzing the unique shape and colour of the construction hardhat. This method further provides a solution for differentiating people on site based on their work type (i.e. worker, foreman, supervisor, and engineer).

f) Introduced a new perspective to indicate and measure the supervisory effect on tool time and performance of a worker by analyzing supervisory presence on a job site when performing the task.

g) Developed a system to automate the classification of non-tool time activities into different categories (i.e. material handling, tool handling, etc.) by using movements in geo zones declared on the job site.

h) Developed a set of application modules to support colour settings, single photo resection, geo zone declaration, 3D point cloud, and tool time and performance evaluation, which can be used as a complete tool kit at project managerial level.

7.5 Research Limitations

There were certain limitations encountered by the researcher from the initial planning stage throughout the research lifecycle. The main limitations were in regards to the research equipment used, and due to construction site conditions. Another limitation that affected the research study at the time of selection of tools was that the selected tool was required to have a unique sound when being used, in order for the designed model to identify it through audio recognition. Currently the Microsoft Kinect device, the adopted data acquiring device in the research, supports only a limited distance in camera depth and covers a limited field of view (i.e. horizontal and vertical) from the RGB. This limitation severely affected the study by making coverage of objects spread out on an indoor job site more difficult.

It is important to include the assumptions made in the research as part of the limitations. We assumed that a properly colour-coded construction hardhat should be worn by all site personnel when they work on site. As a result, the system will fail if someone works on site without a hardhat or wearing an undisclosed hardhat throughout the period. However, the system will not fail for occasional hardhat offs after the initial recognition has been completed.

DOA works well in higher SNR levels. However, in some situations in an actual noisy environment when the SNR is below a specified threshold, the performance of all three methods rapidly deteriorates due to inconsistent or ambiguous estimates. Further, higher noise level (lower SNR) will affect not only the DOA but also the accuracy of tool sound classifiers by distorting the audio signal with its signature features.

Additionally, a DOA model provides only one output for a given audio frame. As a result of that, the system can recognize only one sound event and its direction. In short, the system will fail if two tool sounds occur in the same 0.25ms time frame from different directions.

7.6 Future Research and Recommendations

There is a high potential to conduct further research on various aspects of the proposed automated worker tool time and performance measuring system, because the limited scope of the research conducted in a limited time period could not assess a wider application of proposed measures.

7.6.1 Multiple Kinect monitoring system

Before considering system upgrades it is important to note that as we have clearly indicated with proven results, the proposed system works well in the Kinect physical distance range. However, a typical construction site spreads over a vast range that a single Kinect camera will not be able to cover alone. A multiple Kinect monitoring system would be an ideal solution to overcome this issue. But we should also consider that the computation cost and memory usage, as well as maintaining synchronization between cameras, is of critical importance in order to ensure accuracy. Certainly, multi-camera calibration and an optimum camera positioning system have to be developed in order to network the devices and work in a common area. This type of positioning model will minimize the usage of devices in a field, and as a result of that, computational cost and memory requirement can be reduced.

An equally significant aspect of networking devices is adding another dimension for localizing acoustic sound sources. Having more than one location information for a single sound source from multiple devices will increase the accuracy and robustness of the output.

Another significant factor in a Kinect network is that 3D reconstruction of the construction environment can be created with a larger point cloud dataset. These as-built 3D models can be used to measure the progress of the project and effectively combine with worker tool time information to provide productivity related data.

7.6.2 Expansion of tool sound database

Currently, the proposed system tracks only four tool sounds and there is a higher potential of applying the same concept for more tools used in a typical jobsite. For instance, sounds from a pneumatic nail gun and power screwdrivers could potentially be identified by an audio processing system. At the same time, future work should be focused on the robust time delay estimation with low SNR in order to work in noisy environments.

Another significant factor in improving the audio system is a manually operated learning model that can be implemented in order to strengthen the statistical model by identifying false negative (FN) predictions.

7.6.3 360 view

The Google street view covers 360 degree panoramic view, and a Light Detection and ranging system (LiDAR) mounted on top of the Google car generates a point cloud that gives the car a 360-degree view (Bilan, 2013; Spring, 2007; Vanderbilt, 2012). It would be interesting research to adopt the combined technology from the Google car and Google street view, and develop a 360 degree view (i.e. RGB and depth) of the construction site for both outdoor and indoor environments. This single unit would cover all the blind spots in the site and have the potential to generate a diverse information dataset of workers and progress around the unit.

REFERENCES

Arif, O., & Vela, P. A. (2009). *Kernel covariance image region description for object tracking.* Paper presented at the Image Processing (ICIP), 2009 16th IEEE International Conference on.

