FUZZY QUALITATIVE APPROACH TO ADDRESS UNCERTAINTY IN HUMAN MOTION ANALYSIS

LIM CHERN HONG

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

FUZZY QUALITATIVE APPROACH TO ADDRESS UNCERTAINTY IN HUMAN MOTION ANALYSIS

LIM CHERN HONG

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Lim Chern Hong (I.C./Passport No.: 870403-14-5239)

Registration/Matrix No.: WHA110010

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Fuzzy Qualitative Approach To Address Uncertainty In Human Motion Analysis

Field of Study: Computer Science (Image Processing)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

Human motion analysis is one of the most active researches in computer vision society nowadays due to its wide spectrum of applications. Current researchers have been focused on implementing sophisticated algorithms with the goal to achieve good recognition rate but such work are limited to some constraints or assumptions. As a consequence, these systems are impractical to deploy in real-world environment due to the abounded uncertainties in the human motion analysis pipeline such as human size variation, viewpoint variation, and classification ambiguity. Failing in handling these uncertainties could affect the overall system performance. In this thesis, fuzzy qualitative reasoning is studied to address the above uncertainties.

Human modelling is the enabling step in the human motion analysis system where the identified person from a video camera will be projected and represented in a better model to ease the latter processes such as feature extraction. Improper care on the variation of human size and camera positions from the ground might results in a defect human model such as inconsistent human size, and odd human shape. Such defects will hinder the feature extraction process and the error in this step might be cumulated in the rest of the pipeline and deteriorate the overall system performance. In this thesis, fuzzy qualitative Poisson human model is proposed to generalize the human model in terms of sizes and camera viewpoints.

Besides that, to recognize an action with independent to the human viewpoint is a great challenge in human motion analysis, but remains unsolved due to its inherent difficulty. Most state-of-the-art methods are found to be impractical where multi camera system is required to serve the purpose. In this context, view specific action recognition framework is proposed to capture and construct the view specific action model for the objective to achieve view invariant human action recognition within single camera. In the

framework, a novel human contour namely fuzzy qualitative human contour is proposed for view estimation which helps in the construction of the view specific action model.

Action recognition is the final step in the human motion analysis pipeline where the aim is to infer the action or activity from the video. However, classification ambiguity could abounded in this step such as the confusion in viewpoint, action, and scene context due to some similarity factors. These cases are denoted as non-mutually cases in the thesis as their results could not be fully distinguished from the others. Hence, a crisp or binary classifier may not be so effective to deduce the final output for these cases. As a solution, fuzzy qualitative rank classifier is proposed to model the non-mutually exclusive case in the training step and infer with the multi-label and ranking result. This is intuitively reflecting how human decision is made towards the ambiguous case. In addition, dynamic fuzzy qualitative rank classifier is proposed as the extension to overcome the heuristic method in the learning step.

In summary, the collective impact of the above contributions will constitute to achieve a more practical and feasible framework towards the human motion analysis applications. Particular video surveillance system that ensure the public safety and lead to a better and safer society.

ABSTRAK

Analisa gerakan manusia adalah salah satu daripada penyelidikan-penyelidikan yang aktif dalam cabang visi komputer pada masa kini kerana ia memanfaatkan banyak aplikasi. Penyelidikan terkini lebih cenderung untuk melaksanakan algoritma yang lebih canggih dengan matlamat untuk mencapai kadar pengiktirafan yang tinggi tetapi terhad kepada beberapa kekangan atau andaian. Akibatnya, sistem-sistem ini adalah tidak praktikal untuk dipasang dalam persekitaran dunia yang sebenar disebabkan oleh ketitaktentuan berlaku dalam prosess analisi gerakan manusia seperti variasi saiz manusia, perubahan sudut pandangan, dan kekaburan dalam klasifikasi. Gagal untuk menangani ketidaktentuan ini boleh menjejaskan prestasi sistem secara keseluruhan. Penaakulan kualitatif kabur dikaji untuk menangani ketidaktentuan di atas.

Pemodelan Manusia merupakan langkah yang pertama dalam analisa gerakan manusia. Manusia yang diambil daripada kamera video akan diunjurkan dan diwakili dengan model yang lebih baik untuk memudahkan proses kemudian seperti pengekstrakan ciri-ciri pergerakan. Perlaksanaan yang tidak betul terhadap variasi dalam saiz dan sudut pandangan kamera dari tanah mungkin menyebabkan kecacatan dalam model manusia seperti saiz manusia yang tidak konsisten, dan bentuk manusia yang ganjil. Kecacatan itu akan menghalang proses pengekstrakan ciri-ciri dan kesilapan dalam langkah ini mungkin menjejaskan prestasi sistem secara keseluruhan. Dalam tesis ini, “Fuzzy Qualitative Poisson Human Model” dicadangkan untuk mengumumkan model manusia dari segi saiz dan sudut pandangan kamera.

Selain itu, untuk mengiktiraf tindakan dengan bebas kepada pandangan manusia adalah satu cabaran yang besar dalam analisis pergerakan manusia, tetapi tetap tidak dapat diselesaikan kerana kesukaran yang wujud itu. Kebanyakan negeri-of-the-art kaedah yang didapati tidak praktikal di mana sistem multi kamera dikehendaki berkhidmat mak-

sud itu. Dalam konteks ini, *view specific action recognition framework* dicadangkan untuk mencapai prestasi analisis pergerakan manusia yang tidak tersekat dengan perubahan sudut pandangan dengan menggunakan kamera tunggal sahaja. Dalam rangka kerja ini, kontur manusia iaitu "Fuzzy Qualitative Human Contour" dicadangkan untuk anggaran pandangan yang akan membantu dalam pembinaan model tindakan tertentu.

Pengelasan tindakan manusia adalah langkah terakhir dalam analisis pergerakan manusia di mana tujuannya adalah untuk membuat kesimpulan tentang tindakan atau aktiviti daripada video. Walaubagaimanapun, kekaburan didapati berlaku dalam langkah ini seperti kekeliruan dalam sudut pandangan, tindakan, dan konteks pemandangan. Ini disebabkan oleh beberapa faktor pengeliruan. Kes-kes ini ditandakan sebagai kes "non-mutually exclusive" dalam tesis ini kerana keputusan mereka tidak dapat dibezakan sepenuhnya daripada yang lain. Oleh itu, pengelas binari mungkin tidak begitu berkesan untuk menyimpulkan kes-kes ini. Sebagai penyelesaian, *fuzzy qualitative rank classifier* dicadangkan untuk model kes tidak saling eksklusif dalam langkah latihan dan membuat kesimpulan dengan *multi-label* dan keputusan ranking yang mencerminkan bagaimana keputusan manusia dibuat. Di samping itu, *dynamic fuzzy qualitative rank classifier* dicadangkan sebagai lanjutan untuk mengatasi kaedah heuristik dalam langkah pembelajaran.

Ringkasannya, kesan kolektif sumbangan di atas akan mencapai rangka kerja yang lebih praktikal dan boleh dilaksanakan terhadap aplikasi analisis pergerakan manusia, khususnya dalam sistem pengawasan video yang memastikan keselamatan awam dan membawa kepada masyarakat persekitaran yang lebih baik dan lebih selamat.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Chee Seng Chan for the continuous support of my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

My sincere thanks also goes to Prof. Honghai Liu for offering me the visiting research opportunities in the Intelligent Systems and Robotics Group, University of Portsmouth. It was a great time I spent in there together with Dr. Zhaojie Ju, Dr. Alexander Kadyrov, and Dr. Hongyi Li who have leading me working on diverse exciting projects.

I thank my fellow labmates Dr. Vembarasan Vaitheeswaran, Anhar Risnumawan, Ekta Vats, Wai Lam Hoo, Chee Kau Lim, Mei Kuan Lim, Sze Ling Tang, and Ven Jyn Kok for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the past times.

An honorable mention goes to my families and friends for their understandings and supports on me in completing this project. Without helps of the particular that mentioned above, I would face many difficulties while doing this project.

Last but not least I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

2D-FQstate	Two Dimensional Fuzzy Qualitative State.
DFQRC	Dynamic Fuzzy Qualitative Rank Classifier.
FQ-PHM	Fuzzy Qualitative Poisson-normalized Human Model.
FQHC	Fuzzy Qualitative Human Contour.
FQP	Fuzzy Qualitative Partition.
FQRC	Fuzzy Qualitative Rank Classifier.
FQS	Fuzzy Quantity Space.
FQTM	Fuzzy Qualitative Trained Model.
HiL	High-level.
HMA	Human Motion Analysis.
HOG	Histograms of Oriented Gradients.
KNN	K-nearest Neighbour.
LoL	Low-level.
MiL	Mid-level.
VSAM	View Specific Action Model.

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

HMA has been actively emerging over the years owing to the advancement of video camera technologies and the availability of sophisticated computer vision algorithms in the public domain. It is a popular research since decades due to the high demand in many areas such as surveillance systems (Haering et al., 2008; Hu, Tan, et al., 2004; I. S. Kim et al., 2010; Ko, 2008; Popoola & Wang, 2012), healthcare systems (Anderson et al., 2006; Anderson, Luke, et al., 2009b), human computer interaction (Jaimes & Sebe, 2007), sport analysis (Efros et al., 2003; Loy et al., 2004; Sullivan & Shah, 2008), and others. The objective of the system is to analyse and interpret the human behaviour over time from a video camera. In particular, actions such as fighting and falling can be detected from a video surveillance system. Ideally, the system should trigger an alarm to the respective unit when abnormal behaviour is detected. For instance, as illustrated in Figure 1.1, the Madrid, London and Boston marathon bombing tragedies, happened in 2004, 2005 and 2013, respectively, would not have been worse if an intelligent video surveillance system capable of automatically detecting abnormal human behavior was installed in the public areas.

Figure 1.1: (a) **Madrid train bombings:** On March 11 2004, Madrid commuter rail network was attacked and the explosions killed 191 people, injuring 1,800 others, (b) **London bombings:** July 7 2005 London bombings were a series of coordinated suicide attacks in the central London, which targeted civilians using the public transport system during the morning rush hour, (c) **Boston Marathon bombings:** On April 15 2013, two pressure cooker bombs exploded during the Boston Marathon, killing 3 people and injuring 264. Information source: *http://en.wikipedia.org/*, Image source: *http://images.google.com*.

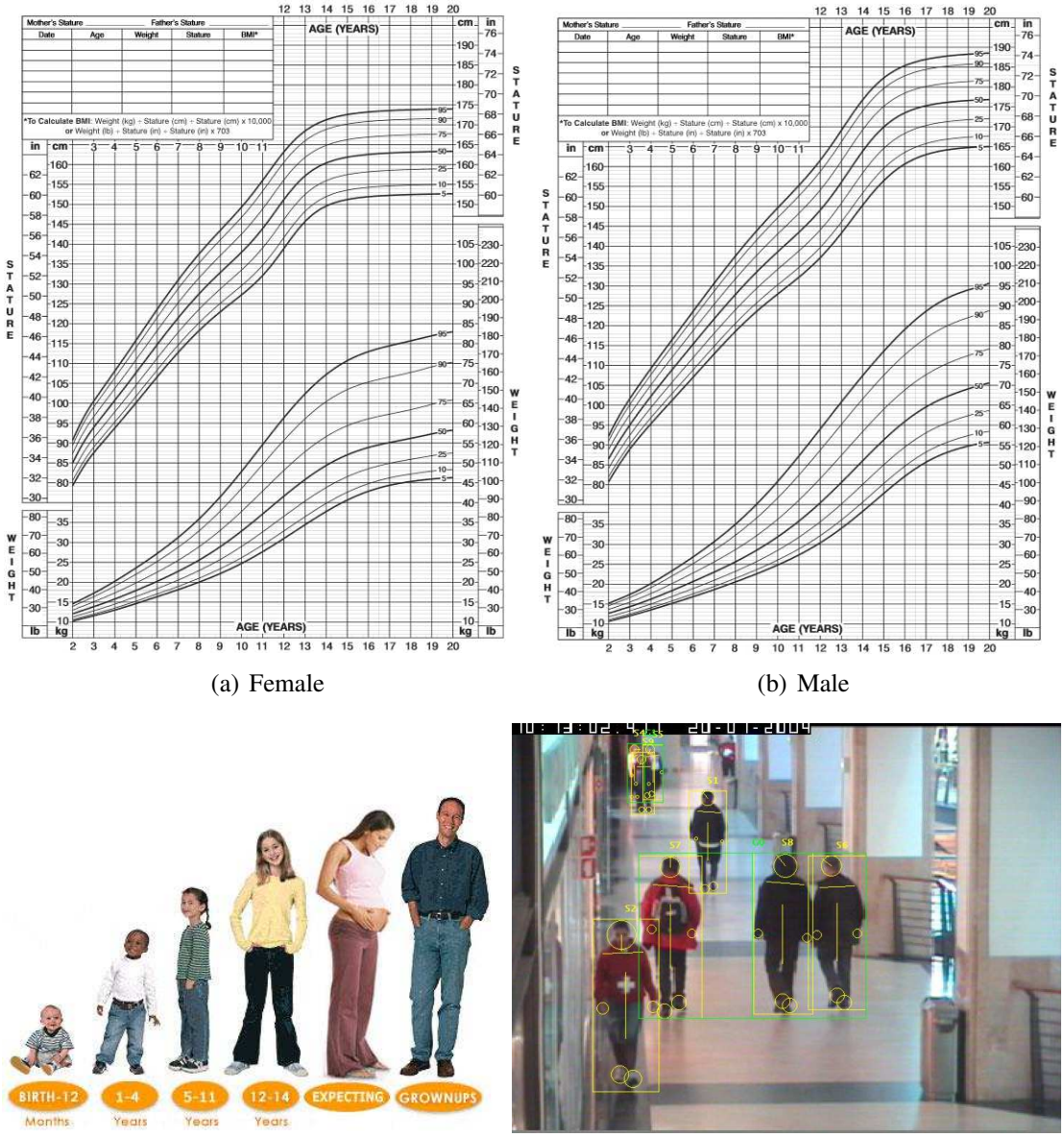
To achieve this, the robustness of the intelligent video surveillance system is a vital concern. False alarm may lead to severe lost between the vendor, user, and the citizen in term of benefit and man power, to have a few. As a result, most of the researchers in this area have tried their best to implement sophisticated HMA algorithms in order to achieve good recognition rate. Because of this, a vast number of research were dedicated to deal with the aforementioned situations which were extensively reviewed in the relevant survey papers (Aggarwal & Ryoo, 2011; Ji & Liu, 2010; Moeslund & Granum, 2001; Poppe, 2010). However most of these solutions are still limited to constraints or assumptions pointed out in Moeslund et al. (2006). For instance, static background, fixed motion pattern, no occlusion and the subject must face the camera at all the time. Unfortunately these systems are impractical to deploy in a real world environment as human in front of the camera can be in various sizes, various viewpoints, and different background conditions. These uncertainties might affect the decision making process (the classification task) and deteriorate the overall system performance.

1.1 Uncertainties in Human Motion Analysis

There are several uncertainties in vision-based HMA that are worth studying in order to enhance the current state-of-the-art solutions. Failure in handling them may lead to poor system performance.

1.1.1 Body Size Variation

The variations in human body size can be categorized into natural and synthetic causes. In specific, natural cause can be explained in term of the biological aspect of a person such as their gender and age (Figure 1.2). By assuming that a person is standing in front of a camera at a fixed position, generally, a male is bigger in size than a female (Figure 1.2(a) and 1.2(b)), and an adult is bigger than a child (Figure 1.2(c)). As for the synthetic



(a) Female (b) Male

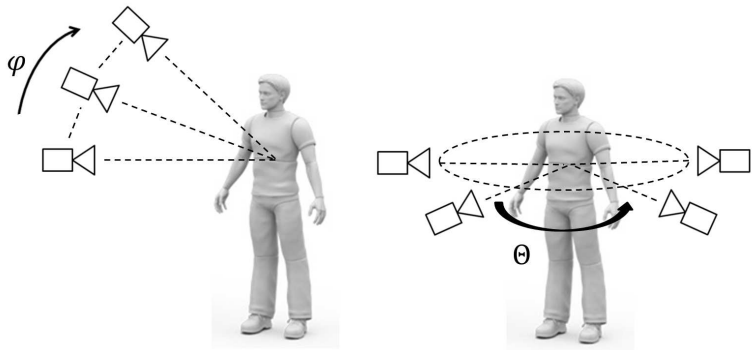
(c) Age (d) Camera

Figure 1.2: 1.2(a) and 1.2(b) represent the weight and height growth chart for female and male. The graphs show that male's growth are bigger in size in term of weight and height compared to that of a female. (Source: Centers for Disease Control and Prevention, *http://www.cdc.gov/growthcharts/reports.htm*). 1.2(c) depicted the size of human according to the ages (Source: *http://images.google.com*). Lastly, 1.2(d) illustrated that different size of human when they are approaching or leaving the camera (Source: CAVIAR Video Sequence Ground Truth, *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/gt.htm*).

reason, it is related to the distance of the person from the camera. A person appears to be bigger in size when he/she approaches the camera, and gradually becomes smaller if he/she moves away from the camera (Figure 1.2(d)).

Such variations are commonly captured in the screen and appear in the video. Without generalization on the size variation, it might affect the projection of the human to build an appropriate human model for further processing. This will limit the system to be size dependent (Cucchiara et al., 2005; Juang & Chang, 2007). Apart from this, naive normalization process such as using bounding box and blob. may result in inappropriate body anatomy representation. In specific, the human body parts could be arbitrary in the projection space like the inconsistency of hand location for different human size, and it could be a tedious job to localize them manually. This problem affects the extraction of robust features which serve as the prerequisite for motion tracking and action recognition. In the end, the cumulated errors may deteriorate the overall system performance.

1.1.2 Viewpoint Variation

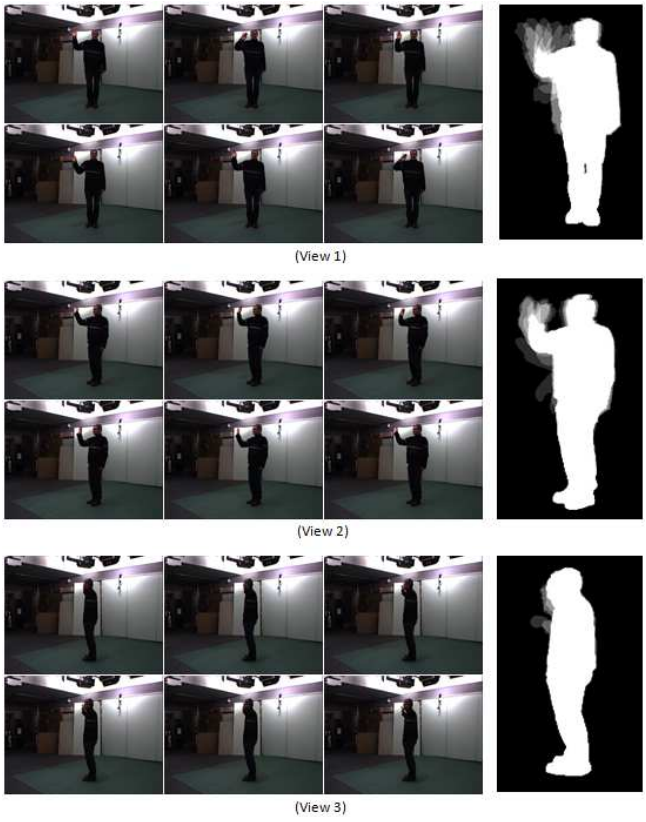


(a) φ angle (b) Θ angle

Figure 1.3: The position of the camera with respect to the viewing object can be parameterized as the combination of two angles (latitude,φ and longitude,Θ).

In a real-world environment, human are not restricted to perform an action at a fixed position, such as always facing the camera. A sophisticated HMA system should be able to take into account the variation in camera viewpoints.

(a) φ angle



(b) Θ angle

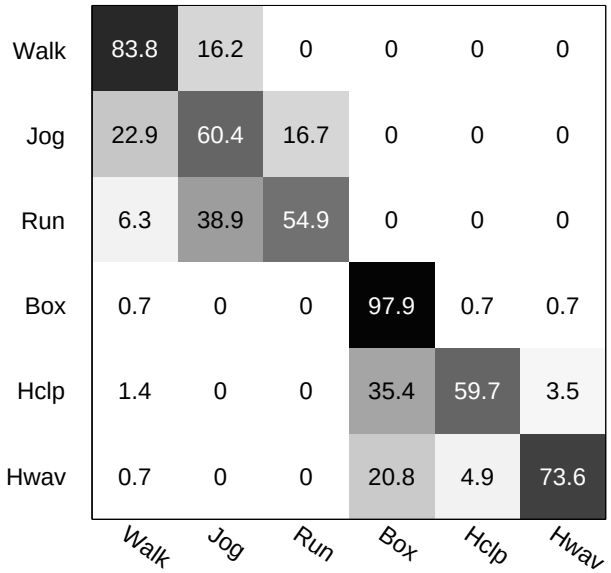
Figure 1.4: The effect of φ and Θ angles in camera viewpoint variation. 1.4(a) shows the examples of a person capture from different camera positions from the ground (φ angle). 1.4(b) illustrates that motion history image of 'wave hand' action from different viewpoints. It can be noticed that the motion patterns are differed from each other in different Θ viewpoint and thus makes view invariant action recognition a daunting task.