Bedard, S., Champagne, B., & Stephenne, A. (1994). *Effects of room reverberation on time-delay estimation performance.* Paper presented at the Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on.

Behzadan, A. H., & Kamat, V. R. (2007). Georeferenced registration of construction graphics in mobile outdoor augmented reality. *Journal of Computing in Civil Engineering, 21*(4), 247-258.

Benavidez, P., & Jamshidi, M. (2011). *Mobile Robot Navigation and Target Tracking System.* Paper presented at the 6th International Conference on System of Systems Engineering: SoSE in Cloud Computing, Smart Grid, and Cyber Security, SoSE 2011, Albuquerque, NM, United states.

Bilan, S. J. (2013). Driving With the Google Car. *Scholastic Action, 36*, 24-24.

Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision, 74*(1), 17-31.

Borcherding, J. D. (1976). Improving productivity in industrial construction. *American Society of Civil Engineers, Journal of the Construction Division, 102*(4), 599-614.

Borcherding, J. D., & Garner, D. F. (1981). Motivation and productivity on large jobs. *Journal of the Construction Division, 107*(3), 443-453.

Bouguet, J. (2010). Camera calibration toolbox for Matlab. Retrieved from http://www.vision.caltech.edu/bouguetj/calib_doc/

Brilakis, I., Park, M., & Jog, G. (2011). Automated vision tracking of project related entities. *Advanced Engineering Informatics, 25*(4), 713-724. doi: http://dx.doi.org/10.1016/j.aei.2011.01.003

Brown, D. (1971). Close range camera calibration. *Journal of Photogrammetric Engineering & Remote Sensing, 37*(8), 855-866.

Burrell, J., & Gay, G. K. (2001). *Collectively defining context in a mobile, networked computing environment.* Paper presented at the Proceedings of CHI Seattle, WA.

Cheng, T., Venugopal, M., Teizer, J., & Vela, P. A. (2011). Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments. *Automation in Construction, 20*(8), 1173-1184.

Chikamasa, T. (2012). Simulink Support for Kinect. *MATLAB Central-File Exchange*. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/32318-simulink-support-for-kinect

Cho, Y. K., Youn, J. H., & Martinez, D. (2010). Error modeling for an untethered ultra-wideband system for construction indoor asset tracking. *Automation in Construction, 19*(1), 43-54.

Choy, E., & Ruwanpura, J. Y. (2006). Predicting construction productivity using situation-based simulation models. *Canadian Journal of Civil Engineering, 33*(12), 1585-1600.

Chu, S., Narayanan, S., & Kuo, C. C. J. (2009). Environmental Sound Recognition With Time Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1142-1158. doi: 10.1109/tasl.2009.2017438

CII. (2010). Guide to activity analysis. *Construction Industry Institute's Implementation Resource 252-2a*, 1-76. Retrieved from https://www.construction-institute.org/scriptcontent/more/ir252_2a_more.cfm

Clavel, C., Ehrette, T., & Richard, G. (2005). *Events Detection for an Audio-Based Surveillance System.* Paper presented at the Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.

Cordova, F., & Brilakis, I. (2008). *On-site 3D vision tracking of construction personnel.* Paper presented at the 16th Annual Conference of the International Group for Lean Construction, IGLC16, Manchester, United kingdom.

Cronk, S., Fraser, C., & Hanley, H. (2006). Automated metric calibration of colour digital cameras. *The Photogrammetric Record, 21*(116), 355-372. doi: 10.1111/j.1477-9730.2006.00380.x

Dhull, S., Arya, S., & Sahu, O. P. (2010). Comparison of Time-Delay Estimation Techniques in Acoustic Environment. *International Journal of Computer Application, 8*(9), 29-31.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). *Behavior Recognition via Sparse Spatio-Temporal Features.* Paper presented at the Proceedings - 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, VS-PETS, Beijing, China.

Dozzi, S. P., & AbouRizk, S. M. (1993). Productivity in Construction *Institute for Research in Construction, National Research Council*. Ottawa, ON, Canada.

Ekahau. (2012). Wi-Fi Based Asset Management and People Tracking Solution For Hospitals and Other Enterprises Retrieved 12/10/2012, 2012, from http://www.ekahau.com

Elkamchouchi, H., & Mofeed, M. A. E. (2005). *Direction-of-arrival methods (DOA) and time difference of arrival (TDOA) position location technique.* Paper presented at the Radio Science Conference, 2005. NRSC 2005. Proceedings of the Twenty-Second National.

ENR. (2011). Don't blame the workers. *Engineering News-Record, Bruce Buckley* Retrieved from http://bit.ly/lChX6r

Escorcia, V., Dávila, M. A., Golparvar-Fard, M., & Niebles, J. C. (2012). *Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras.* Paper presented at the Proceedings of the 2012 Construction Research Congress, West Lafayette, IN, United states.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi: http://dx.doi.org/10.1016/j.patrec.2005.10.010

Ganesan, S. (1984). Construction productivity. *Habitat International, 8*(3–4), 29-42. doi: http://dx.doi.org/10.1016/0197-3975(84)90041-9

Gilbreth, F. B., & Kent, R. T. (1911). Motion study: A method for increasing the efficiency of the workman. *Journal of the Franklin Institute, 171*(4), 429. doi: http://dx.doi.org/10.1016/S0016-0032(11)90182-3

Gong, J., & Caldas, C. H. (2010). Computer vision-based video interpretation model for automated productivity analysis of construction operations. *Journal of Computing in Civil Engineering, 24*(3), 252-263.

Goodrum, P. M., McLaren, M. A., & Durfee, A. (2006). The application of active radio frequency identification technology for tool tracking on construction job sites. *Automation in Construction, 15*(3), 292-302.

Gouett, M. C., Haas, C. T., Goodrum, P. M., & Caldas, C. H. (2011). Activity Analysis for Direct-Work Rate Improvement in Construction. [Case Study]. *Journal of Construction Engineering & Management, 137*(12), 1117-1124. doi: 10.1061/(asce)co.1943-7862.0000375

Grau, D., Caldas, C. H., Haas, C. T., Goodrum, P. M., & Jie, G. (2009). Assessing the impact of materials tracking technologies on construction craft productivity. *Automation in Construction, 18*(7), 903-911.

Habib, A. F. (2008a). *ENGO 431: Chapter 7, Mathematical Model*. Principles of Photogrammetry. Geomatics Engineering. University of Calgary.

Habib, A. F. (2008b). *Single photo resection*. ENGO 431 – Principles of Photogrammetry. Geomatics Engineering. University of Calgary.

Habib, A. F., Pullivelli, A. M., & Morgan, M. F. (2005). Quantitative measures for the evaluation of camera stability. *Optical Engineering, 44*(3), 33605-33601-33608.

He, Z., Zhang, J., & Zeng, W. (2011). *Design and implement of a band-pass FIR filter based on FPGA in multi-channel data acquisition system.* Paper presented at the

Electronic Measurement & Instruments (ICEMI), 2011 10th International Conference on.

Hewage, K. N., & Ruwanpura, J. Y. (2006). Carpentry workers issues and efficiencies related to construction productivity in commercial construction projects in Alberta. *Canadian Journal of Civil Engineering, 33*(8), 1075-1089.

Heydarian, A., Golparvar-Fard, M., & Carlos Niebles, J. (2012). *Automated visual recognition of construction equipment actions using spatio-temporal features and multiple binary Support Vector Machines.* Paper presented at the Construction Research Congress 2012, West Lafayette, IN, United states.

Hightower, J., & Borriello, G. (2001). Location systems for ubiquitous computing. *Computer (USA), 34*(8), 57-66.

Hong, Q., Zhaonan, G., & Xiangli, Z. (2010, 29-31 Oct. 2010). *The Design of FIR Band-Pass Filter with Improved Distributed Algorithm Based on FPGA.* Paper presented at the Multimedia Technology (ICMT), 2010 International Conference on.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* (2nd edn. ed.). New York: John Wiley & Sons, Ltd.

Huadong, W., Siegel, M., & Khosla, P. (1999). Vehicle sound signature recognition by frequency vector principal component analysis. *Instrumentation and Measurement, IEEE Transactions on, 48*(5), 1005-1009. doi: 10.1109/19.799662

Ianniello, J. (1982). Time delay estimation via cross-correlation in the presence of large estimation errors. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 30*(6), 998-1003. doi: 10.1109/tassp.1982.1163992

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., . . . Fitzgibbon, A. (2011). *KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera.* Paper presented at the UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, United states.

Jaselskis, E. J., & El-Misalami, T. (2003). Implementing radio frequency identification in the construction process. *Journal of Construction Engineering and Management, 129*(6), 680-688.

Jenkins, J. L., & Orth, D. L. (2004). Productivity Improvement Through Work Sampling. *Cost Engineering, 46*(3), 27-32.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). *A Biologically Inspired System for Action Recognition.* Paper presented at the 2007 11th IEEE International Conference on Computer Vision, Piscataway, NJ, USA.

Jingdong, C., Yiteng, H., & Benesty, J. (2005). *A comparative study on time delay estimation in reverberant and noisy environments.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on.