In fact, the position of the camera with respect to the viewing object can be parameterized as the combination of two angles, which are the latitude,φ and longitude,Θ (Rogez et al., 2014) as depicted in Figure 1.3. The φ angle represents different camera positions from the ground while the Θ angles denotes different viewpoints captured from the horizontal positions. Example image frames obtained from these types of angles are shown in Figure 1.4. Conventional HMA approaches performed in the way that human is assumed to be always facing the camera which is not practical (Holte et al., 2011; Ji & Liu, 2010). Besides that, the camera may not always set up at a fixed position when installation is done in different environments such as railway station, inside the building or on the street. In specific, they might be set up with different φ angles. If this is not carefully handled in the system, such displacements could affect the overall system performance in interpreting the human action (Lewandowski, Makris, & Nebel, 2010).

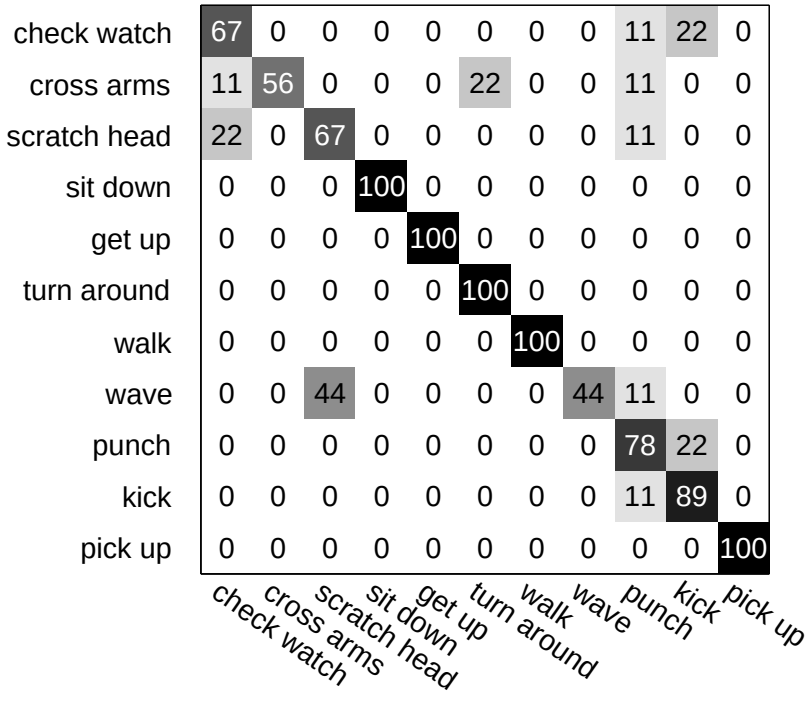
1.1.3 Classification Ambiguity

One of the main reasons that cause the difficulty in achieving good recognition rate in HMA is the ambiguity in classification task. Ambiguity here is defined as the vagueness to accurately recognize an action from the input video due to the similarity factors amongst different actions. Technically, the similarity factors could be in terms of visual-able (e.g. movement of leg) or non visual-able (e.g. space time interest point) features. This is also the reason that researchers in this domain use confusion matrix to tabulate their recognition results as one action might confused with the others due to some the similarity factors. For examples, Figure 1.5 shows the confusion matrices of the action recognition which are adopted from Schuldt et al. (2004) and Y. Yang et al. (2008).

From Figure 1.5, one can notice that there are confusion between “Walk”, “Jog”, and “Run” actions in Figure 1.5(a) as some portions of the recognition rate are scatter among the other two answers instead of the correct one. This is similar to “wave” and “scratch



(a) Schuldt et al. (2004)



(b) Y. Yang et al. (2008)

Figure 1.5: Confusion matrices between actions.

(a) Images from left to right are representing “Walk”, “Jog”, and “Run” actions respectively

(b) Images from left to right are representing “wave” and “scratch head” actions respectively

Figure 1.6: Confusion on the actions. It is noticeable that the characteristics for the actions in 1.6(a) and 1.6(b) are so similar to each others, and thus ambiguity in decision making can happen.

head” actions in Figure 1.5(b). This is a difficult situation as illustrated in Figure 1.6 where even human is having difficulty to correctly tells the correct answer for the action being performed by visual inspection. The reasons are, most of the visible characteristics to differentiate the actions are too similar. Consequently in these scenarios, human tends to provide ambiguous answer instead of a binary (yes / no) one, for examples, “it should be”, “ it maybe”, “I think” where these answers reflected the uncertainty. With this, it implies that the binary classification brutally forced the classification output to belong to solely one class with no tolerance to the uncertainty is not an effective solution.

1.2 Problems Formulation

The aforementioned uncertainties could have attached to the general HMA pipeline (Aggarwal & Ryoo, 2011; Ji & Liu, 2010; Moeslund et al., 2006; L. Wang et al., 2003) as illustrated in Figure 1.7 which will affect the performance of each step. In the worst case,

the cumulated errors will deteriorate the overall system accuracy.

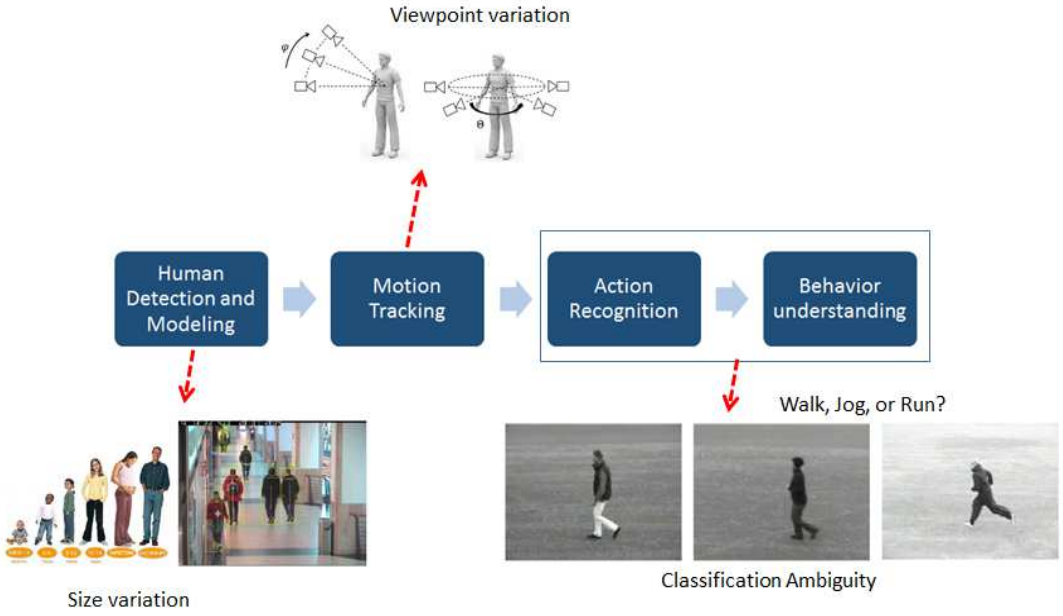


Figure 1.7: Uncertainties that attached to the HMA pipeline.

First, failed in handling the different human size will caused improper projection of human model in the human detection and modeling step. This will affect the feature extraction process due to the incorrect positioned human plane. Most of the researches limit the system to be size dependent (Cucchiara et al., 2005; Juang & Chang, 2007) or extensive training is needed to support the different human size (Dalal & Triggs, 2005) which can be a tedious and time consuming task.

Apart from that, view invariant (viewpoint independent) is the current research trend in vision-based HMA with a few recent published surveys (Holte et al., 2011; Ji & Liu, 2010; Weinland et al., 2011). The aim is to track and model the human motion in a way that no restriction on the human position in front of the camera. To achieve this, many conventional approaches assume that the subjects are always performing an action in a position of frontal-parallel to the camera; and thus builds a 3D action model (Ahmad & Lee, 2006; Anderson, Luke, et al., 2009b; Weinland et al., 2007, 2006) for capturing the actions from different viewpoints or by analysing the multi-view geometry (Ashraf et al., 2013; Yilma & Shah, 2005) as an action template for recognition purpose. This

assumption has few limitations. First of all, in a real-world environment, subjects are not always facing frontal-parallel to the cameras. Secondly, it is unusual to find a multi cameras system in public space that has many overlapping regions. Authorities always tend to cover as much areas as possible with a limited number of cameras for cost effective and limited space. The existence of a region that overlaps with multiple cameras, from different viewing angles and the appearance of a subject performing an action that is frontal-parallel to the cameras are very limited too. Therefore, the approaches that use multiple cameras might not be practical in a real-world environment. Despite of this, the state-of-the-art approach that overcomes the problem by using single camera in view invariant HMA (Lewandowski, Makris, & Nebel, 2010) has a drawback in dealing with the variant in camera positions where extensive training is needed to cope with every camera displacement especially the ϕ angles.

Last but not least, the classification ambiguity could happened in the last step of HMA pipeline which is the classification step to determine the action or behaviour of the human subject(s). Dilemma happened when there are similar actions but different class such as “Walk”, “Jog”, and “Run”. All these actions have the similar characteristic and thus caused the confusion in the classification. Besides that, a surge of interest has sparked in HMA that takes into account the existence of scene context (Ikizler-Cinbis & Sclaroff, 2010; Marszalek et al., 2009). This is because the scenery information is proven to be effective as an extra cue to infer human activity in view independent manner. Nevertheless, by looking into these works (Ikizler-Cinbis & Sclaroff, 2010; Marszalek et al., 2009), the source of uncertainties are still apparent in their approaches that may cause bottleneck in achieving better recognition rate. For instance, the ambiguity in the actions and the scene images might cause confusion in the classification tasks. Due to these ambiguities, conventional inference methods in HMA and even in scene understanding that brutally force the output to be solely belongs to one class may not be a good option.

1.3 Objectives

Based on the introduction and the problem statements, the uncertainties (human size variation, viewpoint variation, and classification ambiguity) exist in HMA and the current state-of-the-art methods are still infeasible to deal with them. The objective of this thesis is to propose a framework that is capable of modelling the uncertainties in a single underlying framework and implement a mechanism to better interpret the ambiguous output.

In order to achieve this, first, fuzzy approaches are studied to understand the feasibility to address the uncertainties in HMA. In the mean time, to discover the benefits of fuzzy qualitative reasoning over the ordinary fuzzy approaches and the motivation to use it in addressing the uncertainties.

Secondly, this thesis propose and examine the effectiveness of view specific action recognition framework in addressing the size and viewpoint variation. This framework is capable to generalize the different human sizes and ϕ angles from the acquired image and perform action recognition independent to the subject's viewpoint.

Besides that, this thesis also intents to propose a better classification method to interpret the ambiguous cases. This work involves the validation of the existence of non-mutually exclusive cases in the real-world problem that caused the ambiguity. Last but not least, to integrate all the above solutions into a single underlying framework and evaluate the performance in HMA.

1.4 Contributions

The main aim of this thesis is to study the potential of fuzzy approaches, in particular the fuzzy qualitative reasoning to address the uncertainties in HMA due to its feasibility in modelling the uncertainties (will explain more in the literature review). With the purpose to understand the topic thoroughly, extensive study had been done on the conventional

HMA compared to fuzzy HMA with the corresponding review paper is accepted in Pattern Recognition (2015). Technically in this thesis, two main contributions to address the uncertainties are shown in Figure 1.8.

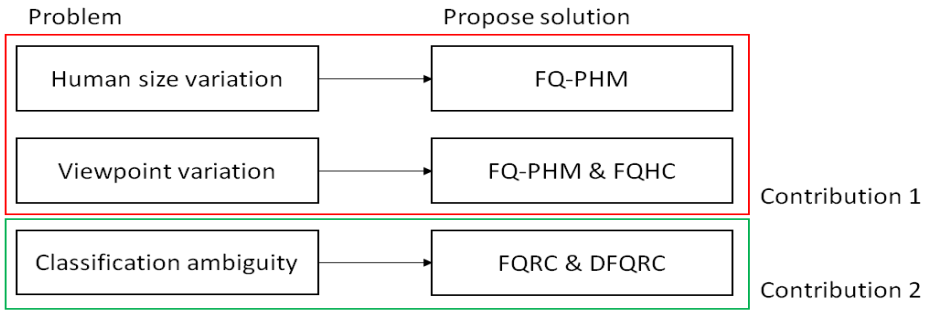


Figure 1.8: Problems to be addressed and the propose solutions.

Contribution 1: In order to achieve the human size and view invariant capability in the single camera system, view specific action recognition framework is introduced in this work as depicted in Figure 1.9. In the process cycle, two main components are necessary in the propose framework, which are the view estimation module and the View Specific Action Model (VSAM). The former is to estimate the viewpoint of a person in front of the camera and the latter is the collection of action models constructed from different viewpoints. The estimated viewpoint will then use to trigger the corresponding VSAM for the inference process.

In the pipeline, a novel representation of human model is proposed namely the Fuzzy Qualitative Poisson-normalized Human Model (FQ-PHM). This is build with the aid of the Poisson solution (Gorelick et al., 2006a,b) and the Fuzzy Quantity Space (FQS) (Liu et al., 2009). The FQ-PHM is a generalized human model in terms of the human size, body anatomy, and the camera position (Φ angle). Extra merit is given to the FQ-PHM as it allows the extraction of proposed human contour called the Fuzzy Qualitative Human Contour (FQHC) that is proven as an effective feature for view estimation. Furthermore, view specific action recognition framework showed that some actions are better recog-

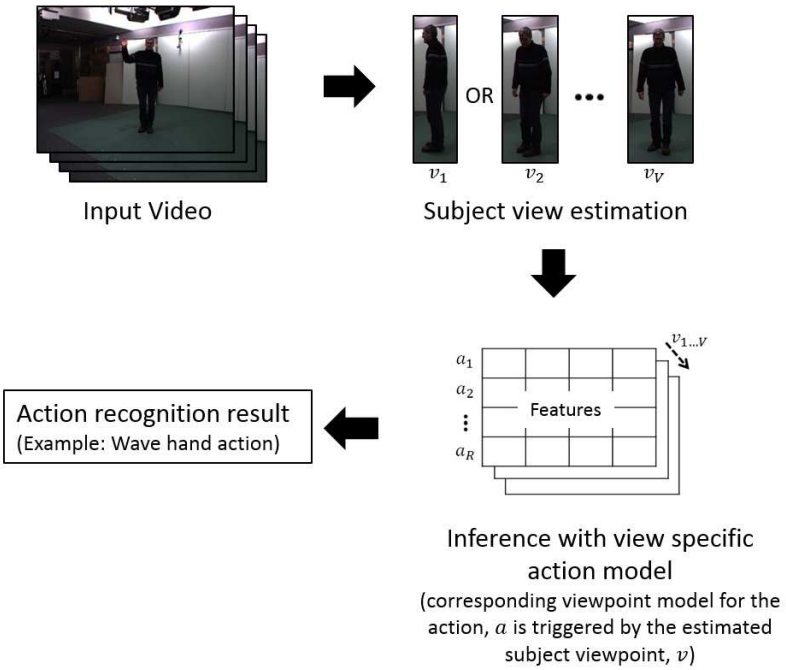


Figure 1.9: Overall framework of the view specific action recognition framework.

nized in certain viewpoints, this is worth to study in depth to enhance the current state-of-the-art methods. This work was accepted in Neural Computing and Applications (2015), and the IEEE International Conference on Fuzzy System (FUZZ-IEEE 2013), Hyderabad, India.

Contribution 2: Sometimes classification result can be ambiguous in HMA and scene context that being used as extra cue for HMA (Ikizler-Cinbis & Sclaroff, 2010; Marszalek et al., 2009). Such ambiguous cases are technically known as the **non-mutually exclusive** cases in this thesis. It can be noticed in HMA, confusion always exist between actions that have similar pattern such as “running”, “walking” and “jogging”. This is similarly in scene understanding where a scene image can be confused between several scene classes when they possess the common characteristics (Boutell et al., 2004; M.-L. Zhang & Zhou, 2007). In order to verify this, an online survey was conducted and participated by a group of people in the range of 12 to 60 years old from different backgrounds such as jobs, environments and countries. The task is to select the image class label that best reflects

the given scene image without prior knowledge of what the ground truth (the answer defined by the researcher) is. Surprisingly the results is in line with the problem statement where non-mutually exclusive case does exists. Thus, binary classification methods are found not so effective to deal with non-mutually exclusive cases. Such finding raises the awareness of computer vision community regarding this very important, but largely neglected issue.

With this in mind, the Fuzzy Qualitative Rank Classifier (FQRC) is proposed to model the non-mutually exclusive case, and develop an inference method that outputs multi-label result with ranking ability instead of crisp or binary result. To the extend, DFQRC is proposed over ordinary FQRC by endowing the capability to learn the model adaptively in the training phase. Qualitative and quantitative evaluations showed the effectiveness and the efficiency of the proposed FQRC and DFQRC in modelling the ambiguity. The ranking inference mechanism is proven close to human reasoning. The proposed works were respectively accepted in IEEE Transactions on Fuzzy Systems (2015) and the IEEE International Conference on Fuzzy System (FUZZ-IEEE 2012), Brisbane, Australia.

1.5 Outline

This chapter provides an overview of the works presented in this thesis with emphasize on the problem statements, objectives and the contributions. Following are the rest of the chapters with brief introduction.

Chapter 2 presents the background studies on the current trends in HMA that lead to the problem statements. In additions, fuzzy human motion analysis is reviewed thoroughly to understand the effectiveness of fuzzy approaches that contributed to modeling the uncertainties in HMA. Nonetheless, it includes the revisit of the FQS with respect to fuzzy qualitative reasoning and 4-tuple membership number representation which were

adopted in the proposed solutions.

Chapter 3 introduces the view specific action recognition framework. The solution in this chapter first uses the view estimation module to estimate the viewpoint of subject in the image frames, then the VSAM is constructed for view independent action recognition purpose. In the view estimation module, a new representation of human model called the FQ-PHM is built with the aid of Poisson solution and the FQS. It achieved the generalization over the human size, body anatomy and camera positions. A novel human contour descriptor, FQHC which can be extracted from the FQ-PHM is proposed and is experimentally proven to works effectively in the view estimation task.

Chapter 4 describe the motivation of study about non-mutually exclusive (ambiguous) cases and the existence of non-mutually case is validated with an online survey. It is conducted by using the popular scene images. To the extend, FQRC is proposed to model the non-mutually exclusive case and output the multi-label ranking result. The architecture of the training step to generate the Fuzzy Qualitative Trained Model (FQTM) and the inference method are explained.

Chapter 5 proposes the extension of FQRC namely DFQRC to overcome the heuristic training in the FQRC. This is done by adaptively learn the 4-tuple fuzzy numbers to build the FQTM in the training step. Comprehensive experiments have been done using DFQRC in scene understanding to evaluate the performance of the DFQRC over the FQRC. In addition, qualitative and quantitative results have proved its effectiveness and efficiency compared to the state-of-the-art methods. Last but not least, it was applied in HMA and produced promising results.

Chapter 6 concludes the works and provides the suggestions for future work.

CHAPTER 2: LITERATURE REVIEW

In the literature review, the transition of conventional HMA to fuzzy HMA is studied. In the studies, the intention of adopting fuzzy approach to address the uncertainties that abounded in HMA has been review critically in regards of Low-level (LoL), Mid-level (MiL), and High-level (HiL) which reflecting the HMA pipeline. The motivation of selecting fuzzy qualitative reasoning in the propose framework is identified and a brief explanation to the approach is included. In overall, the review is conducted as Figure 2.1.

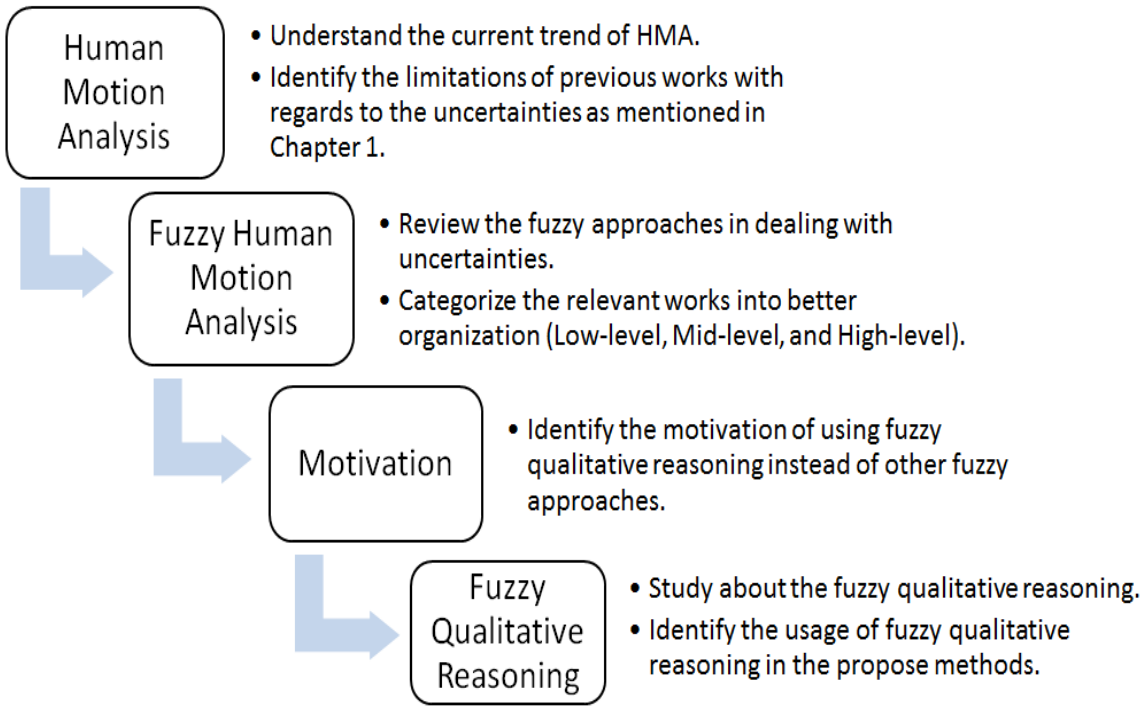


Figure 2.1: The flow of the literature review.

2.1 Human Motion Analysis

HMA in computer vision has been studied extensively for decades due to the demand and promising growth in high specification of camera technology. The importance and popularity of the HMA system has led to several surveys in the literature, as indicated in Table 2.1. One of the earliest surveys was by Aggarwal et al. (1994), focused on

various methods employed in the analysis of the human body motion, which is in non-rigid form. Cédras & Shah (1995) gave an overview on the motion extraction methods using the motion capture systems and focused on action recognition, individual body parts recognition, and body configuration estimation. Aggarwal & Cai (1997) used the same taxonomy as in the survey by Cédras & Shah (1995), but engaging different labels for the three classes, that is further dividing the classes into subclasses yielding a more comprehensive taxonomy. Gavrila (1999) gave an overview on the applications of visual analysis of human movements, and their taxonomy covered the 2D and 3D approaches with and without the explicit shape models.

As the works in this area prosper, public datasets start to gain importance in the vision community to meet different research challenges. The KTH (Schuldt et al., 2004) and the Weizmann (Blank et al., 2005; Zelnik-Manor & Irani, 2001) datasets were the most popular human actions datasets introduced in the early stages. However, neither of the datasets represent the human actions in a real-world environment. In general, each action is performed in a simple manner with just a single actor, static background and fixed viewpoint. KTH however considered a few complex situations such as different lighting conditions, but it is still far away from the real-world complex scenarios. Therefore, other datasets were created such as the CAVIAR, ETISEO, CASIA Action, MSR Action, HOLLYWOOD, UCF datasets, Olympic Sports and HMDB51, BEHAVE, TV Human Interaction, UT-Tower, UT-Interaction, etc. Please refer to Chaquet et al. (2013) for a complete list of the currently available datasets in HMA.

Due to the advancement of the technology, using networks of multiple cameras for monitoring public places such as airports, shopping malls, etc. were emerged. Aggarwal & Cai (1997); Gavrila (1999); Holte et al. (2011); Hu, Tan, et al. (2004); Ji & Liu (2010); Moeslund et al. (2006); Poppe (2010); L. Wang et al. (2003); Weinland et al. (2011) moved ahead to survey on the representation and recognition of the human actions in

multiple-views aspect. Various new datasets were created exclusively for this purpose such as the IXMAS, i3DPost, MuHAVi, VideoWeb and CASIA Action. Last but not the least, Aggarwal & Ryoo (2011); Cristani et al. (2013); Gavrila (1999); Hu, Tan, et al. (2004); Moeslund & Granum (2001); Turaga et al. (2008); L. Wang et al. (2003) surveyed on the various applications of HMA such as the smart surveillance and advanced user interface for human-computer interaction. For the convenience of the readers, the surveys papers and their focuses are summarized in Table 2.1 and 2.2.

Based on the survey, vision-based HMA can be categorized into two main categories which are view-dependent and view-independent approaches (Holte et al., 2011; Ji & Liu, 2010). The former require the subject to fixed his/her viewpoint in front of the camera while the latter is free from this restriction during the image acquiring process.

Most of the works fall into the former category which assumed that all the actions are performed at a fixed viewpoint (Bobick & Davis, 2001; Chan & Liu, 2009; Gorelick et al., 2007; Laptev, 2005). From the study, the pioneer work in space time approach is by Bobick & Davis (2001) where they use temporal information to build two dimensional binary foreground images called Motion-energy image (MEI) and the scalar values of foreground images called Motion-history image (MHI) to represent an action. Template matching is then applied to the pair of MEI and MHI and provides promising results in recognition of a series of ballet actions. Besides that, Gorelick et al. (2007) proposed an approach that utilizes the Poisson solution to estimate the moving torso and protruding limbs for action recognition. Nonetheless, Bregonzio et al. (2009); Laptev (2005); Scovanner et al. (2007) extended the idea of spatial interest points into spatio-temporal domain where a descriptor composed of space time information is built to classify an event. However, the aforementioned works is suffered from the limitation of viewpoint where they assumed that a subject will always performs in a static viewpoint.

While in the second category, many works focused on multi-camera approaches to

Table 2.1: Summarization of the survey papers on HMA.

Paper	Author	Title	Description	Year
Aggarwal et al. (1994)	J.K. Aggarwal, Q. Cai, W. Liao & B. Sabata	Articulated and elastic non-rigid motion: a review	The earliest survey on HMA, focusing on various methods used in the articulated and non-rigid motion.	1994
Cédras & Shah (1995)	C. Cedras & M. Shah	Motion-based recognition: a survey	An overview on various methods for motion extraction: action recognition, body parts recognition and body configuration estimation.	1995
Aggarwal & Cai (1997)	J.K. Aggarwal & Q. Cai	Human motion analysis: a review	Focus on motion analysis of human body parts, tracking moving human from a single view or multiple camera perspectives, and recognizing human activities from video.	1997
Gavrila (1999)	D.M. Gavrila	The visual analysis of human movement: a survey	Discussed various methodologies grouped into 2D approaches with or without explicit shape models as well as 3D approaches.	1999
Pentland (2000)	A. Pentland	Looking at people: sensing for ubiquitous and wearable computing	Reviewed the state-of-the-art of "looking at people" focusing on person identification and surveillance monitoring.	2000
Moeslund & Granum (2001)	T.B. Moeslund & E. Granum	A survey of computer vision-based human motion capture	Overview on the taxonomy of system functionalities: initialization, tracking, pose estimation and recognition.	2001
L. Wang et al. (2003)	L. Wang, W. Hu & T. Tan	Recent Developments in Human Motion Analysis	Focus on three major issues: human detection, tracking and activity understanding.	2003
Hu, Tan, et al. (2004)	W. Hu, T. Tan, L. Wang & S. Maybank	A survey on visual surveillance of object motion and behaviors	Reviewed recent developments in visual surveillance of object motion and behaviors in dynamic scenes and analyzed possible research directions.	2004
Moeslund et al. (2006)	T. B. Moeslund, A. Hilton, & V. Kruger	A survey of advances in vision-based human motion capture and analysis	Discuss recent trends in video-based human motion capture and analysis.	2006
Poppe (2007)	R. Poppe	Vision-based human motion analysis: An overview	HMA with two phases: modelling (concerned with construction of the likelihood function) and estimation (finding the most likely pose given the likelihood surface).	2007
Turaga et al. (2008)	P. Turaga, R. Chellappa, V. Subrahmanian & O. Udrea	Machine recognition of human activities: A survey	Addressed the problem of representation, recognition and learning of human activities from video and related applications.	2008
Ji & Liu (2010)	X. Ji & H. Liu	Advances in view-invariant human motion analysis: A review	Emphasized on the recognition of poses and actions. Three major issues were addressed: human detection, view-invariant pose representation and estimation, and behavior understanding.	2010
Poppe (2010)	R. Poppe	A survey on vision-based human action recognition	Overview on current advances in vision-based human action recognition, addressing challenges faced due to variations in motion performance, recording settings and inter-personal differences. Also, discussed shortcomings of the state-of-the-art and outline promising directions of research.	2010
Candamo et al. (2010)	J. Candamo, M. Shreve, D. Goldgof, D. Sapper, & R. Kasturi	Understanding transit scenes: A survey on human behavior-recognition algorithms	Reviewed automatic behavior recognition techniques, focusing on human activity surveillance in transit applications context.	2010
Aggarwal & Ryoo (2011)	J. K. Aggarwal & M. S. Ryoo	Human activity analysis: A review	Discussed methodologies developed for simple human actions as well as high-level activities.	2011
Weinland et al. (2011)	D. Weinland, R. Ronfard & E. Boyer	A survey of vision-based methods for action representation, segmentation and recognition	Concentrated on the approaches that aim at classification of full-body motions: kicking, punching and waving, and further categorized them according to spatial and temporal structure of actions, action segmentation from an input stream of visual data and view-invariant representation of actions.	2011
Holte et al. (2011)	M.B. Holte, T.B. Moeslund, C. Tran & M.M. Trivedi	Human action recognition using multiple views: A comparative perspective on recent developments	Presented a review and comparative study of recent multi-view 2D and 3D approaches for HMA.	2011
Lara & Labrador (2013)	O. Lara & M. Labrador	A survey on human activity recognition using wearable sensors	Surveys human activity recognition based on wearable sensors. 28 systems were qualitatively evaluated in terms of recognition performance, energy consumption, and flexibility etc.	2013
L. Chen et al. (2013)	L. Chen, H. Wei & J. Ferryman	A survey of human motion analysis using depth imagery	Reviewed the research on the use of depth imagery for analyzing human activity (e.g. the Microsoft Kinect). Also listed publicly available datasets that include depth imagery.	2013
Cristani et al. (2013)	M. Cristani, R. Raghavendra, A. Del Bue & V. Murino	Human behavior analysis in video surveillance: A social signal processing perspective	Analyzed the social signal processing perspective of the automated surveillance of human activities such as face expressions and gazing, body posture and gestures, vocal characteristics etc.	2013
Chaquet et al. (2013)	J. M. Chaquet, E. J. Carmona & A. F.-Caballero	A survey of video datasets for human action and activity recognition	Provide a complete description of the most important public datasets for video-based human activity and action recognition.	2013

Table 2.2: Criterion on which the previous survey papers on HMA emphasized on (1994-2013). Note that those criterion without a 'tick' means the topic is not discussed comprehensively in the corresponding survey paper, but might be touched indirectly in the contents.

Year	Paper	Human Detection	Tracking	Behavior Understanding	Multi-view	Feature extraction	Datasets	Application
1994	Aggarwal et al. (1994)	-	✓	✓	-	✓	-	-
1995	Cédras & Shah (1995)	-	✓	✓	-	✓	-	-
1997	Aggarwal & Cai (1997)	✓	✓	✓	✓	✓	-	-
1999	Gavrila (1999)	✓	✓	✓	✓	✓	-	✓
2000	Pentland (2000)	✓	✓	✓	-	✓	-	-
2001	Moeslund & Granum (2001)	✓	✓	✓	-	✓	-	✓
2003	L. Wang et al. (2003)	✓	✓	✓	✓	-	-	✓
2004	Hu, Tan, et al. (2004)	✓	✓	✓	✓	-	-	✓
2006	Moeslund et al. (2006)	✓	✓	✓	✓	-	-	-
2007	Poppe (2007)	✓	✓	-	-	✓	-	-
2008	Turaga et al. (2008)	✓	-	✓	-	✓	-	✓
2010	Ji & Liu (2010)	✓	-	✓	✓	-	✓	-
2010	Poppe (2010)	-	-	✓	✓	✓	✓	-
2010	Candamo et al. (2010)	✓	✓	✓	-	-	-	-
2011	Aggarwal & Ryoo (2011)	-	-	✓	-	✓	✓	✓
2011	Weinland et al. (2011)	✓	-	✓	✓	✓	✓	-
2011	Holte et al. (2011)	-	-	✓	✓	✓	✓	-
2013	Lara & Labrador (2013)	-	-	✓	-	✓	✓	-
2013	L. Chen et al. (2013)	✓	✓	✓	-	-	✓	-
2013	Cristani et al. (2013)	✓	✓	✓	-	-	-	✓
2013	Chaquet et al. (2013)	-	-	-	-	-	✓	-

achieve view independent action recognition (Ahmad & Lee, 2006; Anderson, Luke, et al., 2009b; Ashraf et al., 2013; Weinland et al., 2007, 2006; Yilma & Shah, 2005). The drawback of using multi-camera approach is, it is only applicable to closed controlled environment such as a calibrated space or room, and these system may be impractical to deploy in an open environment such as airport or street. In these multi-camera methods, 2D models are extended into 3D models to reconstruct the human shape in a volumetric space. For examples, Anderson, Luke, et al. (2009b) built a 3D representation of human using multiple cameras and called voxel person. The features such as the voxel person's centroid; eigen-based height; and the similarity of voxel person's primary orientation and the ground plane normal are then extracted from the voxel person to infer the

falling activity. On the other hand, Weinland et al. (2007) proposed a new framework that model actions using 3D occupancy grids, which are built from multiple viewpoints in an exemplar-based HMM. For recognition, the 3D exemplars are used to produce 2D image information for matching purpose.

On the other hand, the epipolar geometry between two views (Kimura & Saito, 2001) is also a popular method to analyse the body posture from multiple cameras at different viewing angle. For example, Ashraf et al. (2013) utilized the epipolar geometry to obtain a fundamental matrix between two fixed cameras and the concept of fundamental ratios is investigated which are invariant to camera intrinsic parameters in view invariant action recognition (Ashraf et al., 2013). Besides that, Yilma & Shah (2005) proposed the extension of the standard epipolar geometry to support dynamic scenes where the cameras are movable to study the action. Although the above mentioned approaches achieved significant results in view independent action recognition, multi-camera approaches are only applicable to closed-controlled environment and it is impractical and expensive to deploy in a real-world environment.

This problem is addressed by Lewandowski, Makris, & Nebel (2010) where they emphasize the importance for implementing view independent action recognition with single camera. In their work, a torus-like descriptor Lewandowski, Makris, & Nebel (2010) is proposed which takes advantage of Temporal Laplacian Eigenmaps (Lewandowski, Martinez-del Rincon, et al., 2010) and the Decomposable Generative Model (Lee & Elgammal, 2007) to recognize action in view invariant manner. However, their approach faces the difficulty in dealing with the variant in camera positions, which is the ϕ angle (please refer to section 1.1.2). Consequently, extensive training is needed to cope with that.

Apart from the above mentioned works, there is a raise in the interest for the combination of scene context in HMA (Ikizler-Cinbis & Sclaroff, 2010; Marszalek et al., 2009)

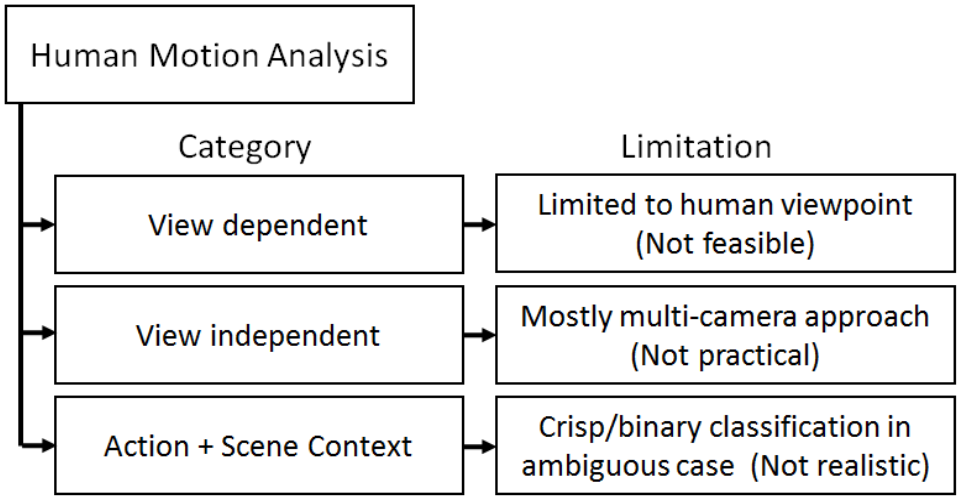


Figure 2.2: Summary of HMA and its respective categories and limitations.

recently. The scenery information was proven to be effective as an extra cue to infer an activity in view independent manner. Nonetheless, the source of uncertainty exists in these approaches and causes the bottleneck in achieving good performance. This is mainly due to the ambiguity among the actions and the scene images. It implies that the final result could be non-mutually exclusive where the testing sample could belongs to more than one action or image classes. This indicates that the conventional crisp or binary classifier (K-Nearest Neighbour, Support Vector Machine, etc.) may not effective in this ambiguous case because they brutally force an output to belongs to solely one class. The summary of all these problems and its respective limitations are shown in Figure 2.2.

Fuzzy set theory (Zadeh, 1965) which endowed with the capability of modelling the uncertainties has lead HMA to a new research direction in fuzzy HMA to deal with the above limitations. In this literature review, a detailed review of the works in fuzzy HMA will be presented in the next section with respect to how fuzzy approaches deal with the uncertainties abounded in the HMA system.

2.2 Fuzzy Human Motion Analysis

Various uncertainties can happens in almost every computer vision application (Huntsberger et al., 1986) especially when dealing with the real-world problems such as HMA system (Chan et al., 2011). Apparently, failed in handling these uncertainties can cause catastrophe such as system failure or bad performance. In order to solve this, many researchers have focused to apply fuzzy approaches in HMA. This is because fuzzy set theory is endowed with the capability to model the uncertainties.

Fuzzy set theory since its inception in 1965 (Zadeh, 1965), has played an important role in handling uncertainties and is successfully integrated into various applications, for example the subway system in Sendai, Japan; washing machine; digital camera and so on. Fuzzy set is used to represent a class of object with the membership grade (Zadeh, 1965). More often than not, in our nature surrounding, there are objects which generally hard to be distinguish distinctly with neither 'Yes' or 'No'. In other word, crisp or binary answer. Instead, fuzzy set is used to measure the degree of the belonging of that object to the related classes with interval 0 to 1. The flexibility of the sets theory also empowers the use of set operations to optimize the meaning of fuzzy sets such as intersection, union, complement, and many other advance operations for fuzzy relation (Zadeh, 1988). With the advancement in fuzzy set theory, various fuzzy approaches have been proposed which significantly contributes to the HMA such as type-1 and type-2 Fuzzy Inference System (FIS) (Karnik et al., 1999; Mendel & John, 2002; Zadeh, 1988), fuzzy clustering (Krishnapuram & Keller, 1993; Pal et al., 2005), and fuzzy qualitative reasoning (Shen & Leitch, 1993; Shen et al., 1993), etc.

In this literature, the focus will be primarily on the solutions that utilized the fuzzy approaches towards HMA. Particularly regarding the early years of the fuzzy set oriented approaches for HMA, individuating how the fuzzy set may improve the HMA, envisaging

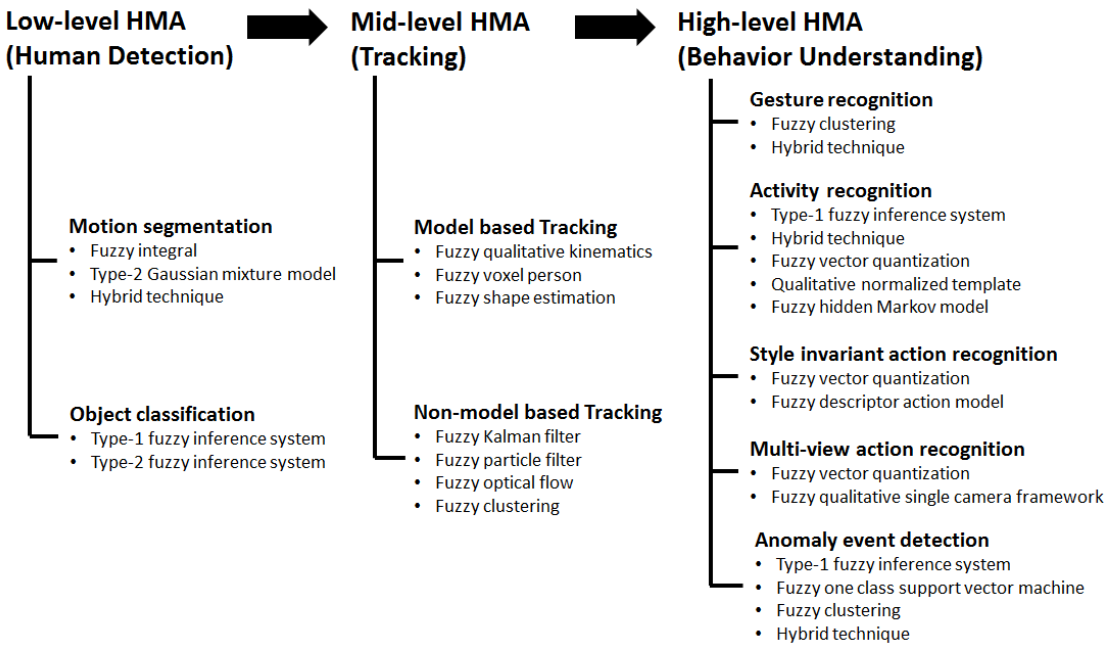


Figure 2.3: Overall taxonomy of the review in fuzzy HMA. It is organized according to the pipeline of HMA from Low-level to High-level with subcategories of the fuzzy approaches that have been employed in the literature.

and delineating the future perspectives. This is in contrast to the past research works where stochastic solutions were the predominant discussions.