Katz, I., Saidi, K., & Lytle, A. (2008). *The role of camera networks in construction automation.* Paper presented at the ISARC 2008 - Proceedings from the 25th

International Symposium on Automation and Robotics in Construction, Vilnius, Lithuania.

Khoury, H. M., & Kamat, V. R. (2009a). Evaluation of position tracking technologies for user localization in indoor construction environments. *Automation in Construction, 18*(4), 444-457.

Khoury, H. M., & Kamat, V. R. (2009b). High-precision identification of contextual information in location-aware engineering applications. *Advanced Engineering Informatics, 23*(4), 483-496.

Khoury, H. M., & Kamat, V. R. (2009c). *Indoor User Localization for Rapid Information Access and Retrieval on Construction Sites.* Paper presented at the Proceedings of the 15th Annual Workshop of the European Group for Intelligent Computing in Engineering (EG-ICE), European Group for Intelligent Computing in Engineering, Plymouth, UK.

Kiziltas, S., Akinci, B., Ergen, E., Pingbo, T., & Gordon, C. (2008). Technological assessment and process implications of field data capture technologies for construction and facility/infrastructure management. *Electronic Journal of Information Technology in Construction, 13*, 134-154.

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on, 24*(4), 320-327. doi: 10.1109/tassp.1976.1162830

Lawley, D. N., & Maxwell, A. E. (1962). Factor Analysis as a Statistical Method. *Journal of the Royal Statistical Society. Series D (The Statistician), 12*(3), 209-229. doi: 10.2307/2986915

Liberda, M., Ruwanpura, J., & Jergeas, G. (2003). *Construction Productivity Improvement: A Study of Human, Management and External Issues.* Paper presented at the Construction Research Congress, Winds of Change: Integration and Innovation in Construction, Proceedings of the Congress, March 19, 2003 - March 21, 2003, Honolulu, HI., United states.

Lim, & Alum, J. (1995). Construction productivity: Issues encountered by contractors in Singapore. *International Journal of Project Management, 13*(1), 51-58. doi: http://dx.doi.org/10.1016/0263-7863(95)95704-H

Lim, Choi, B. S., & Lee, J. M. (2006). *An Efficient Localization Algorithm for Mobile Robots based on RFID System.* Paper presented at the SICE-ICASE, International Joint Conference, Busan, South Korea.

Lindenmeyer, C. R. (2001). How to design and conduct a computer aided work sampling with Microsoft excel. Retrieved from http://www.c-four.com/docs/How%20To%20Do%20a%20CAWSE%20Study.pdf

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

McTague, B., & Jergeas, G. (2002). Productivity Improvements on Alberta Major Construction Projects *Construction Productivity Improvement Report/Project Evaluation Tool*.

MESA, I. (2013). SwissRanger™ SR4000 Overview Retrieved from http://www.mesa-imaging.ch/prodview4k.php?cat=3D%20Camera

Microsoft. (2012). Kinect for Windows Retrieved 12/10, 2012, from http://www.microsoft.com/en-us/kinectforwindows/

Microsoft. (2013). Kinect for Windows Human Interface Guidelines v1.5.0. *msdn*. Retrieved from http://msdn.microsoft.com/en-us/library/jj663791.aspx

Ming, J., Kot, A. C., & Er, M. H. (1998). *Performance study of time delay estimation in a room environment [microphone arrays]*. Paper presented at the Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on.

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, *78*(3), 691-692. doi: 10.2307/2337038

Noor, I. (1998). Measuring construction labor productivity by daily visits. *AACE International Transactions*, PR16-PR21.

Peddi, A., Huan, L., Bai, Y., & Kim, S. (2009). *Development of human pose analyzing algorithms for the determination of construction productivity in real-time.* Paper presented at the Building a Sustainable Future - Proceedings of the 2009 Construction Research Congress, Seattle, WA, United states.

Peng, C., So, T., Stage, F., & St. John, E. (2002). The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999. *Research in Higher Education*, *43*(3), 259-293. doi: 10.1023/a:1014858517172

Ranasinghe, U., Ruwanpura, J. Y., & Liu, X. (2012). Streamlining the Construction Productivity Improvement Process with the Proposed Role of a Construction Productivity Improvement Officer. *Journal of Construction Engineering and Management, 138*(6), 697-706.

Ray, S. J., & Teizer, J. (2012). *Real-time posture analysis of construction workers for ergonomics training.* Paper presented at the Construction Research Congress 2012, West Lafayette, IN, United states.

Remondino, F., & Fraser, C. (2006). Digital camera calibration methods: considerations and comparisons. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 36*(5), 266-272.

Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). *Action MACH: A spatio-temporal maximum average correlation height filter for action recognition.* Paper presented at the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Anchorage, AK, United states.

Rouas, J. L., Louradour, J., & Ambellouis, S. (2006). *Audio Events Detection in Public Transport Vehicle.* Paper presented at the Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE.

Sattineni, A., & Azhar, S. (2010). *Techniques for tracking RFID tags in a BIM model.* Paper presented at the 2010 - 27th International Symposium on Automation and Robotics in Construction, ISARC 2010, Bratislava, Slovakia.

Schilit, B., Adams, N., & Want, R. (1995). *Context-aware computing applications.* Paper presented at the Workshop on Mobile Computing Systems and Applications, Los Alamitos, CA, USA.

Schneider, M. (2003). *Radio frequency identification (RFID) technology and its applications in the commercial construction industry.* MSc, University of Kentucky, USA.

Shoelson, B. (2008). Detecting Circles in an Image. *Mathworks File Exchange*. Retrieved from http://blogs.mathworks.com/pick/2008/05/23/detecting-circles-in-an-image/

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., . . . Blake, A. (2011). *Real-Time Human Pose Recognition in Parts from a Single Depth Image.* Paper presented at the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, NJ, USA.

Silva, L., & Ruwanpura, J. Y. (2011). *Improved crew productivity quality, communication safety and quality via virtual supervision.* Paper presented at the Modern methods and advances in structural engineering and construction, IESEC-6 Zurich.

Smisek, J., Jancosek, M., & Pajdla, T. (2011). *3D with Kinect.* Paper presented at the Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on.

Smith, A. G. (1987). *INCREASING ONSITE PRODUCTION.* Paper presented at the 31st Annual Meeting of the American Association of Cost Engineers., Atlanta, GA, USA.

Spring, T. (2007). How the Web Works: Google's Street-Scene Machine. *PC World, 25,* 131-131.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). *Human Activity Detection from RGBD Images*. Paper presented at the AAAI Workshop - Technical Report, San Francisco, CA, United states.

Teizer, J., Caldas, C. H., & Haas, C. T. (2007). Real-time three-dimensional occupancy grid modeling for the detection and tracking of construction resources. *Journal of Construction Engineering and Management, 133*(11), 880-888.

Teizer, J., Lao, D., & Sofer, M. (2007). *Rapid automated monitoring of construction site activities using ultra-wideband.* Paper presented at the 24th International Symposium on Automation and Robotics in Construction, ISARC 2007, Kochi, India.

Teizer, J., Venugopal, M., & Walia, A. (2008). Ultrawideband for Automated Real-Time Three-Dimensional Location Sensing for Workforce, Equipment, and Material Positioning and Tracking. *Transportation Research Record*, 56-64.

Tucker, R. L., Rogge, D. F., Hayes, W. R., & Hendrickson, F. P. (1982). IMPLEMENTATION OF FOREMAN-DELAY SURVEYS. *Journal of the Construction Division, 108*(CO4), 577-591.

Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., & Sarti, A. (2007). *Scream and gunshot detection and localization for audio-surveillance systems.* Paper presented at the Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on.

Vanderbilt, T. (2012). Let the Robot Drive: The Autonomous Car of the Future Is Here. *Wired*, 5.

Varma, K. (2002). *Time-Delay-Estimate Based Direction-of-Arrival Estimation for Speech in Reverberant Environments.* Master of Science, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Varma, K., Ikuma, T., & Beex, A. A. (2002). *Robust TDE-based DOA estimation for compact audio arrays.* Paper presented at the Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002.

Wang, J., & Liu, J. C. L. (2005). *Interference minimization and uplink relaying for a 3G/WLAN network.* Paper presented at the Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS InternationalWorkshop on Self-Assembling Wireless Networks (SNPD/SAWN), Towson, Baltimore.

Weerasinghe, I. P., Ruwanpura, J. Y., Boyd, J. E., & Habib, A. F. (2012). *Application of Microsoft Kinect sensor for tracking construction workers.* Paper presented at the Construction Research Congress 2012, West Lafayette, IN, United states.

Whiteside, J. D. (2006). Construction Productivity. *AACE International Transactions*, ES81-ES88.

Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., & Rehg, J. M. (2007). *A scalable approach to activity recognition based on object use.* Paper presented at the 2007 11th IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil.

Wu, W., Yang, H., Chew, D. A. S., Yang, S., Gibb, A. G. F., & Li, Q. (2010). Towards an autonomous real-time tracking system of near-miss accidents on construction sites. *Automation in Construction, 19*(2), 134-141.

Xia, L., Chen, C., & Aggarwal, J. K. (2011). *Human Detection Using Depth Information by Kinect.* Paper presented at the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2011), Los Alamitos, CA, USA.