For simplicity, the review is categorized into three broad levels which are important to achieve a successive HMA system. The respective uncertainties abounded in each of these level of processing is discussed with the corresponding fuzzy solution. The three broad levels are: LoL, MiL, and HiL HMA, as depicted in Figure 2.3. The LoL HMA is the background foreground subtraction which contributes in the pre-processing of the raw images to discover the areas of interest such as the human region. MiL HMA is the object tracking. In this level, it serves as the means to prepare data for human pose estimation and activity recognition. HiL HMA is the behavior understanding where the objective is to correctly classify the human motion patterns into activity categories; for example, walking, running, wave hands and so on.

2.2.1 Low-level

In HMA, by ignoring the image acquiring stage, the foremost task will be the low-level vision which is also considers the human detection step that includes motion segmentation and object classification. Human detection is the initial step in almost every low-level vision-based HMA system before the higher level processing steps such as tracking and behavior understanding can be performed. Technically, human detection aims at locating and segmenting the human regions from the other subjects in the image. This process usually involves the motion segmentation followed by the object classification.

2.2.1 (a) Motion segmentation

Motion segmentation aims at separating the moving objects from the natural scenes. The extracted motion regions are vital for the next level of processing, e.g. it relaxes the tracking complexity as only the pixels with changes are considered in the process. However, some critical situations in the real-world environment such as the illumination changes, dynamic scene movements (e.g. rainy weather, waving tree, rippling water and so on), camera jittering, and shadow effects make it a daunting task.

Background subtraction is one of the popular motion segmentation algorithms that has received much attention in the HMA system. This is due to the usefulness of its output that is capable of preserving the shape information, as well as helps in extracting motion and contour information (Bobick & Davis, 2001; Lewandowski, Makris, & Nebel, 2010; Weinland et al., 2006). In general, background subtraction is to differentiate between the image regions which have significantly different characteristics from the background image (normally denoted as the background model). A good background subtraction algorithm comprises of a background model that is robust to the environmental changes, but sensitive to identify all the moving objects of interest. There are some fuzzy approaches

that endowed this capability in the background subtraction which will be discussed as follows.

Fuzzy integral

Information fusion from a variety of sources is the most straightforward and effective approach to increase the classification confidence, as well as removing the ambiguity and resolving the conflicts in different decisions. Rationally in background modeling, the combination of several measuring criteria (also known as the features or attributes) can strengthen the pixel's classification as background or foreground. However the basic mathematical operators used for aggregation such as the minimum, maximum, average, median, 'AND', and 'OR' operators provide crisp decisions and utilize only a single feature that tends to result in false positive (H. Zhang & Xu, 2006). In contrast, the fuzzy integrals take into account the importance of the coalition of any subset of the criteria (El Baf et al., 2008b).

Figure 2.4: Comparison between the Sugeno and the Choquet fuzzy integral methods for background subtraction El Baf et al. (2008b). First row: The original image. Second row: the output from the Sugeno fuzzy integral on the left and the Choquet fuzzy integral on the right.

In general, the fuzzy integral is a non-linear function that is defined with respect to the fuzzy measure such as a belief or a plausibility measure (Tahani & Keller, 1990), and is employed in the aggregation step. As the fuzzy measure in the fuzzy integral is defined on a set of criteria, it provides precious information about the importance and relevance of the criteria to the discriminative classes. Thus it achieves feature selection with better classification results. H. Zhang & Xu (2006) proposed to use the Sugeno integral (Marichal, 2000) to fuse color and texture features in their works for better classification of the pixel that belongs to either background or foreground, while Balcilar & Sonmez (2013); El Baf et al. (2008a,b) improved the work by replacing the Sugeno integral with the Choquet integral (Murofushi & Sugeno, 1989). The main reason is that the Choquet integral which was adapted for cardinal aggregation, was found to be more suitable than the Sugeno integral that assumed the measurement scale to be ordinal (Narukawa & Murofushi, 2004; Sugeno & Kwon, 1995). The corresponding results for the comparison between the Sugeno integral and the Choquet integral are shown in Figure 2.4. The background modeling process using the fusion of color and texture features have shown to achieve better detection of the moving targets against cluttered backgrounds, backgrounds with little movements, shadow effects as well as illumination changes.

Type-2 Gaussian mixture model

The studies on the background subtraction (Cheung & Kamath, 2004; Piccardi, 2004) have shown that the Gaussian Mixture Model (GMM) is one of the popular approaches used in modeling the dynamic background scene. It solves the limitation in the unimodal model (single Gaussian) which is unable to handle the dynamic backgrounds such as waving tree and water rippling. The expectation-maximization algorithm is normally used in the initialization step of the GMM to estimate the parameters from a training sequence using the Maximum-likelihood (ML) criterion. However, due to insufficient or

noisy training data, the GMM may not be able to accurately reflect the underlying distribution of the observations. This is because exact numbers must be used in the likelihood computation and unfortunately, these parameters are bounded by uncertainty. In order to take into account the uncertainty, the fuzzy set theory was explored.

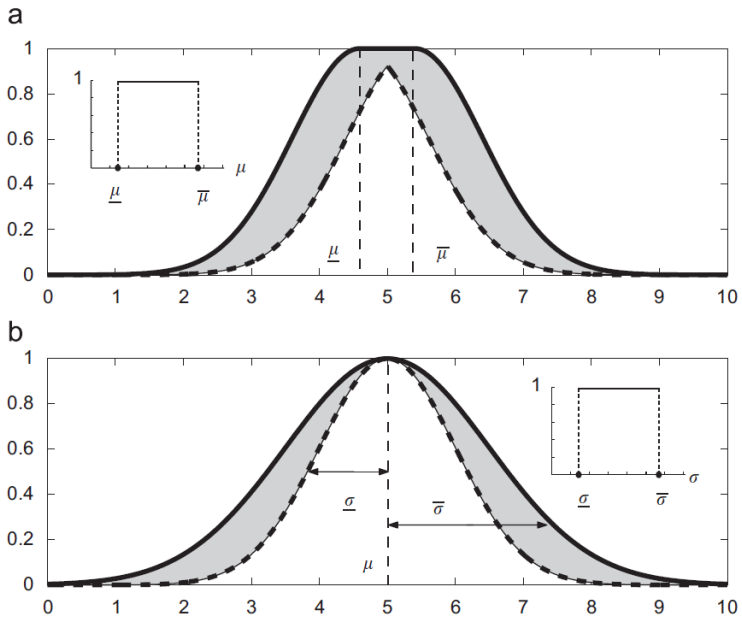


Figure 2.5: Example of the type-2 fuzzy membership function of the Gaussian model with (a) uncertain mean, μ and (b) uncertain standard deviation, σ, having uniform possibilities. The shaded region is the Footprint of Uncertainty (FOU). The thick solid and dashed lines denote the lower and upper membership functions Zeng et al. (2008).

However, there has been an argument that type-1 fuzzy set, which is an ordinary fuzzy set (Zadeh, 1965), has limited capability in modeling the uncertainty. This is because the membership function for the type-1 fuzzy set is not associated with uncertainty. Therefore, type-2 fuzzy sets (Mendel & John, 2002) emerged from the type-1 fuzzy set by generalizing it to handle more uncertainty in the underlying fuzzy membership function. As a whole, the type-2 fuzzy membership function is itself a fuzzy set and referring to Figure 2.5, it can be noticed that the uncertainty in the fuzzy membership function is represented in the shaded area known as the Footprint of Uncertainty (FOU). With the capability of type-2 fuzzy set to handle higher dimensions of uncertainty, it was adopted in Zeng et al. (2008) to represent the multivariate Gaussian with an uncertain mean vector

or a covariance matrix. In more detail, it was assumed that the mean and the standard deviation vary within the intervals with uniform possibilities (Figure 2.5), instead of crisp values as in the conventional GMM.

Several works (Bouwmans et al., 2009; El Baf et al., 2008c, 2009) have been reported that utilized the type-2 fuzzy GMM to deal with insufficient or noisy data, and resulted in better background subtraction model. In the later stage, Zhao et al. (2012) made an improvement on these works with the inclusion of spatial-temporal constraints into the type-2 fuzzy GMM by using the Markov Random Field.

Hybrid technique

Although the fuzzy approaches provide superior performance in background subtraction, most of these approaches have a common problem, that is how to optimize the parameters in their algorithms. These parameters can be the intrinsic parameters such as the interval values of the membership function, or the threshold value for the inference step. Optimizing these parameters usually increases the overall system performance. However, such steps require human intervention (El Baf et al., 2008a,b; H. Zhang & Xu, 2006). For example, the trial and error process to determine a classification threshold value is a tedious job, computationally expensive and subjective (Sigari et al., 2008).

Fortunately, such limitations can be handled by using hybrid techniques, i.e. the combination of fuzzy approaches with machine learning methods. Lin et al. (2000) applied neural fuzzy framework to estimate the image motion. The back-propagation learning rule from a five-layered neural fuzzy network was used to choose the best membership functions so that the system is able to adapt to different environments involving occlusions, specularity, shadowing, transparency and so on. Besides that, Maddalena & Petrosino (2010) introduced a spatial coherence variant incorporated with the self-organizing neural network to formulate a fuzzy model to enhance the robustness against

false detection in the background subtraction algorithm. Z. Li et al. (2012) used both the particle swarm optimization and the kernel least mean square to update the system parameters of a fuzzy model, and Calvo-Gallego et al. (2013) employed a tuning process using the Marquardt-Levenberg algorithm within a fuzzy system to fine-tune the membership function. In order to determine the appropriate threshold value for the classification task, Shakeri et al. (2008) proposed a novel fuzzy-cellular method that helps in dynamically learning the optimal threshold value.

2.2.1 (b) Object classification

The outcome from the motion segmentation usually results in a rough estimation of the moving targets in a natural scene. These moving targets in a natural scene can be shadow, vehicle, flying bird and so on. Before the region is further processed at the next level, it is very important to verify and refine the interest object by eliminating the unintended objects. In this section, we discuss some fuzzy approaches that are beneficial in the human object classification.

Type-1 fuzzy inference system

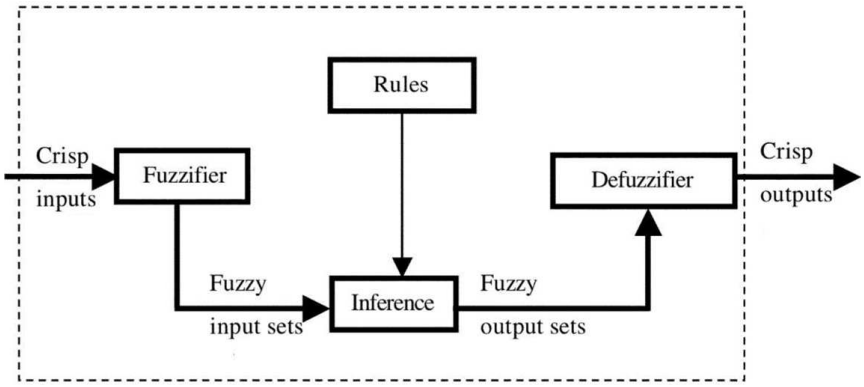
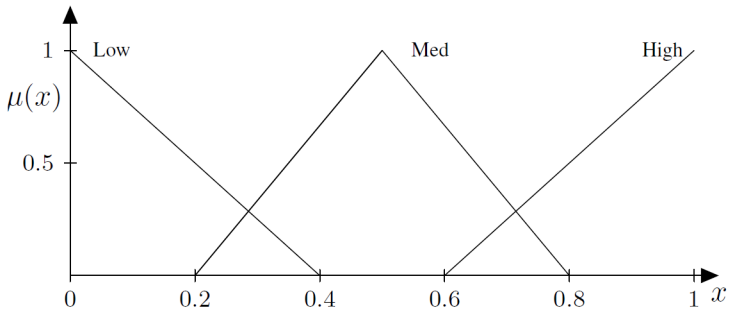


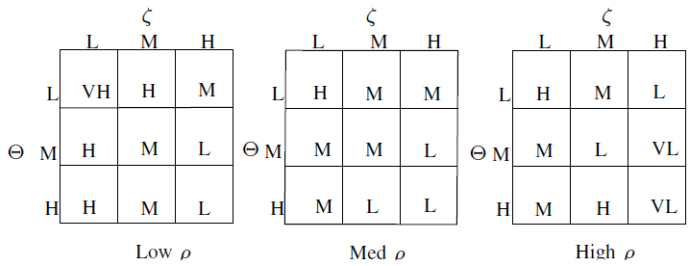
Figure 2.6: Type-1 Fuzzy Inference System (Mendel et al., 2006).

The Type-1 Fuzzy Inference System (FIS) (Yager & Zadeh, 1992) is a complete fuzzy decision making system that utilizes the fuzzy set theory. It has been successfully

applied in numerous applications for commercial and research purposes. Its popularity is due to the capability to model the uncertainty and the sophisticated inference mechanism that greatly compromises the vague, noisy, missing, and ill-defined data in the data acquisition step. Figure 2.6 shows the overall framework of a typical Type-1 FIS, where it includes three important steps: fuzzification, inference, and defuzzification. The fuzzification step maps the crisp input data from a set of sensors (features or attributes) to the membership functions to generate the fuzzy input sets with linguistic support (Zadeh, 1988). Then, the fuzzy input sets go through the inference steps with the support from a set of fuzzy rules to infer the fuzzy output sets. Finally, the fuzzy output sets are defuzzified into the crisp outputs.



(a)



(b)

Figure 2.7: (a) Example of the membership function for the distance feature where $\mu(x)$ denotes the membership value, and x is the distance value. (b) The fuzzy rules for the fuzzy input for three features (Distance, ρ; Angle, Θ and Cord to Arc Ratio, ζ), and its corresponding fuzzy output (VL=Very low, L=Low, M=Med, H=High, VH=Very High) Mahapatra et al. (2013).

In human detection, the FIS is an effective and direct approach to distinguish between the human and non-human with different features (Chowdhury & Tripathy, 2014; Mahapatra et al., 2013; See et al., 2005). As an example, Mahapatra et al. (2013) ex-

tracted three features from the contours of the segmented region, such as the distance to the centroid, angle, and cord to arc ratio, and input them into the FIS. The corresponding fuzzy membership function and a set of fuzzy rules were used to infer the fuzzy output as depicted in Figure 2.7. The fuzzy outputs (VL, L, M, H, VH) were then defuzzified into the crisp outputs, and used to perform human classification. For example, if the crisp output is found to be less than the threshold value, then it is recognized as a human and vice versa.

Besides that, X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006) studied in depth about the problems encountered in the human classification tasks, such as the situations where the unintended objects are attached to the classified human region. This problem often occurs in the silhouette based classification output. In general, silhouette is the binary representation of the segmented regions from the background subtraction techniques, where in HMA, human silhouette has proved its sufficiency to describe the activities captured by the video (Bobick & Davis, 2001; Lewandowski, Makris, & Nebel, 2010; Weinland et al., 2006). For example. a chair that is being moved by a person can be misclassified as a part of the segmented region, and included as part of the silhouette image. In order to solve this, X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006) applied the FIS to perform an adaptive silhouette extraction in the complex and dynamic environments. In their works, they used multiple features such as the sum of absolute difference (SAD), fraction of neighbor blocks, and distance between blocks and human body centroid. A set of fuzzy rules were generated, for instance, "IF SAD is SMALL, AND the fraction of neighboring silhouette blocks belong to the human body is LARGE, AND the distance from the centroid is SMALL, THEN the new block is more likely to be a human silhouette block". Depending upon the application, the FIS is capable of modeling different sources of features by generating the appropriate fuzzy membership functions and the fuzzy rules.

Type-2 fuzzy inference system

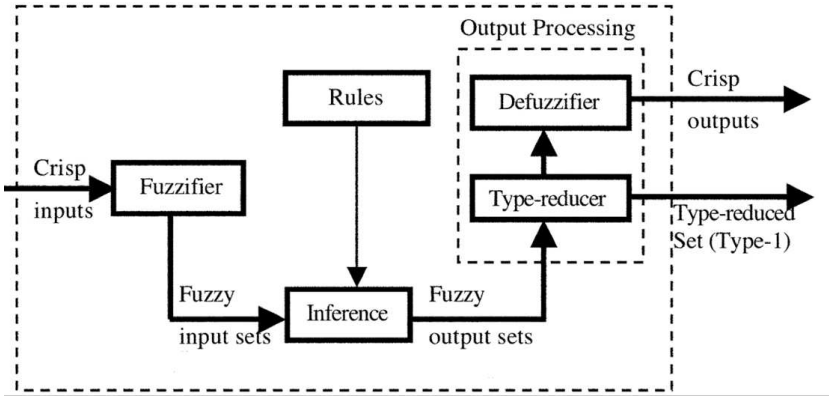


Figure 2.8: Type-2 Fuzzy Inference System Mendel et al. (2006).

To a certain extent, the overall performance of the system from X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006) may be degraded due to the misclassification of the objects in the proposed type-1 FIS. Taking this into account, Yao et al. (2012) employed the interval type-2 FIS (Liang & Mendel, 2000) which is capable of handling higher uncertainty levels present in the real world dynamic environments.

In general, as aforementioned, the type-2 FIS differs from the type-1 FIS in terms of the type-2 FIS offers the capability to support higher dimensions of uncertainty. The main focus in the type-2 FIS is the membership function that is used to represent the input data, where the membership function itself is a fuzzy set with FOU bounded in an ordinary membership function. In consequences, the input data is first fuzzified into type-2 input fuzzy sets, and then go through the inference process where the rules can be similar as the type-1 FIS. Before the defuzzification step takes place, the type-2 output fuzzy sets must be reduced from type-2 to type-1 output fuzzy set. This is processed by using a type-reducer, as depicted in Figure 2.8.

Using the same features as X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006), Yao et al. (2012) proposed to fuzzify the input feature values into the type-2 fuzzy sets using the singleton fuzzification method (Karnik & Mendel, 1998). Consequently, it produces the interval type-2 membership functions for the inference pro-

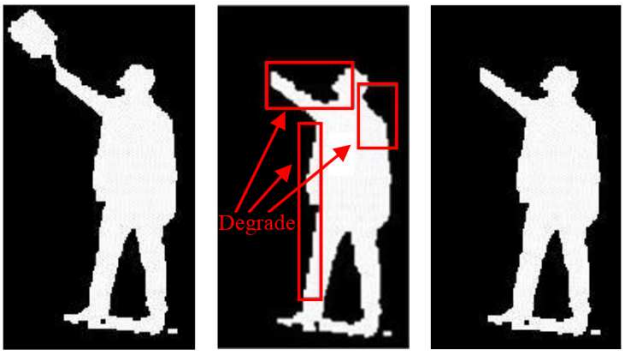


Figure 2.9: Background subtraction on the image of a person raising a book. (a) Extracted silhouette by using the GMM, but is unable to eliminate the unintended object (book). (b) Extracted silhouette after using type-1 FIS to detach the book from the human, but degraded as a result X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006). (c) Extracted silhouette after using type-2 FIS where the result is much smoother.

cess. Their approach was tested on a set of images captured from the real world environment that contains single person, multi-person and the crowded scenes, respectively. The ground truth data was captured from the cameras deployed around their laboratory (i.e. a smart living room) to analyze people's regular activities. Their proposed work showed that the type-2 FIS provides much better results as compared to the type-1 FIS (Figure 2.9). The summary of the works using the fuzzy approaches to solve the uncertainties in LoL HMA is presented in Table 2.3.

Table 2.3: Summarization of research works in LoL HMA using the Fuzzy approaches.

LoL processing	Problem statements / Sources of Uncertainty	Authors	Why fuzzy?	Approach
Motion segmentation	Critical situations such as illumination changes, dynamic scene movements, camera jittering, and shadow effects confuse the pixels belonging to the background model or the foreground object.	Balcilar & Sonmez (2013); El Baf et al. (2008a,b); H. Zhang & Xu (2006)	Information fusion from a variety of sources using the fuzzy aggregation method relaxes the crisp decision problem that causes confusion in the specific class.	Fuzzy integral
	Insufficient and noisy training data do not accurately reflect the distribution in an ordinary Gaussian Mixture Model (GMM) background modelling process.	Bouwmans et al. (2009); El Baf et al. (2008c, 2009); Zhao et al. (2012)	The uncertainty in GMM is bounded with interval mean and standard deviation instead of the crisp values. Type-2 fuzzy set is utilized to handle higher dimensions of uncertainty within the type-1 membership itself.	Type-2 Fuzzy GMM
	Difficulty in determining the optimum parameters in the fuzzy system such as the membership function or the threshold value for the decision making process in the background subtraction algorithms.	Calvo-Gallego et al. (2013); Z. Li et al. (2012); Lin et al. (2000); Maddalena & Petrosino (2010); Shakeri et al. (2008)	Integration of the machine learning techniques with the fuzzy approaches allow the system to learn the optimum parameters that leads to better overall system performance and the feasibility to adapt to various situations depending on the task in hand.	Hybrid technique
Object classification	The confusion between the human and non-human objects, and the unintended objects attached to the human region causes the uncertainty in the classification tasks.	X. Chen, He, Anderson, et al. (2006); X. Chen, He, Keller, et al. (2006); Chowdhury & Tripathy (2014); Mahapatra et al. (2013); See et al. (2005)	Type-1 FIS is able to model the uncertainty in the features data as the membership function, and perform inference using the fuzzy rules to achieve better classification results.	Type-1 FIS
	The insufficiency of the type-1 FIS causes the misclassification of the objects and the degradation in the silhouette extraction.	Yao et al. (2012)	Type-2 fuzzy set offers the capability to support higher dimensions of uncertainty where in this case, the smoother classification results can be obtained.	Type-2 FIS

2.2.2 Mid-level

Once we have successfully located the human in the frame, the next step is to track the human movements over time for the higher level interpretation. Tracking is a crucial step in HMA as it forms the basis for data preparation for HiL HMA tasks such as action recognition, anomaly event detection and so on. The aim of the tracking algorithm is to reliably track the object of interest such as the human body from a sequence of images, and it can be categorized either the model based or the non-model based motion tracking.

2.2.2 (a) Model based tracking

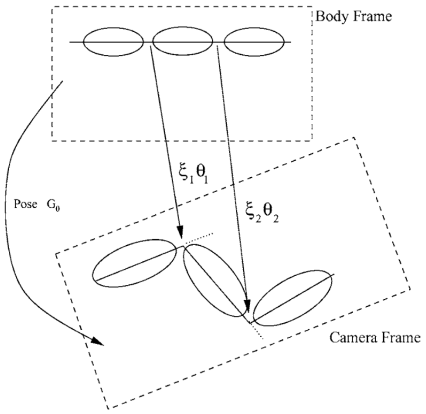
In the model based human motion tracking, the human body models such as the stick figures, 2D and 3D motion description models are adopted to model the complex, non-rigid structure of the human body (Guo et al., 1994; Iwai et al., 1999; Ju et al., 1996; Kakadiaris & Metaxas, 1996; Leung & Yang, 1995; Niyogi & Adelson, 1994; Rehg & Kanade, 1995; Rohr, 1994; Silaghi et al., 1998; Wachter & Nagel, 1997). Readers can

refer to Aggarwal & Cai (1997); Gavrila (1999); Moeslund & Granum (2001); L. Wang et al. (2003) for the detailed reviews. The stick figure model represents the human body as a combination of sticks or line segments connected by the joints (Guo et al., 1994; Iwai et al., 1999; Leung & Yang, 1995; Silaghi et al., 1998), while the 2D models represents the human body using 2D ribbons or blobs (Ju et al., 1996; Leung & Yang, 1995; Niyogi & Adelson, 1994). 3D models are used to depict the human body structure in a more detailed manner using cones, cylinders, spheres, ellipses etc. (Kakadiaris & Metaxas, 1996; Rehg & Kanade, 1995; Rohr, 1994; Wachter & Nagel, 1997).

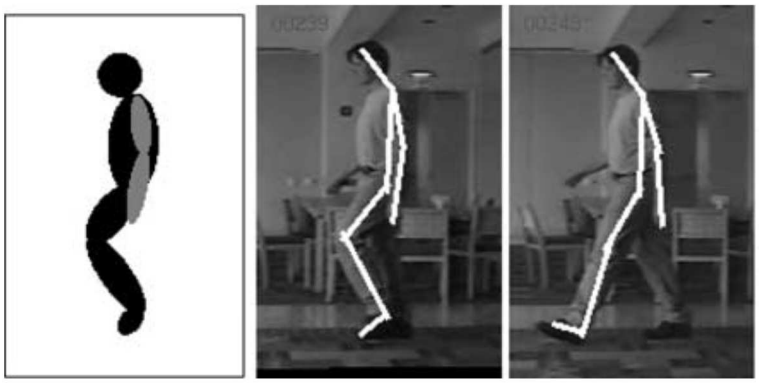
However, tracking human in video sequences is not an easy task. The human body has a complex non-rigid structure consisting of a number of joints (e.g. the leg is connected to the foot by the ankle joint) and each body part can therefore move in a high degree of freedom around its corresponding joints. This often results in self-occlusions of the body parts. 3D models are able to handle such scenarios, but there are other factors that can affect the tracking performance such as the monotone clothes, cluttered background and changing brightness Ning et al. (2004). Therefore, the fuzzy approaches such as the fuzzy qualitative kinematics, the fuzzy voxel person, and the fuzzy shape estimation are explored in the model based human motion tracking algorithms to handle the uncertainties.

Fuzzy qualitative kinematics

A variety of works in the model based human motion tracking have employed the kinematic chain (Guo et al., 1994; Iwai et al., 1999; Ju et al., 1996; Kakadiaris & Metaxas, 1996; Leung & Yang, 1995; Niyogi & Adelson, 1994; Rehg & Kanade, 1995; Rohr, 1994; Silaghi et al., 1998; Wachter & Nagel, 1997). Bregler et al. Bregler et al. (2004) demonstrated a comprehensive visual motion estimation technique using the kinematic chain in a complex video sequence, as depicted in Figure 2.10. However, the crisp representation



(a)

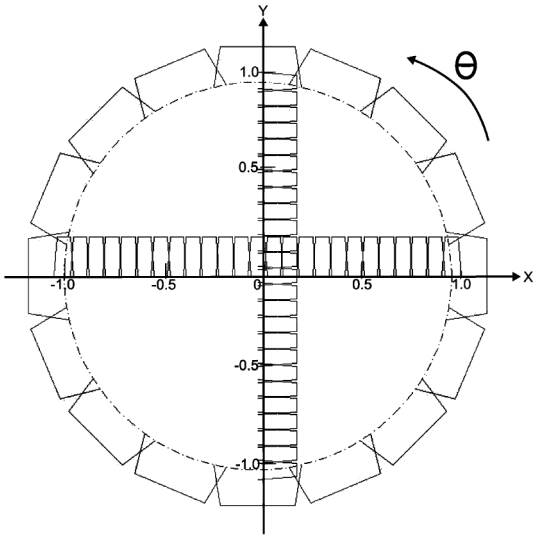


(b)

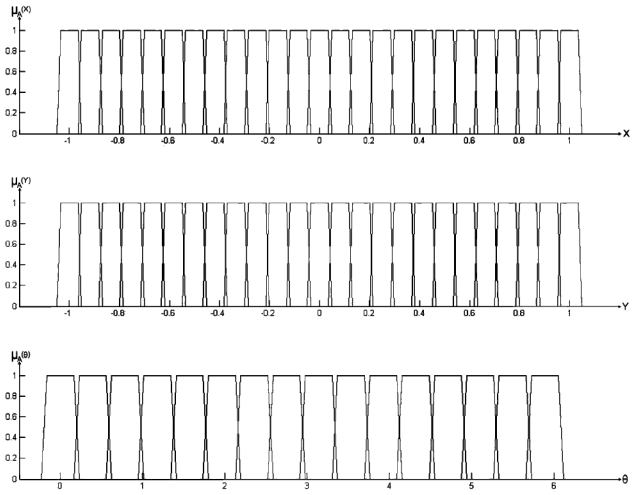
Figure 2.10: (a) Kinematic chain defined by twist (Bregler et al., 2004), and (b) The estimated kinematic chain on the human body while performing the walking action.

of the kinematic chain has a limitation. It suffers from the precision problem (Liu, 2008) and the cumulative errors can directly affect the performance of the higher level tasks. Therefore, a better strategy is required to model the kinematic chain, and to this end, the fuzzy qualitative kinematics has been proposed.

To begin with, the fuzzy qualitative reasoning (Chan et al., 2011; Shen & Leitch, 1993) is a form of approximate reasoning that can be defined as the fusion between the fuzzy set theory (Zadeh, 1965) and the qualitative reasoning (Kuipers, 1986). The qualitative reasoning operates with the symbolic 'quantities', while the fuzzy reasoning reasons with the fuzzy intervals of varying precisions, providing a means to handle the uncertainty in a natural way. Therefore, the fuzzy qualitative reasoning incorporates the advantages of both the approaches to alleviate the hard boundary or the crisp values of the ordinary



(a)



(b)

Figure 2.11: (a) Description of the Cartesian translation and the orientation in the conventional unit circle replaced by the fuzzy quantity space. (b) Element of the fuzzy quantity space for every variable (translation (X, Y), and orientation θ) in the fuzzy qualitative unit circle is a finite and convex discretization of the real number line Chan & Liu (2009).

measurement space. For instance, Liu et al. (2009) applied this in the Fuzzy Qualitative Trigonometry (Figure 2.11) where the ordinary Cartesian space and the unit circle are substituted with the combination of membership functions yielding the fuzzy qualitative coordinate and the fuzzy qualitative unit circle. Extension from this, a fuzzy qualitative representation of the robot kinematics (Liu, 2008; Liu et al., 2008a) was proposed. The work presented a derivative extension to the Fuzzy Qualitative Trigonometry Liu et al. (2009). Motivated by these approaches, Chan et al. (2008) proposed a data quantization

process based on the Fuzzy Qualitative Trigonometry to model the uncertainties during the kinematic chain tracking process; and subsequently constructed a generic activity representation model.

Fuzzy voxel person

Figure 2.12: Voxel person constructed using multiple cameras from different viewpoints of the silhouette images that resolved the occlusion problem in the single camera system. However, due to the location of the cameras and the person's positions, the information gathered using the crisp voxel person model can be imprecise and inaccurate. Therefore, the fuzzy voxel person representation was proposed Anderson, Luke III, et al. (2009).

As aforementioned, the 3D models provide more useful information than the 2D models as the features (height, centroid, orientation, etc.) in the 3D space are camera-view independent. Inspired by this, Anderson, Luke, et al. (2009a,b) demonstrated a method to construct a 3D human model in voxel (volume element) space using the human silhouette images called the voxel person (Figure 2.12). However, due to the location of the cameras and the object's positions, the gathered information using the crisp voxel person model can be sometimes imprecise and inaccurate. The crisp technique works well if and only if there are sufficient number of cameras. But unfortunately, it is hard to find more than a couple of cameras in the same area due to the high cost involved and the limited space area.

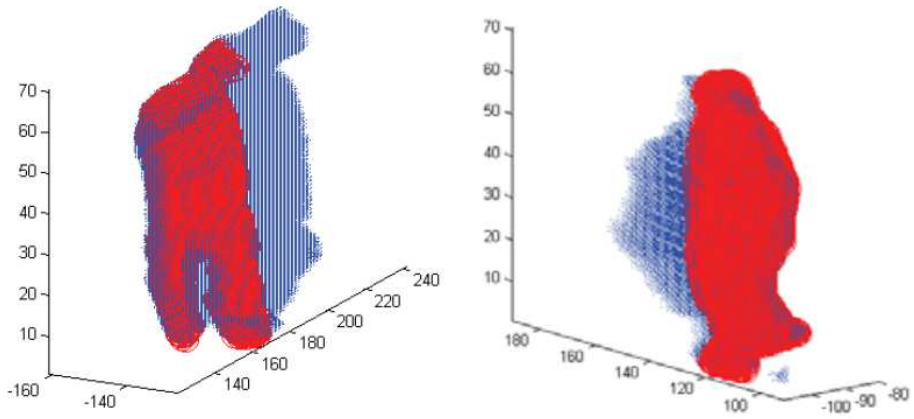


Figure 2.13: The proposed fuzzy voxel person to obtain an improved crisp object. Red areas are the improved voxel person and the blue areas are the rest of the original crisp voxel person Anderson, Luke III, et al. (2009). This picture is best viewed in colors.

Therefore, fuzzy voxel person was utilized in Anderson, Luke III, et al. (2009) by employing only a few cameras and a minimal prior knowledge about the object. The FIS was used to determine the membership degree of the voxel person, reflecting how likely it belongs to the actual object. Extreme body joints viewing conditions were taken into account and it was observed that the fuzzy acquired results were much better than the crisp approach, both qualitatively (as shown in Figure 2.13) as well as quantitatively (Anderson, Luke III, et al., 2009). This concept of the fuzzy voxel person was incorporated in a number of works (Anderson, Luke, et al., 2009a,b).

Fuzzy shape estimation

The regions of interest extracted from the background subtraction algorithm are normally represented using different shape models, such as ribbons and blobs for 2D images, while cones, cylinders, spheres, ellipses etc. for the 3D images. Here, we will concentrate mainly on the blob representation. For a tracking system with reliance on the shape estimation, problems arise due to imperfect image segmentation techniques. This is because of the image irregularities, shadows, occlusions, etc. that results in multiple blobs generation for a single object. Besides that, in the multiple objects tracking, recovering from the

overlapping regions is a big challenge. In order to solve this, García et al. (2002); Garcia et al. (2011) applied FIS to update both the trajectories and the shape estimated for the targets with a set of image regions. These image regions are represented using the blobs extracted from each frame. Following the general steps of the FIS, heuristic features were extracted from the detected blobs, and used as inputs to the FIS to assess the confidence values assigned to each blob to update the estimators describing the targets' shape and the tracks. With this, the tracking can be locked if the confidence of the target shape is low. This is to prevent the tracking to deviate from the real path caused by the cumulated errors such as the uncertain shape. The tracking resumes once the confidence of the object shape is high.

2.2.2 (b) Non-model based tracking

In non-model based tracking, the objects detected are represented using the random dispersed points instead of the rigid shape models (e.g. stick figure, blob, cylinder, etc.). The association amongst the points that contribute to the motion tracking are based on the hypothesis which takes into account the object's characteristics and behavior. This is a complex problem to be formulated because of the presence of occlusions, misdetections, new object entries etc. that may lead to permanent tracking error. Fuzzy approaches such as the fuzzy Kalman filter, fuzzy particle filter, fuzzy optical flow and fuzzy clustering are widely employed in the non-model based object tracking, where they explicitly take into account the uncertainties to establish the point correspondence between the object motions.

Fuzzy Kalman filter

Kalman filter, the popular optimal estimator capable of operating recursively on the streams of noisy input data (Kalman, 1960), is a popular choice for tracking a moving

object. It has been successfully applied in several previous works on the human motion tracking (Kakadiaris & Metaxas, 1996; Kohler, 1997; Marins et al., 2001; G. F. Welch, 2009; Yun et al., 2005; Yun & Bachmann, 2006). There are three basic steps involved in the Kalman filtering for human motion tracking: initialization, prediction and correction (G. Welch & Bishop, 1995). Often the complex dynamic trajectories due to the changes in the acceleration of human motion are not feasible to be modeled by the linear systems. Therefore, instead of the basic Kalman filters, the Extended Kalman filters are used which are capable of modeling the non-linear states. However, all these Kalman filtering algorithms suffer from the divergence problem if the theoretical behavior of a filter and its actual behavior do not agree. The divergence due to modeling errors is a critical issue in the Kalman filtering process.

In order to solve this, the FIS was adopted in the Kalman filtering (G. Chen et al., 1998; Kobayashi et al., 1998; Sasiadek & Khe, 2001; Sasiadek & Wang, 1999; Sasiadek et al., 2000; Senthil et al., 2006) to detect the bias of measurements and prevent the divergence. The new Kalman filter is called as the fuzzy adaptive Kalman filter. Takagi-Sugeno fuzzy model is used to detect the divergence and the uncertainty of the parameters in the Kalman filter such as the covariance and the mean value are modeled as membership function with the corresponding fuzzy rules for inference. To this extent, P. Angelov et al. (2008) proposed the evolving Takagi-Sugeno fuzzy model (P. Angelov & Filev, 2005; P. P. Angelov & Filev, 2004) which can be seen as the fuzzy weighted mixture of the Kalman filter for object tracking in the video streams, and the performance is better than the ordinary Kalman Filter.

Fuzzy particle filter

Similar to the Kalman filters, particle filters offer a good way to track the state of a dynamic HMA system. In general, if one has a model of how the system changes

with time, and possible observations made in particular states, the particle filters can be employed for tracking. However, as compared to the Kalman filters, the particle filters offer a better tracking mechanism as they provide multiple predictions or hypothesis (i.e. as many as hypothesis as the number of particles) to recover from the lost tracks, which helps to overcome the problems related to the complex human motion. One must note that there is a tradeoff between system precision and computational cost in the particle filter framework, i.e. more number of particles improves the system precision, but also increases the computational cost and vice versa.

As a remedy to the above mentioned problems, a new sequential fuzzy simulation based particle filter was proposed in H. Wu et al. (2008) to estimate the state of a dynamic system with noises described as fuzzy variables using the possibility theory. In most of the current particle filtering algorithms, the uncertainty of the tracking process and the measurement of noises are expressed by the probability distributions, which are sometimes hard to construct due to the lack of statistical data. Therefore, it is more suitable to compute the possibility measure using the fuzzy set theory for modeling the uncertain variables with imprecise knowledge. H. Wu et al. (2008) found that their proposed fuzzy logic based particle filter outperforms the traditional particle filter even when the number of particles is small. Another variant of this work is Yoon et al. (2013), where an adaptive model is implemented in the fuzzy particle filter with the capability to adjust the number of particles by using the result from the measurement step, and improve the speed of an object tracking algorithm. Apart from that, Chan & Liu (2009); Chan et al. (2008) handled the tradeoff between the system precision and the computational cost by employing data quantization process that utilizes the Fuzzy Quantity Space (Liu et al., 2009). In general, the work quantize the particles into finite fuzzy qualitative states. As such, the system able to model the offset of the tracking errors, while retaining the precision when relatively low number of particles are selected to perform the tracking task. Last but not

the least, the FIS has also contributed in the particle filters (Kamel & Badawy, 2005; Y.-J. Kim et al., 2007) and achieved better accuracy with lower computational cost.

Fuzzy optical flow

Optical flow (Beauchemin & Barron, 1995; Horn & Schunck, 1981) is another popular motion tracking algorithm. It is an efficient technique for approximating the object motion in two consecutive video frames by computing the intensity variations between them. However, the removal of the incoherent optical flow field is still a great challenge. This is because the incoherent regions can be treated as random noises in the optical flow field due to the sources of disturbances in a natural scene (e.g. dynamic background). Fuzzy hostility index was introduced in Bhattacharyya & Maulik (2013); Bhattacharyya et al. (2009) to overcome this issue and thus improving the time efficiency of the flow computation. The fuzzy hostility index (Bhattacharyya et al., 2007) measures the amount of homogeneity or heterogeneity of the neighborhood pixel in the optical flow field. The more homogeneous is the neighborhood of a pixel, the less is the pixel hostile to its neighbor. This implies that a denser neighborhood indicates a more coherent optical flow neighborhood region. To deal with the uncertain conditions, soft computing is applied where the hostility index computed from the neighborhood pixels is represented as a fuzzy set, where the membership values lie between 0 and 1. This method has shown the capability to track fast moving objects from the video sequences efficiently.

Fuzzy clustering

Clustering is an unsupervised machine learning solution that learns the unlabeled data by grouping the similar ones into the corresponding groups autonomously. Inspired from this, multi-object cluster trackings (Heisele et al., 1997; Pece, 2002) were introduced with the belief that the moving targets always produce a particular cluster of pixels with

similar characteristics in the feature space, and the distribution of these clusters changes only little between the consecutive frames. Xie et al. (2004) proposed a fast fuzzy c-means (FCM) clustering tracking method which offers a solution towards the high complexity and the computational cost involved in the conventional methods on multi-object tracking, and also the hard clustering algorithms such as the k-means that causes failure in the case of severe occlusions and pervasive disturbances. FCM is also recognized as the soft clustering algorithm where it applies data partition to allocate each sample data into more than one clusters with the corresponding membership values which is more meaningful and stable than the hard clustering algorithms. In Xie et al. (2004), the component quantization filtering was incorporated with FCM to provide faster processing speed. Table 2.4 summarizes the intuition of using the fuzzy approaches in MiL HMA.

Table 2.4: Summarization of research works in MiL HMA using the Fuzzy approaches.