Yang, J., Arif, O., Vela, P. A., Teizer, J., & Shi, Z. (2010). Tracking multiple workers on construction sites using video cameras. *Advanced Engineering Informatics, 24*(4), 428-434.

Zeng, H. (2010). *A 3D coordinate transformation algorithm.* Paper presented at the 2nd Conference on Environmental Science and Information Application Technology, ESIAT 2010, Wuhan, China.

Zhang, Y., & Abdulla, W. H. (2005). A Comparative Study of Time-Delay Estimation Techniques Using Microphone Arrays. *School of Engineering Report No. 619*. Retrieved from http://homepages.engineering.auckland.ac.nz/~wabd002/Technical%20Reports/Technical%20Report%20619_Yushi.pdf

Zhu, Y. J., Pan, W. Q., & Luo, Y. L. (2010). 3D measurement system based on computer-generated gratings. In J. Tan & X. Wen (Eds.), *6th International Symposium on Precision Engineering Measurements and Instrumentation* (Vol. 7544). Bellingham: Spie-Int Soc Optical Engineering.

Appendix-A: Calculations of Single Photo Resection (SPR)

This document with regards to the section 4.5.7 and describes the results obtained from the SPR calculation. Sixteen number of marked ground control points were captured in the test field from the Kinect RGB camera as shown in the following figure. The centroids of the circular shapes in the have been extracted from the Hough transform algorithm. The origin of the building coordinate system is the centroid of the circle in the left bottom corner in the image and the orientation of coordinates are shown in the following figure.

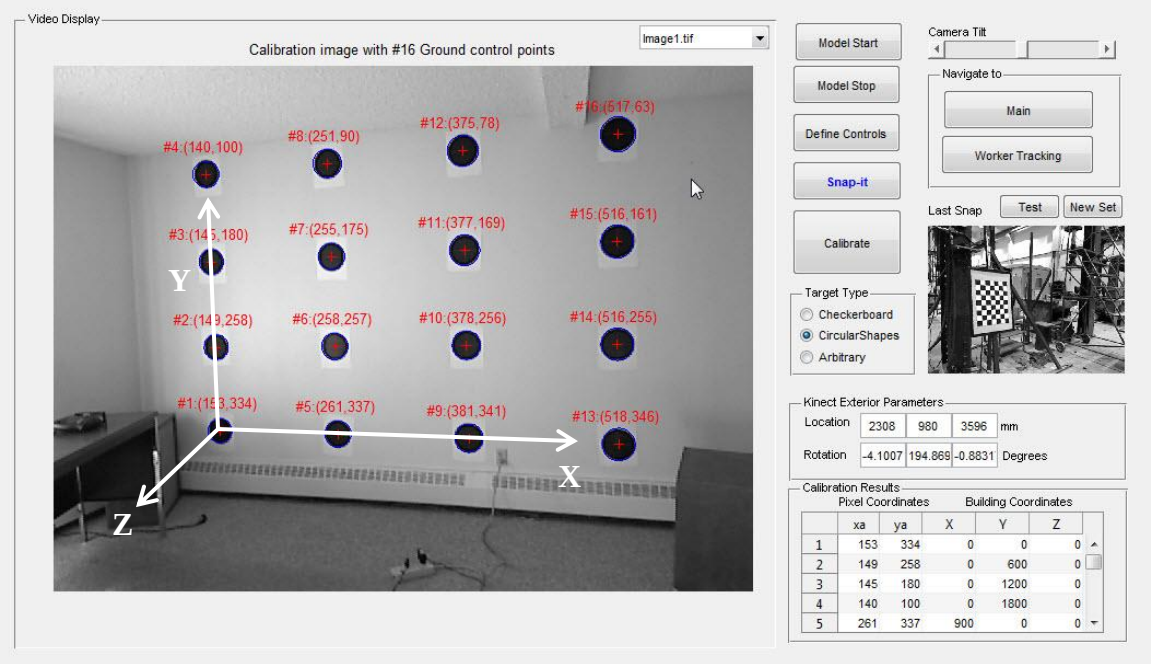


Figure: SPR Application

Following table shows the measured pixel coordinates and image coordinates (in mm) of 16 ground control points displayed in the above image. Further 3D coordinates of ground control points w.r.t. the building coordinate system are also listed in the following table.