MiL processing	Problem statements / Sources of Uncertainty	Authors	Why fuzzy?	Approach
Model based tracking	Crisp representation of the kinematic chain suffers from the precision problem, and the cumulative errors can directly affect the performance of the tracking process.	Chan & Liu (2009); Chan et al. (2008); Liu et al. (2008a,b, 2009)	Integration of the fuzzy set theory and the fuzzy qualitative reasoning in the kinematic chain representation provides a means of handling the uncertainty in a natural way. Fuzzy qualitative kinematics solves the precision problem by eliminating the hard boundary problem in the measurement space that can tolerate the offset errors.	Fuzzy qualitative kinematics
	Due to the location of cameras and object's positions, the information gathered using crisp voxel person model can be imprecise and inaccurate. Crisp approach works fine in multi-camera environment, but it is not feasible due to high cost and limited space.	Anderson, Luke, et al. (2009a,b); Anderson, Luke III, et al. (2009)	Fuzzy voxel person is able to model different types of uncertainties associated with the construction of the voxel person by using the membership functions, employing only a few cameras and a minimal prior knowledge about the object.	Fuzzy voxel person
	In shape based (blob) tracking, the imperfect image segmentation techniques result in multiple blobs generation for a single object because of the image irregularities, shadows, occlusions, etc. While in the multiple object tracking, recovering from the overlapping regions is a big challenge.	García et al. (2002); Garcia et al. (2011)	FIS is applied to perform the fuzzy shape estimation to achieve a better tracking performance by taking into account the uncertainty in shape estimation. If the shape is uncertain, the tracking will be locked and it will be recovered once the confidence becomes higher. This is to prevent the tracking errors caused by the uncertain shapes.	Fuzzy shape estimation
Non-model based tracking	Conventional Kalman filter algorithms suffer from the divergence problem and it is difficult to model the complex dynamic trajectories.	Aggarwal & Cai (1997); G. Chen et al. (1998); Gavrila (1999); Hu, Tan, et al. (2004); I. S. Kim et al. (2010); Ko (2008); Kobayashi et al. (1998)	Fuzzy Kalman filters are capable of solving the divergence problem by incorporating the FIS, and are more robust against the streams of random noisy data inputs.	Fuzzy Kalman filter
	Particle filters suffer from the tradeoff between the accuracy and computational cost as its performance usually relies on the number of particles. This means more number of particles will improve the accuracy, but at the same time increases the computational cost.	Chan & Liu (2009); Chan et al. (2008); Kamel & Badawy (2005); Y.-J. Kim et al. (2007); H. Wu et al. (2008); Yoon et al. (2013)	The fuzzy particle filter effectively handles the system complexity by compromising the low number of particles that were used while retaining the tracking performance.	Fuzzy particle filter
	Random noises in optical flow field due to the sources of disturbances in a natural scene (e.g. dynamic background) affects the tracking performance.	Bhattacharyya & Maulik (2013); Bhattacharyya et al. (2009)	Fuzzy hostility index is used in the optical flow to filter the incoherent optical flow field containing random noises in an efficient manner.	Fuzzy optical flow
	In the conventional methods for multi-object tracking, hard clustering tracking algorithms such as the K-means are used, and involve high complexity and computational cost. Also, they fail in the case of severe occlusions and pervasive disturbances.	Xie et al. (2004)	FCM tracking algorithm offers more meaningful and stable performance by using soft computing techniques. The integration of component quantization filtering with FCM tracking algorithm provides faster processing speed.	Fuzzy clustering

2.2.3 High-level

The final aim of the HMA system is to perform the human behavior understanding. This level can be extended into several processes, for examples, gesture recognition, describing an activity, and reacting to an event. All these processes are usually incorporated into a real-time system to assist humans in specific tasks such as surveillance purposes, industrial applications, human-computer interaction, military action and robotics mission (Aggarwal & Ryoo, 2011; Cristani et al., 2013; Gavrila, 1999; Hu, Tan, et al., 2004; Moeslund & Granum, 2001; Turaga et al., 2008; L. Wang et al., 2003). LoL and MiL HMA serve as the preliminary steps for this level. In HiL, this review is regarding the

feasibility of the fuzzy approaches to achieve better performance with emphasis on: (a) hand gesture recognition, (b) activity recognition, (c) style invariant action recognition, (d) multi-view action recognition, and (e) anomaly event detection.

2.2.3 (a) Hand gesture recognition

Gesture recognition aims at recognizing meaningful expressions of the human motion, involving the hands, arms, face, head, or body. The applications of gesture recognition are manifold (Lyons et al., 1999), ranging from the sign language to medical rehabilitation and virtual reality. The importance of gesture recognition lies in building efficient and intelligent human-computer interaction applications (Y. Wu & Huang, 1999) where one can control the system from a distance for a specific task, i.e. without any cursor movements or screen touching. Besides that, nowadays, there exists successful commercialized gesture recognition devices such as the Kinect: a vision-based motion sensing device, capable of inferring the human activities. Unfortunately, in a gesture recognition system, the complex backgrounds, dynamic lighting conditions and sometimes the deformable human limb shapes can lead to high level of uncertainties and ambiguities in recognizing the human gestures. Also, "pure" gestures are seldom elicited, as people typically demonstrate "blends" of these gestures (Mitra & Acharya, 2007). Among all the solutions, the fuzzy clustering algorithms and the integration of fuzzy approaches with machine learning methods are often incorporated to deal with such difficult situations and achieve better system performance. In this section, we review the relevant works with emphasis on the hand gesture recognition.

Fuzzy clustering

Among the well-known clustering techniques are K-means, GMM, hierarchical model, and FCM. However, in the probabilistic based clustering algorithms (e.g. K-means,

GMM, and hierarchical model), the data allocation to each cluster is done in a crisp manner, that is each data element can belong to exactly one cluster. In contrast, the fuzzy clustering algorithm (e.g. FCM), soft computing is applied in the sense that the data partition alleviates the data allocation where each data can belong to more than one clusters and associated with a set of membership values. This solution works better in the challenging environments such as the complex backgrounds, dynamic lighting conditions, and the deformable hand shapes with real-time computational speeds (X. Li, 2003; Verma & Dev, 2009; J. Wachs et al., 2002; J. P. Wachs et al., 2005).

Using the FCM, J. Wachs et al. (2002); J. P. Wachs et al. (2005) worked on a fast respond telerobotic gesture-based user interface system. The nature of FCM in relaxing the hard decision allowed the use of smaller portions of the training set and thus shorter training time was required. Empirically, it has proved to be sufficiently reliable and efficient in the recognition tasks with the achievement on high accuracy and real-time performance. X. Li (2003) further improved the work by J. Wachs et al. (2002) in the skin segmentation problem using the color space to solve the skin color variation. Besides spatial information, temporal information is also important in the gesture inference process. In Verma & Dev (2009), the spatial information of hand gesture using the FCM was trained in order to determine the partitioning of the trajectory points into a number of clusters with the fuzzy pseudo-boundaries. In general, each trajectory point belongs to each cluster specified by a membership degree. Then, the temporal data is obtained through the transitions between the states (cluster of trajectory points) of a series of finite state machines to recognize the gesture motion.

Hybrid technique

A few works (Al-Jarrah & Halawani, 2001; Binh & Ejima, 2005; Várkonyi-Kóczy & Tusor, 2011) on fusing the fuzzy approaches with machine learning solutions have been

reported in the gesture recognition. Al-Jarrah & Halawani (2001) used the adaptive neuro-fuzzy inference system to recognize the gestures in Arabic sign language. This work was motivated by the transformation of human knowledge into a FIS, but does not produce the exact desired response due to the heuristic or non-sophisticated membership functions and the fuzzy rules generation. Thus, there was a need to fine-tune the parameters in the FIS to enhance its performance, and the adaptive neuro-fuzzy inference system provided this flexibility by applying a learning procedure using a set of training data.

Binh & Ejima (2005) introduced a new approach towards gesture recognition based on the idea of incorporating the fuzzy ARTMAP (Carpenter et al., 1992) in the feature recognition neural network (Hussain & Kabuka, 1994). The proposed method reduced the system complexity and performed in real-time manner. Nonetheless, Várkonyi-Kóczy & Tusor (2011) presented an approach with several novelties and advantages as compared to other hybrid solutions. They introduced a new fuzzy hand-posture model using a modified circular fuzzy neural network architecture to efficiently recognize the hand posture. As a result, the robustness and reliability of the hand-gesture identification was improved, and the complexity and training time involved in the neural networks was significantly reduced.

2.2.3 (b) Activity recognition

Activity recognition is an important task in the HiL HMA systems. The goal of activity recognition is to autonomously analyze and interpret the ongoing human activities and their context from the video data. For example, in the surveillance systems for detecting suspicious actions, or in sports analysis for monitoring the correctness of the athletes' postures. In recent times, the fuzzy approaches such as type-1 FIS, fuzzy HMM, and hybrid techniques have proved to be beneficial in the human activity recognition, with capability of modeling the uncertainty in the feature data. Nonetheless, Fuzzy Vector

Quantization (FVQ) and Qualitative Normalized Template (QNT) provide the capability to handle the complex human activities occurring in our daily life such as walking followed by running, then running followed by jumping, or a hugging activity where two or more people are involved. In this section, the applications of these fuzzy approaches in the activity recognition will be discussed.

Type-1 fuzzy inference system

The FIS can be efficiently used to distinguish the human motion patterns and recognize the human activities with its capability of modeling the uncertainty and the fusion of different features in the classification process. In the literature of activity recognition, there exists some works (Le Yaouanc & Poli, 2012; Yao et al., 2014) that employed the FIS to classify different human activities.

Both Le Yaouanc & Poli (2012); Yao et al. (2014) took into account the uncertainties in both the spatial and temporal features for efficient human behavior recognition. Their method aims at handling high uncertainty levels and the complexities occurring in the real world applications. Le Yaouanc & Poli (2012) used the spatial and temporal geometry features to study the importance of the spatio-temporal relations such as '*IsMoving*', '*IsComingCloseTo*', '*IsGoingAway*', '*IsGoingAlong*' with the objective to provide a qualitative interpretation of the behavior of an entity (e.g. a human) in real-time. Another work Yao et al. (2014) adopted the spatio-temporal features such as the silhouette slices and the movement speed in video sequences as the inputs to the FIS. Extra merit in this work is that they learn the membership functions of the FIS using the FCM which prevents the intervention of human in generating the fuzzy membership function heuristically.

Hybrid technique

Owing to the demands of the development of enhanced video surveillance systems

that can automatically understand the human behaviors and identify dangerous activities, Acampora et al. (2012) introduced a semantic human behavioral analysis system based on the hybridization of the neuro-fuzzy approach. In their method, the kinematic data obtained from the tracking algorithm is translated into several semantic labels that characterizes the behaviors of various actors in a scene. To achieve this, the behavioral semantic rules were defined using the theory of time delay neural networks and the fuzzy logic, to identify a human behavior analyzing both the temporal and the contextual features. This means that they analyze how a human activity changes with respect to time along with how it is related to the contexts surrounding the human. Their hybrid method outperformed other approaches and showed high level of scalability and robustness.

Another work Hosseini & Eftekhari-Moghadam (2013) presented a fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video, integrating the Decision Tree with the FIS. A flexible system was designed using the fuzzy rules, that can be used with least reliance on the predefined feature sequences and domain knowledge. The FIS was designed as a classifier taking into account the information from a set of audio-visual features as its crisp inputs and generate the semantic concepts corresponding to the events occurred. From the fuzzification of the feature vectors derived from the training data, a set of tuples were created, and using the Decision Tree, the hidden knowledge among these tuples as well as the correlation between the features and the related events were extracted. Then, traversing each path from the root to the leaf nodes of the Decision Tree, a set of fuzzy rules were generated which were inserted in the knowledge base of the FIS and the occurred events were predicted from the input video (i.e. soccer video) with good accuracy.

Fuzzy vector quantization

In order to learn the complex actions, Gkalelis et al. (2008) represented the human

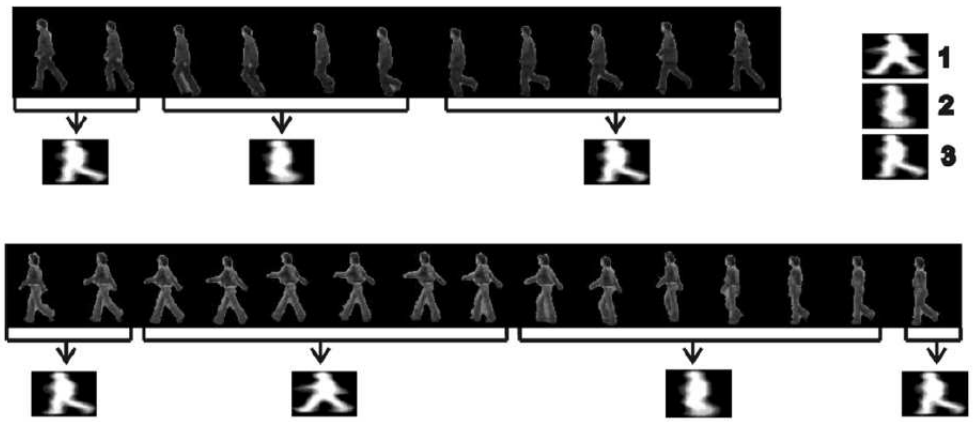


Figure 2.14: Movements of running (top) and walking (bottom) activities, as well as the associated dynemes which are learned from the FCM Gkalelis et al. (2008).

movements as a combination of the smallest constructive unit of human motion patterns called the dyneme (Figure 2.14). It is the basic movement patterns of a continuous action. In the bottom of action hierarchy, dyneme is defined as the smallest constructive unit of human motion; while one level above is the movement which is perceived as a sequence of dynemes with clearly defined temporal boundaries and conceptual meaning. Dyneme can be learned in an unsupervised manner and in Gkalelis et al. (2008), the FCM was chosen. Then, fuzzy vector quantization (FVQ) Karayiannis & Pai (1995) as a function that regulates the transition between the crisp and the soft decisions was employed to map an input posture vector into the dyneme space. Finally, each movement was represented as a fuzzy motion model by computing the arithmetic mean of the comprising postures of a movement in the dyneme space. Their algorithm provides good classification rates and exhibits adequate robustness against partial occlusions, different styles of movement execution, viewpoint changes, gentle clothing conditions and other challenging factors.

Qualitative normalized template

Utilizing the concept of fuzzy qualitative robot kinematics (Liu, 2008; Liu et al., 2008a), Chan & Liu (2009) and (Chan et al., 2008) built a generative action template, called the Qualitative normalized template (QNT) to perform the human action recogni-

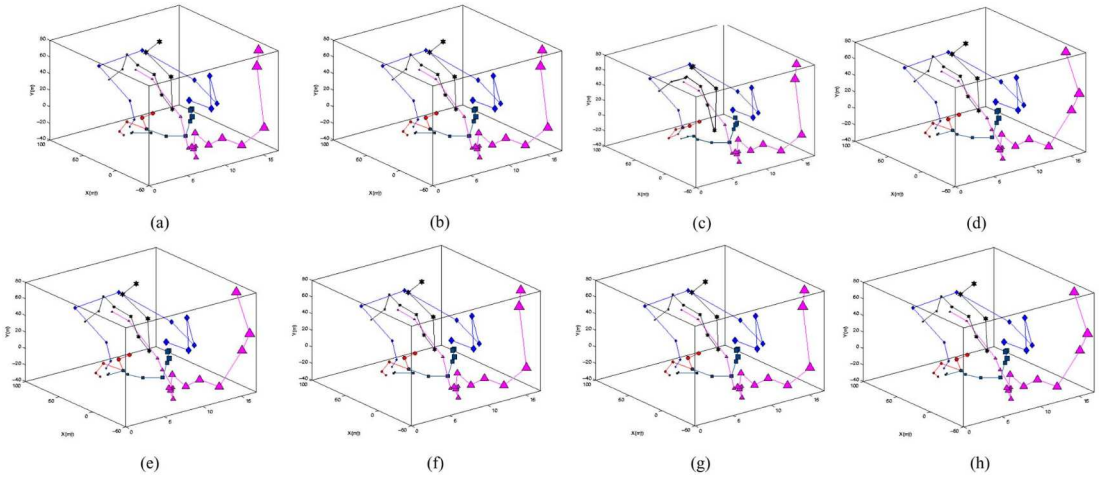


Figure 2.15: Visualization of the QNT model: each of the five activities (walking, running, jogging, one-hand waving (wave1) and two-hands waving(wave2)) from eight subjects (a)-(h) in the quantity space Chan & Liu (2009).

tion. First of all, the training data that represents a typical activity is acquired by tracking the human anatomical landmarks in the image sequences. In their work, a data quantization process was employed to handle the tradeoffs between the tracking precision and the computational cost. Then, the QNT as illustrated in Figure 2.15 was constructed according to the fuzzy qualitative robot kinematics framework (Liu, 2008; Liu et al., 2008a). An empirical comparison with the conventional hidden Markov model (HMM) and fuzzy HMM using both the KTH and the Weizmannn datasets has shown the effectiveness of the proposed solution (Chan & Liu, 2009).

Fuzzy Hidden Markov Model

Hidden Markov model (HMM) (Elliott et al., 1995) is the statistical Markov model with the state being not directly visible, but the output that is dependent on the state is visible. HMM have been widely employed in the human action recognition (Bobick & Wilson, 1995; Campbell & Bobick, 1995; Oliver et al., 2000; Wilson & Bobick, 1999; Yamato et al., 1992). These works have well demonstrated the modeling and recognition of the complex human activities using HMM. In the training stage of HMM, expectation maximization algorithm is adopted. However, in the conventional HMM, each obser-

vation vector is assigned only to one cluster. Mozafari et al. (2012) pointed out that assigning different observation vectors to the same cluster is possible and if their observation probabilities become the same, consequently, the classification performance may decrease. Therefore, HMM was extended to fuzzy HMM where in the training stage, the distance from each observation vector to each cluster center is computed and the inverse of the distance is considered as the membership degree of the observation vector to the cluster. Mozafari et al. (2012) utilized this concept for human action recognition and the experiment results demonstrate the effectiveness of the fuzzy HMM in human action recognition, with good recognition accuracy for the similar actions such as "walk" and "run".

2.2.3 (c) Style invariant action recognition

A robust action recognition algorithm must be capable of recognizing the actions performed by different person in different styles. Commonly, different person have different styles of executing the same action which can be categorized according to the physical differences (such as human appearances, sizes, postures, etc.) and the dynamic differences (speed, motion pattern, etc.). In order to model such variations, several notable works have been reported incorporating the fuzzy approaches.

Fuzzy vector quantization

Iosifidis et al. (2011) adopted the concept of FVQ and the dyneme, and proposed a novel person specific activity recognition framework to cope with the style invariant problem. The method is mainly divided into two parts: firstly, the ID of the person is identified, and secondly, the activity is inferred from the person specific fuzzy motion model (Gkalelis et al., 2008). It was found that the different styles in action execution endowed the capability to distinguish one person from the another. Therefore, Iosifidis, Tefas, &

Pitas (2012a) developed an activity-related biometric authentication system by utilizing the information of different styles by different people. Improvement was made in the computation of the cumulative fuzzy distances between the vectors and the dynemes that outperforms L_1, L_2, and Mahalanobis distances which were used previously in Gkalelis et al. (2008).

2.2.3 (d) Multi-view action recognition

The capability of multi-view action recognition is emerging as an important aspect for advanced HMA systems. In the real world environment, human are free to perform an action at any angle with no restriction of being frontal parallel to the camera and most of the previous works treat it as a constraint or limitation in their system. This problem has received increasing attention in the HMA research and some of the notable works have been reported (Ji & Liu, 2010; Lewandowski, Makris, & Nebel, 2010; Weinland et al., 2006). Besides that, fuzzy approaches such as the FVQ, and fuzzy qualitative reasoning are also applied in the study of multi-view action recognition which will be discussed in the following subsections.

Fuzzy vector quantization

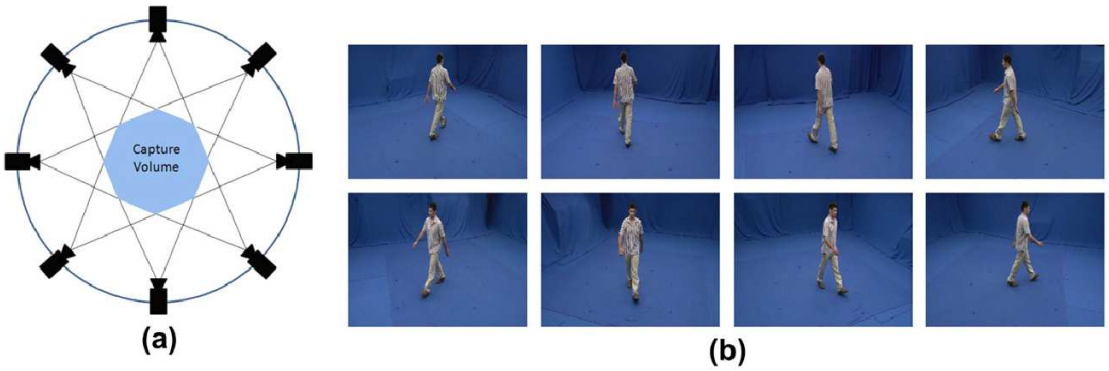


Figure 2.16: (a) A converging eight-view camera setup and its capture volume, and (b) an eight-view video frame Iosifidis, Tefas, Nikolaidis, & Pitas (2012).

Iosifidis, Tefas, Nikolaidis, & Pitas (2012); Iosifidis, Tefas, & Pitas (2012a); Iosifidis et al. (2013) extended Gkalelis et al. (2008) to support multi-view action recognition. The motion patterns obtained from different cameras, as in Figure 2.16, were clustered to determine the number of multi-view posture primitives called the multi-view dynemes. Similar to Gkalelis et al. (2008), FVQ was utilized to map every multi-view posture pattern to create the multi-view dyneme space. This new multi-view fuzzy movement representation is motion speed and duration invariant which generalizes over variations within one class and distinguishes between the actions of different classes. In the recognition step, Fourier view invariant posture representation was used to solve the camera viewpoint identification problem before the action classification was performed. Nonetheless, they tackled the problem of interaction recognition i.e. human action recognition involving two persons (Iosifidis, Tefas, & Pitas, 2012b).

2.2.3 (e) Anomaly event detection

Anomaly detection refers to the problem of finding patterns in the input data that do not conform to the expected behavior. In our daily life, anomaly detection is important to infer the abnormal behavior of a person, such as an action or an activity that is not following the routine or deviated from the normal behavior (Hu, Tan, et al., 2004; Kratz & Nishino, 2009; S. Wu et al., 2010). For example, in the healthcare domain to prevent unfavorable events from occurring such as the risk of falling down of the patients, and in the surveillance systems, to automatically detect the crime activities.

Type-1 fuzzy inference system

As humans gain more knowledge, they are able to make better decisions; similarly if the FIS is provided with sophisticated knowledge (i.e. fuzzy rules), it can deal with the real world problems in a better manner. FIS has been employed in various works

for anomaly event detection such as the elderly fall detection in Anderson et al. (2006); Anderson, Luke, et al. (2009a,b), to address the deficiencies and the inherent uncertainty related to modeling and inferring the human activities. The works emphasized that the non-interpretable likelihood value or the ad-hoc training of the activity models in the conventional approaches is impractical in the area of human action recognition. Therefore, a confidence value (fuzzy membership degree) that can be reliably used to reject unknown activities is more convenient.

Anderson, Luke, et al. (2009b) proposed a novel fuzzy rule based method for monitoring the wellness of the elderly people from the video. In this paper, the knowledge base (fuzzy rules as depicted in Figure 2.17) was designed under the supervision of nurses for the recognition of falls of the elderly people. Under this framework, the rules can be easily modified, added or deleted, based on the knowledge about the cognitive and functional abilities of the patients. This work was an extension of Anderson, Luke, et al. (2009a) where the linguistic summarizations of the human states (three states: upright, on-the-ground and in-between) based on the voxel person and the FIS were extracted, extended using a hierarchy of the FIS and the linguistic summarization for the inference of the patients' activities. Their technique works well for fall detection, but the question is if this framework can be extended to different activities. The answer is yes where, Anderson et al. (2008) extended the work to support the additional common elderly activities such as standing, walking, motionless-on-the-chair, and lying-motionless-on-the-couch, with the inclusion of the knowledge about the real world for the identification of the voxels that corresponds to the wall, floor, ceiling, or other static objects or surfaces. Two new states were included to recognize these activities i.e. on-the-chair and on-the couch. These states were different from the previous three states (upright, on-the-ground and in-between) as they were based on the voxel person interacting with a static object in the scene. Further, the fuzzy rules were extended to six new fuzzy rules designed for

identifying on-the-chair and on-the-couch activities.

Rule		Centroid	Eigen Height	Normal Similarity		Upright	In Between	On the Ground
1		H	H	H		L	V	V
2		M	H	H		L	L	V
3		L	H	H		V	L	L
4		H	M	H		V	H	V
5		M	M	H		V	H	L
6		L	M	H		V	H	H
7		M	L	H		V	L	H
8		L	L	H		V	V	M
9		H	H	M		L	V	V
10	If	M	H	M	Then	L	L	V
11		L	H	M		L	H	V
12		H	M	M		L	H	V
13		M	M	M		L	H	V
14		L	M	M		V	H	L
15		L	M	L		V	L	H
16		L	L	M		V	L	M
17		H	H	L		H	V	V
18		M	H	L		M	V	V
19		L	H	L		L	L	V
20		H	M	L		M	L	V
21		M	M	L		L	L	V
22		L	M	L		L	H	V
23		M	L	L		V	H	L
24		L	L	L		V	L	H

Figure 2.17: Rule table of the human states (Upright, In Between, On the Ground) with V=Very low, L=Low, M=Medium, and H=high which are used to infer the human activities Anderson, Luke, et al. (2009b).

Fuzzy one class support vector machine

The fuzzy one class support vector machine (FOCSVM) is an efficient algorithm often used in fall detection systems to distinguish a falling from other activities such as walking, bending, sitting or lying. Yu et al. (2011) proposed a robust fall detection system using FOCSVM with novel 3D features. In their method, a voxel person was first computed, then the video features obtained from the variation of a persons' 3D angle and centroid information were extracted from the sequences of voxel persons which were used to train the FOCSVM classifier. As compared to the traditional one class support vector machine, FOCSVM obtained more accurate fall detection result with tight decision boundaries under a training dataset with outliers. The success of the proposed method is evident from the experiments on the real video sequences, with less non-fall samples being misclassified as falls by the classifier with imperfect training data.

Fuzzy clustering

In order to perform fall detection in multiple camera framework, fuzzy clustering algorithms (e.g. FCM, Gustafson and Kessel Clustering, or Gath and Geva Clustering) along with the fuzzy K-nearest neighbor algorithms were employed in Wongkhuenkaew et al. (2013). In particular, Hu moment invariant features were computed from the 2D silhouette images and principal component analysis was utilized to select the principal components. The fuzzy clustering algorithms were used to generate the multi-prototype that represent the action classes such as standing or walking, sitting or bending, lying and lying forward. Fuzzy K-nearest neighbor was then used to deduce the corresponding action classes. For example, if the detected action was "lying" or "lying forward", it was considered as the falling activity.

Hybrid technique

A hybrid model of the FIS and the Fuzzy Associative Memory (FAM) was incorporated in Z. Wang & Zhang (2008), which basically receives an input and assigns a degree of belongingness to a set of rules. Z. Wang & Zhang (2008) considered the angles of human limbs as the inputs to the FAM with three rules defining the abnormal movement types. FAM then assigns a degree of membership to each rule and determines the anomalous or normal events based on a specific threshold. Juang & Chang (2007) also used the neural fuzzy network hybrid model, compensating the lacking of the learning ability of the fuzzy approaches to recognize human poses (e.g. standing, bending, sitting, and lying). Their system with simple fuzzy rules is capable of detecting the emergencies caused by the accidental falls or when a person remains in the lying posture for a period of time. The works evidently show the flexibility of the fuzzy approaches in the alteration or extension of its knowledge base to adapt to newly encountered real world problems.

Hu, Xie, et al. (2004) proposed fuzzy self-organizing neural network (fuzzy SOM)

Table 2.5: Summarization of research works in HiL HMA using the Fuzzy approaches.

HiL processing	Problem statements / Sources of Uncertainty	Authors	Why fuzzy?	Approach
Hand gesture recognition	Complex backgrounds, dynamic lighting conditions and sometimes deformable human limbs' shape leads to ineffective clustering outcome with the conventional crisp clustering algorithms.	X. Li (2003); Verma & Dev (2009); J. Wachs et al. (2002); J. P. Wachs et al. (2005)	FCM relaxes the learning and recognition of gesture by using soft computing technique. This reduces the errors caused by the crisp decisions and increases the system efficiency.	Fuzzy clustering
	Difficulty in determining the optimum parameters in the fuzzy system such as membership function or the threshold value for the decision making in gesture recognition algorithms.	Al-Jarrah & Halawani (2001); Binh & Ejima (2005); Várkonyi-Kóczy & Tusor (2011)	Integration of the fuzzy approaches with machine learning algorithms help in learning the important parameters for the fuzzy system adaptively based on the training data.	Hybrid technique
Activity recognition	The uncertainty in the feature data affects the performance of human activity recognition.	Le Yaouanc & Poli (2012); Yao et al. (2014)	FIS effectively distinguishes the human motion patterns and activity recognition with its flexibility in customizing the membership functions and the fuzzy rules with tolerance to the vague feature data.	Type-1 FIS
	Difficult to determine the optimum membership functions and the fuzzy rules in the FIS for human activity recognition.	Acampora et al. (2012); Hosseini & Eftekhari-Moghadam (2013)	Integration of fuzzy logic with machine learning techniques allows the generation of the optimum membership function and fuzzy rules to infer the human behavior.	Hybrid technique
	Solving continuous human movements or complex activities over time is a difficult problem. For instance, walk then run. Most of the state-of-the-art methods assumed the activity to be uniform and simple.	Gkalelis et al. (2008)	Fuzzy Vector Quantisation (FVQ) incorporated with FCM is used to model the human movements and provides the flexibility to support complex continuous actions.	FVQ
	The usage of sophisticated tracking algorithms in the action recognition suffers from the tradeoff between the computational cost and accuracy.	Chan & Liu (2009); Chan et al. (2008, 2010)	Qualitative Normalized Template (QNT) relaxes the complexity of the representation of the human joints that uses sophisticated tracking algorithms, achieving the efficiency and robustness in complex activity recognition.	QNT
	Conventional Hidden Markov Model (HMM) is unable to model the uncertainties in the training stage which reduces the classification performance.	Mozafari et al. (2012)	Fuzzy HMM models apply soft computing in the training stage which effectively increases the performance in the classification of similar actions such as "walk" and "run".	Fuzzy HMM
Style invariant action recognition	A similar action can be performed with different styles by different person that causes difficulty in the learning and recognition process.	Iosifidis et al. (2011); Iosifidis, Tefas, & Pitas (2012a)	Style invariant action recognition can be achieved by using person specific fuzzy movement model which is trained using FVQ.	FVQ
Multi-view action recognition	Humans are not restricted to perform an action at a fixed angle from the camera.	Iosifidis, Tefas, Nikolaidis, & Pitas (2012); Iosifidis, Tefas, & Pitas (2012a,b); Iosifidis et al. (2013)	Multi-view posture patterns are generated by utilizing FVQ to build a multi-view fuzzy motion model in order to support view invariant human action recognition.	FVQ
Anomaly event detection	The difficulty of extension of a framework to deal with new issues and support new activities.	Anderson et al. (2006, 2008); Anderson, Luke, et al. (2009a,b)	FIS is flexible in customization where the knowledge base (fuzzy rules) can be modified, added, or removed to adapt to various situations such as falling activities.	Type-1 FIS
	The imperfect training data (e.g. some samples would be outliers) affect the classification performance in the fall detection system.	Yu et al. (2011)	FOCSVM is used to reflect the importance of every training sample, by assigning each training data with the membership degree. With this, a good accuracy and decision boundaries are obtained under a training dataset with outliers.	FOCSVM
	Most of the existing elderly fall detection systems are performed in the single camera environment which provides limited information for the inference process.	Wongkhuenkaew et al. (2013)	Fuzzy clustering algorithms (e.g. FCM, Gustafson and Kessel Clustering, or Gath and Geva Clustering) incorporated with Hu moment invariant features and principle component analysis were employed to learn the multi-prototype action classes in the multiple camera environment.	Fuzzy clustering
	Difficulty in determining the optimum parameters in the fuzzy system.	Hu, Xie, et al. (2004); Juang & Chang (2007); Z. Wang & Zhang (2008)	Integration of the fuzzy approaches with machine learning algorithms allows the learning of optimum fuzzy membership functions and fuzzy rules that can adapt to newly encountered problems.	Hybrid technique

to learn the activity patterns for anomaly detection in visual surveillance. Their method aims at automatically constructing the activity patterns by self-organizing learning instead of predefining them manually. Traditionally, individual flow vectors were used as inputs to the neural networks. In the proposed method, whole trajectory was taken as an input, simplifying the structure of the neural networks to a great extent. Fuzzy SOM further improved the learning speed and accuracy of the anomaly detection problem, as demonstrated with the support of experimental results. To understand better, a summary of research works in HiL HMA using the fuzzy approaches is shown in Table 2.5.

2.3 Motivation to the propose works

First, the practice of using multi-camera approaches in view independent HMA as mentioned in the literature (section 2.1) is a popular issue. Motivated from this, view specific action recognition framework that uses single camera is proposed. In the framework, fuzzy qualitative reasoning is adopted to cope with the size and viewpoint variations in the processing pipeline. According to Rudoy & Zelnik-Manor (2012), the information of the viewpoint of a person is very important as the pattern of similar action performed from different viewpoints are vary. They verified this by showing that the better viewpoints are those where the action is easy to recognize and conclude that the selection of viewpoint does improve the action recognition rate. This is the inspiration to attempt the action analysis from different viewpoints to achieve view invariant human action recognition.

Secondly, although there are numerous works proposed in the fuzzy HMA (section 2.2), most of the works are still focusing on the crisp or binary outcome where the defuzzification step is still mandatory in their works. Sometimes it might not be the best solution due to the reason that ambiguity might abounded in the final output. This is because uncertainties exist in most of the real-world applications as mentioned in chapter 1 that causes the confusion in final classification task. Human action recognition and scene

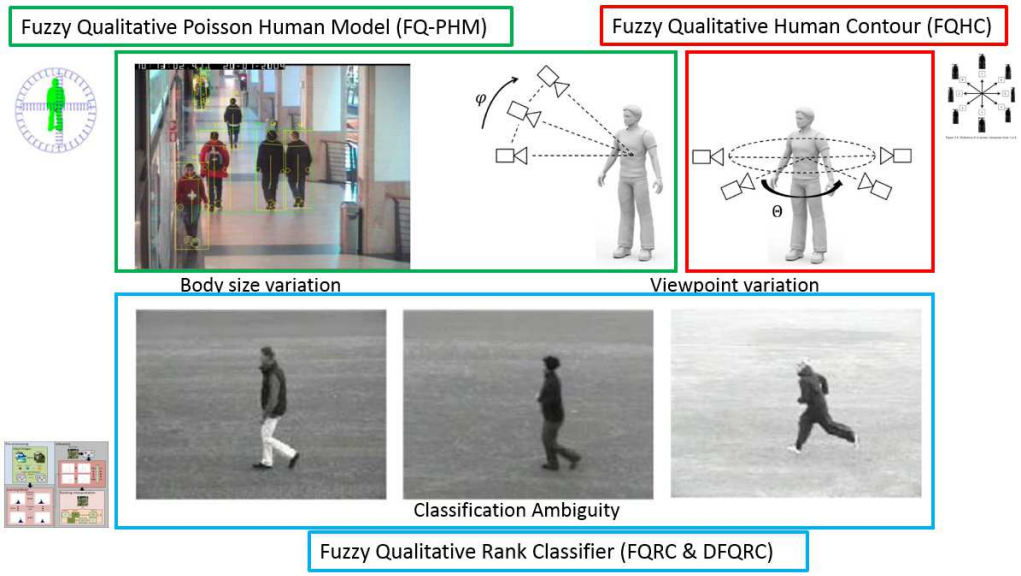


Figure 2.18: Fuzzy qualitative reasoning to address the uncertainties.

understanding are fall into this category. With this, brutally force an output to be one of the possible class might deteriorate the system performance. Such circumstance lead us to propose the use of fuzzy qualitative reasoning instead of the others fuzzy approaches. Figure 2.18 shows the corresponding fuzzy qualitative approach that proposed to deal with each uncertainty respectively.

2.4 Fuzzy Quantity Reasoning

From numbers of fuzzy approaches that had been applied to improve the performance in HMA, fuzzy qualitative reasoning is the state-of-the-art approach recently. A brief revisit of FQS which is built from this theory will be explained here as this thesis utilized FQS in many ways such as to build FQ-PHM for feature extraction, the implementation of FQRC and DFQRC. In general, FQS is introduced by Liu et al. (2009) to replace the conventional Cartesian space into the fuzzy qualitative Cartesian space. This is motivated by the fuzzy qualitative reasoning that proposed by Shen & Leitch (1993).

A FQS is generated by a finite discrimination of the underlying range of each variable of a system being modelled. The FQS will have the desirable properties of finiteness and

coverage, as long as the system contains a finite number of variables. Granularity in the FQS is obtained by the arbitrariness of the discrimination of the numeric ranges of system variables that are assumed to be of interest. Hence, a subset of a numeric range can be translated to one qualitative value according to what is needed in a particular modelling process, such that the extensions of a single qualitative intention may be rather different. The adoption of fuzzy subsets has a direct distinct advantage over the traditional crisp representations when considering granularity.

In fact, if one intends to describe the qualitative values of system variables only in terms of the crisp subsets of the underlying real range of the variables, the mapping from the real range to a quantity space will result in the search for the limits of the real numbers served as the boundaries between (dis-jointly) adjacent qualitative values within the quantity space. This usually incurs severe difficulties in determining these limits (Shen & Leitch, 1993). The fuzzy representation of qualitative values is more general than ordinary (crisp) interval representations, since it can represent not only the information stated by a well-determined real interval but also the knowledge embedded in the soft boundaries of the interval. Thus, FQS removes, or largely weakens (if not completely resolving), the boundary interpretation problem, achieved through the description of a gradual rather than an abrupt change in the degree of membership of which a physical quantity is mapped onto a particular qualitative value. It is, therefore, closer to the common sense intuition of the description of a qualitative value. The interval values are denoted as a fuzzy tuple.

2.4.1 Fuzzy Tuple

This definition on a FQS is given in a general form such that the operations performed within such a quantity space, consisting of normal and convex fuzzy numbers with arbitrary forms of distribution. As a matter of fact, operations on fuzzy qualitative values are

based upon the extension principle outlined in Shen & Leitch (1993). This principle is invoked every time an arithmetic operation is performed and requires expensive calculation. Also, the computational implementation of the calculation with arbitrary membership distributions of fuzzy numbers can only be done in a discrete domain obtained by sampling the original continuous distribution. The use of the extension principle with sampled membership distributions generates a considerable increase in the discrete samples of the result, and furthermore, only some of the resulting samples are correct. Fortunately, more efficient ways to characterise fuzzy numbers have been developed. This utilises a parametric approximation of the membership function where the membership distribution of a normal convex fuzzy number is approximated by the 4-tuple number, $[a\ b\ \alpha\ \beta]$.

An example of fuzzy tuple is shown in Figure 2.19, and defined as,

$$\mu_A(x) = \begin{cases} 0 & x < a - \tau \\ \tau^{-1}(x - a + \tau) & x \in [a - \tau, a] \\ 1 & x \in [a, b] \\ \beta^{-1}(b + \beta - x) & x \in [b, b + \beta] \\ 0 & x > b + \beta \end{cases} \tag{2.1}$$

A FQS formed in this way makes it possible to build a bridge between 'sets' and 'value' because representation allows a real number, a real interval, a fuzzy number, and a fuzzy interval to be uniformly described. Thus, the qualitative category representation and the ordinal representation can be combined in a natural way. For example, the real number 4 can be denoted by a real interval [4 4], which in turn, can be represented by a 4-tuple fuzzy number [4 4 0 0], whilst this fuzzy number is a special fuzzy subset of the real line.

Similarly, the real interval [3.8, 4] can be represented by the fuzzy description [3.8

4 0 0], and the strict fuzzy number 'approximately 4' may be expressed by [4 4 3 3]. In this way, when there does exist a precise qualitatively distinct landmark value, this value can also be represented in the form of a 4-tuple fuzzy number. Furthermore, even if the landmarks are only partially known, say, in terms of the lower and upper (exact) boundaries of the range within which a landmark value falls, such knowledge can still be encoded by the 4-tuple version of a real interval as shown in Figure 2.19.

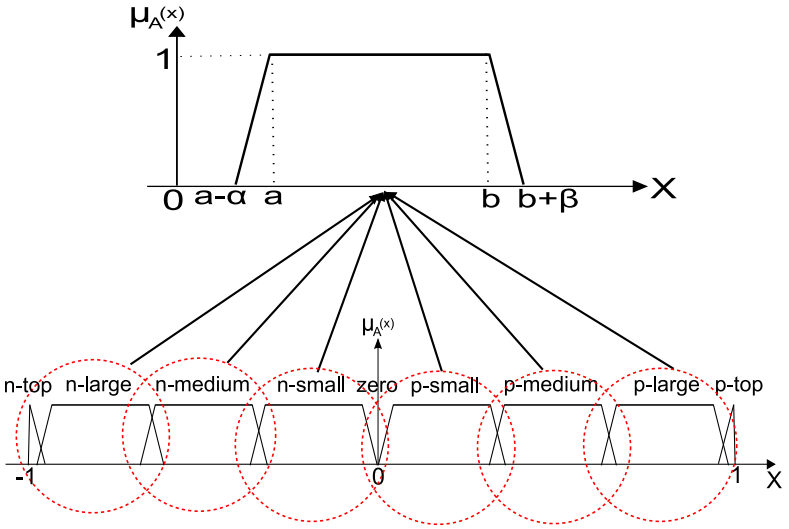


Figure 2.19: 4-tuple fuzzy quantity space.