Measured pixel coordinates, image coordinates and ground coordinates

Point #	Pixel Coordinates		Ground Coordinates (mm)			Image Coordinates (mm)	
	x	y	X	Y	Z	x	y
1	152.8	334.1	0	0	0	-0.925	-0.521
2	149.5	257.6	0	600	0	-0.944	-0.097
3	145.1	179.5	0	1200	0	-0.968	0.335
4	140.2	99.6	0	1800	0	-0.995	0.777
5	260.8	336.8	900	0	0	-0.328	-0.535
6	257.8	256.8	900	600	0	-0.344	-0.093
7	254.9	174.8	900	1200	0	-0.360	0.361
8	251.2	89.7	900	1800	0	-0.381	0.832
9	380.6	340.5	1800	0	0	0.335	-0.556
10	378.3	255.6	1800	600	0	0.323	-0.087
11	376.5	169.0	1800	1200	0	0.313	0.393
12	375.0	78.2	1800	1800	0	0.304	0.896
13	517.7	346.3	2700	0	0	1.094	-0.588
14	516.5	254.7	2700	600	0	1.087	-0.081
15	516.4	160.8	2700	1200	0	1.087	0.438
16	516.9	62.9	2700	1800	0	1.090	0.980

Initial approximation

In order to find initial approximations of EOPs we assumed a vertical photograph (i.e. ω, φ=0). Estimate the $\theta = \tan^{-1}(\frac{b}{a})$ using following equation.

$$X = a_0 + ax - by$$
$$Y = b_0 + bx + ay$$

Where, (x, y) are the image coordinates and the (X, Y) are corresponding ground coordinates. The calculated θ is used as the initial approximation of the angle κ. Hence rotation matrix can be calculated. The initial approximations of Xo, Yo are taken as the a_0, and b_0 values of the above equation, and for Zo, average of Z of ground control points measured from the Kinect depth map is used .

Initial approximations to unknown parameters

Unknown	Initial Estimate
X (mm)	1424.19
Y (mm)	794.22
Z (mm)	4000.00
ω (rad)	0.00
ϕ (rad)	0.00
κ (rad)	0.73

Modified exterior orientation parameters after each iteration

Unknown	Iter1	Iter2	Iter3	Iter4	Iter5	Iter6	Iter7	Iter8
X (mm)	2523.84	2268.50	2310.09	2310.71	2309.82	2309.29	2308.93	2308.67
Y (mm)	361.06	1624.57	1104.89	973.41	980.72	980.32	980.34	980.33
Z (mm)	4907.51	3219.99	3604.84	3597.75	3595.29	3595.48	3595.61	3595.71
ω (rad)	0.0697	-0.1567	-0.0910	-0.0698	-0.0717	-0.0716	-0.0716	-0.0716
ϕ (rad)	0.2745	0.2672	0.2612	0.2603	0.2601	0.2600	0.2599	0.2598
κ (rad)	0.0001	-0.0199	-0.0132	-0.0154	-0.0154	-0.0154	-0.0154	-0.0154

Unknown	Iter9	Iter10	Iter11	Iter12	Iter13	Iter14	Iter15	Iter16
X (mm)	2308.48	2308.35	2308.26	2308.19	2308.15	2308.11	2308.09	2308.07
Y (mm)	980.33	980.33	980.33	980.33	980.32	980.32	980.32	980.32
Z (mm)	3595.77	3595.82	3595.85	3595.88	3595.90	3595.91	3595.92	3595.92
ω (rad)	-0.0716	-0.0716	-0.0716	-0.0716	-0.0716	-0.0716	-0.0716	-0.0716
ϕ (rad)	0.2598	0.2597	0.2597	0.2597	0.2597	0.2597	0.2597	0.2597
κ (rad)	-0.0154	-0.0154	-0.0154	-0.0154	-0.0154	-0.0154	-0.0154	-0.0154

Final adjusted values for the exterior orientation parameters

Unknown	Final Adjusted Values
X (mm)	2308
Y (mm)	980
Z (mm)	3596
ω (rad)	-0.0716
ϕ (rad)	3.4013
κ(rad)	-0.0154

Estimate of the variance component

Iteration	Variance	Iteration	Variance	Iteration	Variance	Iteration	Variance
Iter1	9.11E-04	iter5	2.53E-05	Iter9	2.55E-05	Iter13	2.56E-05
Iter2	3.12E-05	iter6	2.54E-05	Iter10	2.55E-05	Iter14	2.56E-05
Iter3	1.66E-05	iter7	2.54E-05	Iter11	2.56E-05	Iter15	2.56E-05
Iter4	2.40E-05	iter8	2.55E-05	Iter12	2.56E-05	Iter16	2.56E-05

Posterior variance-covariance (dispersion) matrix of the parameters: last iteration

	X (mm)	Y(mm)	Z(mm)	ω(rad)	ϕ(rad)	κ(rad)
X (mm)	408.404	-47.410	-168.549	0.010	0.110	-0.002
Y(mm)	-47.410	718.573	4.320	-0.193	-0.013	0.014
Z(mm)	-168.549	4.320	98.469	5.92E-05	-0.047	0.001
ω(rad)	0.010	-0.193	5.92E-05	5.20E-05	2.86E-06	-3.79E-06
ϕ(rad)	0.110	-0.013	-0.047	2.86E-06	2.98E-05	-6.33E-07
κ(rad)	-0.002	0.014	0.001	-3.79E-06	-6.33E-07	2.15E-06

Residuals associated with the image coordinate measurements

XY	Iter1	Iter2	Iter3	Iter5	Iter9	Iter12	Iter14	Iter16
x1	0.0133	-0.0043	-0.0075	-0.0018	-0.0017	-0.0016	-0.0016	-0.0016
y1	0.0092	-0.0016	-0.0066	-0.0078	-0.0077	-0.0077	-0.0077	-0.0077
x2	0.0220	0.0018	-0.0007	0.0033	0.0034	0.0035	0.0035	0.0035
y2	-0.0261	-0.0031	-0.0025	-0.0036	-0.0036	-0.0036	-0.0036	-0.0036
x3	0.0310	0.0013	0.0003	0.0031	0.0032	0.0032	0.0032	0.0032
y3	-0.0198	-0.0037	0.0000	0.0003	0.0003	0.0003	0.0003	0.0003
x4	0.0431	-0.0031	-0.0018	0.0003	0.0004	0.0004	0.0004	0.0004
y4	0.0305	-0.0006	0.0036	0.0062	0.0061	0.0061	0.0061	0.0061
x5	-0.0028	-0.0015	-0.0010	0.0021	0.0021	0.0021	0.0021	0.0021
y5	0.0099	0.0050	0.0019	-0.0016	-0.0016	-0.0016	-0.0016	-0.0016
x6	-0.0052	0.0016	0.0018	0.0041	0.0041	0.0041	0.0041	0.0041
y6	-0.0266	-0.0025	-0.0010	-0.0031	-0.0031	-0.0031	-0.0031	-0.0031
x7	-0.0008	0.0047	0.0053	0.0071	0.0070	0.0070	0.0070	0.0070
y7	-0.0190	-0.0073	-0.0041	-0.0034	-0.0034	-0.0034	-0.0034	-0.0034
x8	0.0046	0.0022	0.0036	0.0054	0.0054	0.0054	0.0054	0.0054
y8	0.0394	-0.0021	-0.0006	0.0039	0.0038	0.0038	0.0038	0.0038
x9	-0.0138	-0.0073	-0.0035	-0.0041	-0.0042	-0.0042	-0.0042	-0.0042
y9	-0.0009	0.0085	0.0073	0.0013	0.0012	0.0012	0.0012	0.0012
x10	-0.0248	-0.0056	-0.0039	-0.0044	-0.0045	-0.0045	-0.0045	-0.0045
y10	-0.0308	-0.0005	0.0017	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014
x11	-0.0276	-0.0020	-0.0019	-0.0022	-0.0022	-0.0023	-0.0023	-0.0023
y11	-0.0176	-0.0083	-0.0064	-0.0050	-0.0050	-0.0050	-0.0050	-0.0050
x12	-0.0224	0.0030	0.0020	0.0022	0.0021	0.0021	0.0021	0.0021
y12	0.0516	-0.0015	-0.0041	0.0028	0.0028	0.0029	0.0029	0.0029
x13	0.0121	-0.0040	0.0017	-0.0039	-0.0040	-0.0039	-0.0039	-0.0039
y13	-0.0285	0.0038	0.0049	-0.0041	-0.0043	-0.0044	-0.0044	-0.0044
x14	-0.0067	-0.0036	-0.0025	-0.0074	-0.0074	-0.0074	-0.0074	-0.0074
y14	-0.0416	0.0001	0.0027	-0.0014	-0.0015	-0.0015	-0.0015	-0.0015
x15	-0.0126	0.0033	0.0000	-0.0041	-0.0041	-0.0041	-0.0041	-0.0041
y15	-0.0091	-0.0008	-0.0014	0.0012	0.0013	0.0013	0.0013	0.0013
x16	-0.0093	0.0134	0.0061	0.0025	0.0025	0.0025	0.0025	0.0025
y16	0.0794	0.0124	0.0033	0.0136	0.0138	0.0139	0.0139	0.0139

Appendix B: Snapshots of the worker tracking GUI

Following figure illustrates a series of snapshots of handling worker occlusion while moving in the work site.