2.4.2 Construction of Fuzzy Quantity Space

In the recent trends, 4-tuple fuzzy numbers have been utilized in constructing the fuzzy quantity space (Liu et al., 2009) that endowed with the capability to model the uncertainties. As mentioned in (Liu et al., 2009), fuzzy quantity space is replacing the conventional Cartesian space into fuzzy qualitative Cartesian space and has been contributed in many ways towards motion analysis (Chan & Liu, 2009; Chan et al., 2010; Liu et al., 2008b) to alleviate the discrete representation which capable of modelling the uncertainties found in respective works. Here, a brief explanation on the construction of FQS will be described in terms of its architecture and the advantages. To begin with, let's denote FQS with Q

which is composed from an orientation component Q^o and the translation component Q^t,

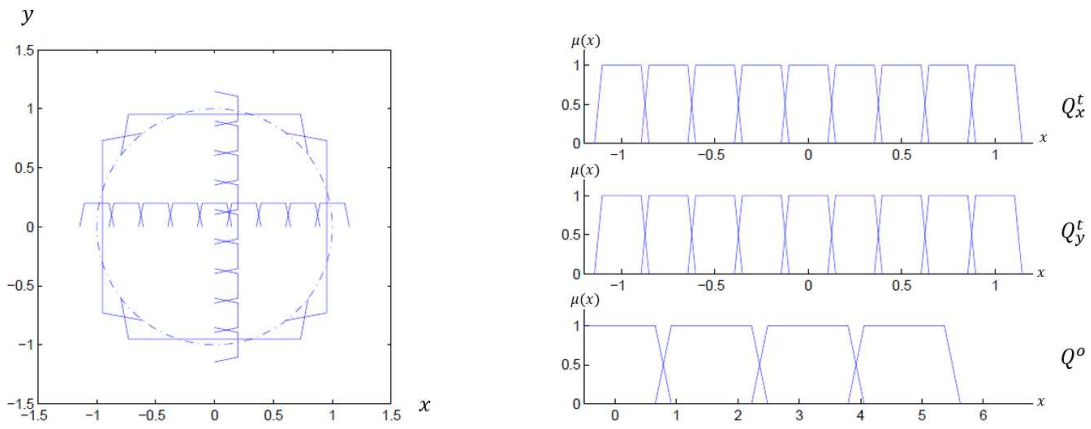
$$Q = \{Q^o, Q^t\} \tag{2.2}$$

where,

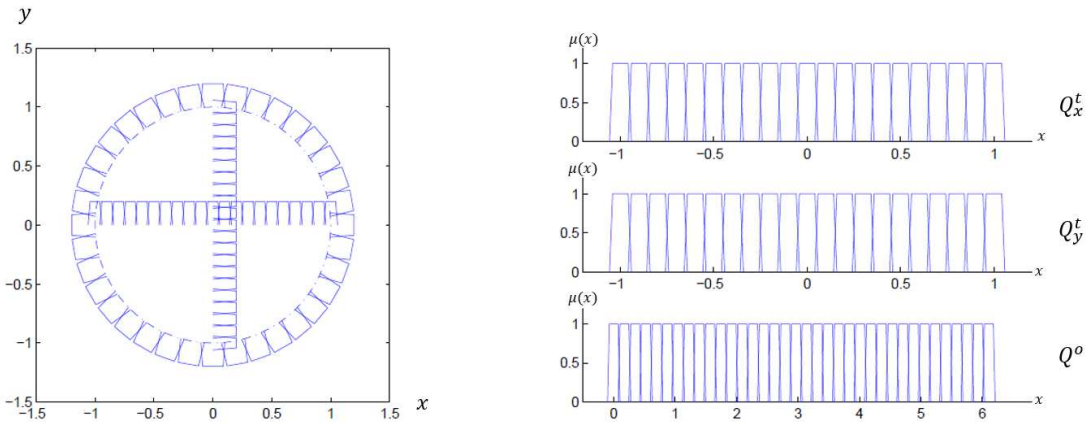
$$\begin{aligned} Q^o &= \{QS_o(\theta_m)\}, \quad \text{where } m = 1,2,3,\cdots,M \\ Q^t &= \{QS_t(l_n)\}, \quad \text{where } n = 1,2,3,\cdots,N \end{aligned} \tag{2.3}$$

$QS_o(\theta_m)$ denotes the state of an angle m, $QS_t(l_n)$ denotes the state of a distance l_n , M and N are the number of the elements of the two components. Figure 2.20 shows some examples of the Q_o and Q_t with different number of the respective components. The position measurement of $P(QS_o(\theta_m), QS_t(l_n))$ is determined by both the characteristics of the fuzzy tuple of $QS_o(\theta_m)$ and $QS_t(l_n)$. For example, an origin is represented as $P_0 = (X_0, Y_0) = ([0\ 0\ 0\ 0], [0\ 0\ 0\ 0])$.

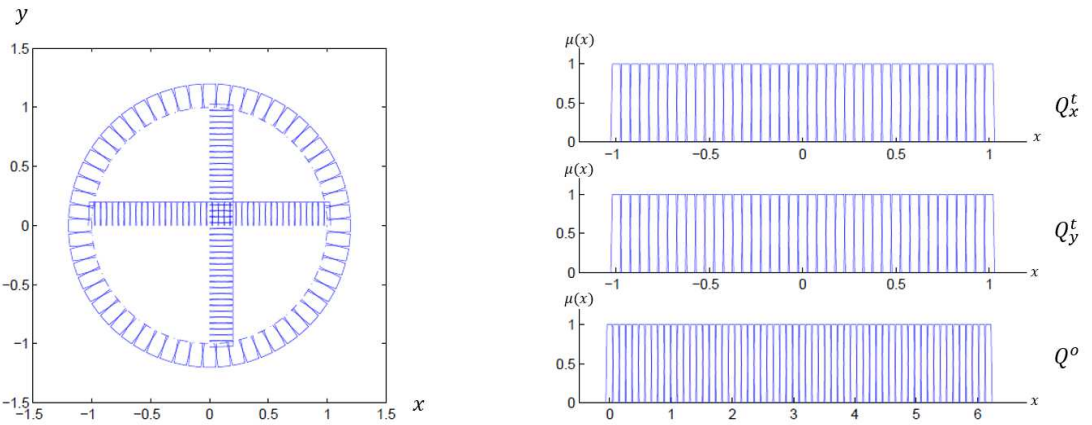
Due to the capability of FQS to model the crisp values in a more general bounded fuzzy interval, the FQS is used to the FQ-PHM in this work. Each measurement of the human model is normalized into the qualitative states which benefits the extraction of the features in a finite well-determined interval manner. Besides that, it is also empower the modelling of ambiguous case, in specific, the modelling of the data distribution in non-mutually exclusive manner for FQRC and DFQRC to obtained a generative model that captured the ambiguity in classification task.



(a) $N = 4, M = 4$



(b) $N = 10, M = 36$



(c) $N = 20, M = 60$

Figure 2.20: Examples of the fuzzy quantity space with different number of components N and M.

CHAPTER 3: VIEW SPECIFIC HUMAN ACTION RECOGNITION

3.1 Introduction

Humans are not restricted to perform an action at a fixed camera viewpoint which means a human subject can be in vary size or viewpoints. This caused the difficulties in HMA as the uncertainties could abounded in the image acquired from the video camera such as different human sizes and viewpoints. As a recall, such uncertainties could hinder the human detection and modeling and motion tracking step in the HMA pipeline (please refer to Figure 1.7).

In general, human detection and modelling is the first step in HMA pipeline with the objective to project the human segment discovered from the image frame into a more meaningful representation. The intention is to obtain a generalized human model which in ideal case it will normalized over the uncertain situations and provides the feasibility in feature extraction task. This is a vital process to minimize the errors in the next processing steps which is the motion tracking. Based on the literature review, conventionally, skeleton, bounding box, blob, cylinder, or cone are used to represent the human body (Aggarwal & Cai, 1997; Aggarwal & Ryoo, 2011). However, due to the uncertainties that are abounded in this modelling step, particularly the human size and the viewpoint, a more sophisticated human model that is capable of model these uncertainties is required.

In motion tracking, view invariant is the current trend as human subject that acquire from the video camera can be from different angles, and the algorithm that is capable of obtaining the motion information from different angles is required and more practical. However, most state-of-the-art view-invariant research (Ahmad & Lee, 2006; Holte et al., 2011; Ji & Liu, 2010; Weinland et al., 2007, 2006; Yilma & Shah, 2005) are found deviated from the practical solution where the acquiring of data from various viewpoint angles

is still mandatory in the real time processing. These works assumed that the subjects always perform actions in a position of the frontal-parallel to the camera; and require to build a 3D action model for action recognition purpose. This assumption has a few limitations. First of all, in a real-world environment, subjects are not always frontal-parallel with respect to each of the cameras. Secondly, finding a multiple cameras system in the public space that covers many overlapping regions is uncommon. Therefore, the 3D action model built based on the assumption above may not be very practical in a real-world environment.

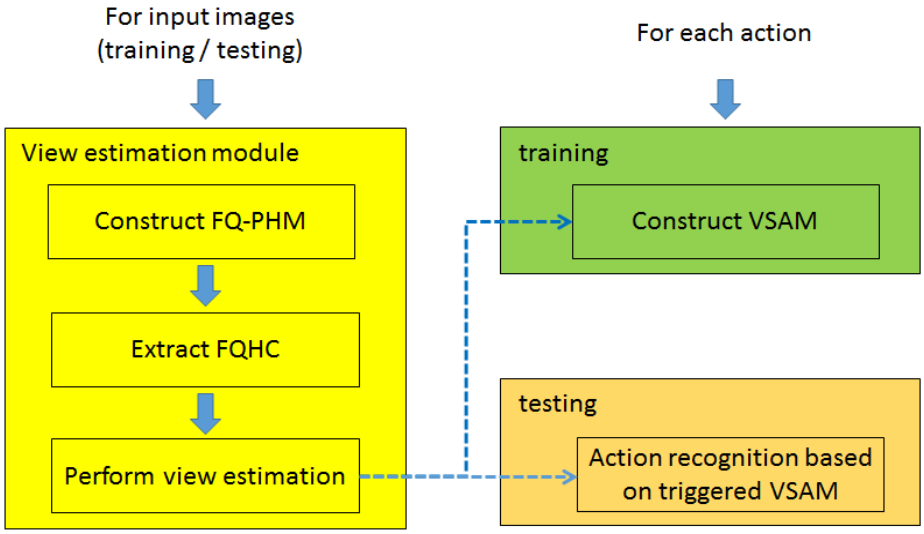


Figure 3.1: The overall flow to construct view specific action recognition framework.

Given the problems above, hence, there is a need to implement a HMA system for multiple views within a single camera to achieve view invariant action recognition. As a solution, a view specific action recognition framework is proposed where the action models learned from different viewpoints are used for the recognition task within a single camera system. What makes this work different from the previous one is, the action model is first build from the different viewpoints with the extracted features instead of correlating the information from multiple cameras in the processing. To achieve this, view estimation module and VSAM are two important components. In this framework, FQ-PHM is constructed as a solution for the modelling problem. In addition, FQHC

can be extracted from FQ-PHM to learn and helps to construct the VSAM for the use of performing action recognition in view independent manner. The overview of the flow to construct view specific action recognition framework is presented in Figure 3.1.

3.2 Proposed Viewpoint Estimation Module

Viewpoint estimation module is an important component in the proposed view specific action recognition framework to identify the viewpoint of the subject in front of the camera. This information is then used to construct the corresponding VSAM for action recognition task. This approach eliminates the require of multiple camera installed in the specific area. To achieve this, a human contour descriptor namely FQHC is extracted from the FQ-PHM which is generalized over variation of size, body anatomy, and camera positions to learn the person viewpoints at the initial stage.

3.2.1 Fuzzy Qualitative Poisson-normalized Human Model

It is a difficult task to locate and project the person from the image frame to another desire space without knowing the intrinsic parameters for camera calibration. This space could be in any form that build with the intention to represent the human segment in a more meaningful manner for a specific task, such as feature extraction. However, due to the variations in human size and the camera positions, for instance the φ angle, this may hinder the process of generating a sophisticated human model and extraction of the robust features. Besides that, without a proper human model, the system is commonly consumed high computational cost in the processing as it requires extensive training to sustain for all the aforementioned variations (Dalal & Triggs, 2005; Lewandowski, Makris, & Nebel, 2010).

As a solution, human modelling is performed after the human segment is identified in an image frame. There are many ways to do it as mentioned in the literature

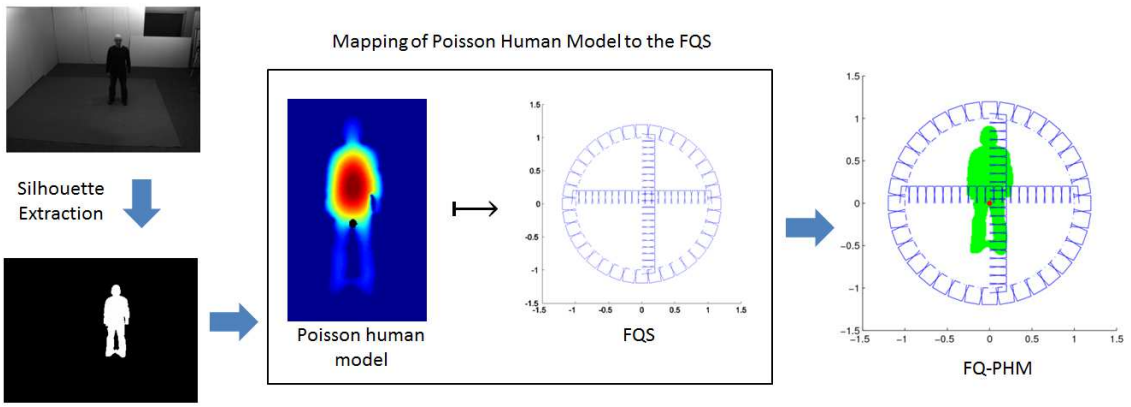


Figure 3.2: The overall pipeline to generate FQ-PHM.

(skeleton, bounding box, blob, cylinder, or cone), but most of these methods constrained by some limitations such as fixed human sizes or fixed camera positions. For examples, Lewandowski, Makris, & Nebel (2010) is limited to a fixed φ angle, and similarly to works using Histograms of Oriented Gradients (HOG) for human detection (Dalal & Triggs, 2005). To the extend, extensive training is required to obtain a sophisticated trained model to cope with these variations where collecting the training images for each variation is a very tedious and time consuming job. Instead, FQ-PHM is proposed to cope with these uncertainties at initial stage before the feature extraction step is conducted. The final goal is to achieve a generalized human model that will eliminates as much as possible the variations between the human subjects in terms of size , body anatomy, and camera positions. To achieve this, the FQ-PHM is proposed to normalize the human segment into the FQS with helps of Poisson solution (Gorelick et al., 2006a). The overall pipeline to generate FQ-PHM is showed in Figure 3.2. This is the prerequisite for extracting the FQHC to perform viewpoint estimation.

3.2.1 (a) Poisson annotated human model

The conventional way of performing normalization on a human segment is commonly based, solely on the longest measurement of the body and represent it with bounding box,

blob, cylinder, etc. (Aggarwal & Cai, 1997; Aggarwal & Ryoo, 2011). However, these works did not consider the uncertainties caused by the human size variation such as the body anatomy. This is because, humans are different from one and another not only on their overall body size, but also their body parts. For example, some people may have longer upper body and longer legs. Overlook on this variation might results in inappropriate body modelling that might causes bad performance in latter feature extraction step, in particularly for those features that related to the body parts. In order to obtain a more appropriate human model, Poisson solution (Gorelick et al., 2006a) is applied to locate a reference points, $\mathbf{r}$ on the human body which is used as an indicator to precisely normalize the other body parts into the FQS. The normalized human model with $\mathbf{r}$ is generalized in terms of the position of body landmarks. For example, the upper body and the lower body of a person is always located in fixed qualitative states once the human size is normalized into the FQS. Ideally, they will never over go each other in the FQS and the desire feature can be extracted from this human model correctly.

To begin with, given a human silhouette, S, $\mathbf{r}(x,y)$ is obtained by

$$\mathbf{r}(x,y) = \max(U(x,y)) + C \tag{3.1}$$

with $U(x,y) \in S$ is computed by solving a Poisson equation of the form $\Delta U(x,y) = -1$, where the Laplacian of U is defined as $\Delta U = U_{xx} + U_{yy}$ subject to Dirichlet boundary conditions $U(x,y) = 0$ at the bounding contour ∂S. C is the constant or function that determine the part of the human body, for instance, human lower torso part is chosen in here which is empirically defined with

$$C = y + (3 * (L/5)) \tag{3.2}$$

where L is the vertical length of the human body in the image. Such definition of C

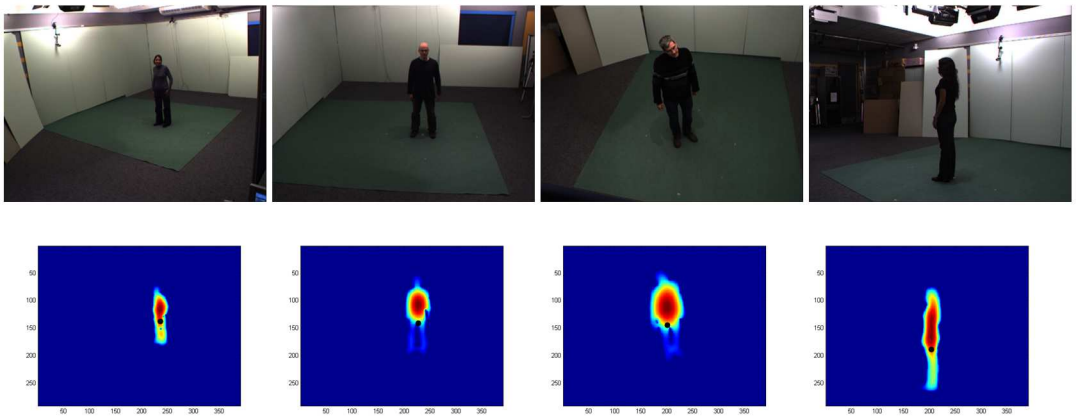
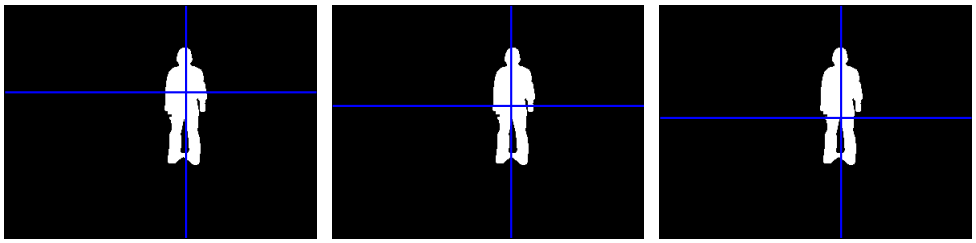


Figure 3.3: The lower part of torso estimated using (3.1) and denoted as a black dot. It is proven that the proposed method works on human with variation of sizes, body anatomy, and postures where the black dot precisely located at their lower torso part (this image is best view with colour).

is adopted because from human inspection, the lower torso part is commonly located at the $3/5$ of the overall body length. The effectiveness of this computation to locate the reference point is shown with some examples in Figure 3.3.

The principle that empowers the notation of this $\mathbf{r}$ point originates from the Poisson solution, U that was proposed by Gorelick et al. (2007, 2006a,b). The value of U increases quadratically as it approach to the centre which is the nature of Poisson equation (Gorelick et al., 2006a). The level sets of U represent smoother versions of the bounding contour with the external protrusions, (where in human context, it refer to the limbs and head) disappearing at relatively low values of U. This is different from the distance transform, which smoothens the shape near concavities while introducing discontinuities near convex sections of the contour. Also unlike the distance transform in which every value is determined by a single contour point (the nearest), the values assigned by the Poisson equation take into account many points on the boundaries and so they reflect more global properties of the silhouette. In human representation, this is giving prudent information as ideally the highest value from the Poisson solution is at the middle of the torso part (Figure 3.4(a)). This is because in general, the torso is the largest part of human body.

However, the highest value of Poisson solution is not directly used in the context.



(a) $\max(U(x,y))$ (b) Conventional mid-point (c) $r(x,y)$

Figure 3.4: Comparison of the methods used to perform segmentation of body parts. 3.4(a) uses maximum value of Poisson solution, $\max(U(x,y))$ while 3.4(b) uses conventional mid-point computation for body parts segmentation which are found to be inappropriate as portion of the hand is cross over the lower body segment. 3.4(c) precisely segment the upper body and lower body as well as the left and right of the body portion.

The reason can be observed from Figure 3.4(a) where the use of $\max(U(x,y))$ results in inappropriate body segmentation. This is because the hand is partly crossover the lower body segment. Thus, lower torso part is chosen in this context because it is a feasible landmark to appropriately separate the body into top and bottom parts, and as well left and right parts. This eases the process to extract the features that rely on the body parts such as the limbs.

Based on Figure 3.3, the reference point is able to correctly separate the body parts and it is invariant to different size, height, and the posture of the human body. Extra merit for this Poisson normalized human model is it endowed the capability to normalize the human body that is capable of precisely locates the specific body parts. Despite of this, it is insufficient for an effective human modelling as the human representation is still affected with the variation in size. For the normalization and modelling part, the Poisson annotated human silhouette S' is then mapped into the FQS.

3.2.1 (b) Fuzzy quantity space mapping

Once the reference points **r** of a human in an image frame is identified, the next step is to represent this human segment into the FQS which is generalized over different body sizes and camera positions. In this thesis, FQS is adopted to represent the human segment and

called the FQ-PHM. S' is map into the FQS with the reference point $\mathbf{r}(x,y)$ as the origin point, P_o in the FQS. To begin with, the conventional Cartesian space and unit circle are being replaced by fuzzy qualitative quantity spaces with a limit of N and M components.

$$\begin{aligned} \lim_{n\to N} C_t(n) &= QS(qp_l) \\ \lim_{m\to M} C_o(m) &= QS(qp_\theta) \end{aligned} \tag{3.3}$$

where n is the number of qualitative state that resides in the x and y translation, while m is the number of qualitative state that resides on the orientation in the FQS (i.e., N and M respectively represent the number of translation and orientation states employed in the quantity spaces to represent the FQS). As $n \to N$ and $m \to M$, the limits of $C_t(n)$ and $C_o(m)$ will approach to a set of N qualitative state for a translation component and a set of M qualitative state for an orientation component. Note that the range of N and M are application dependent and user defined. Empirically in the construction of FQ-PHM, $N = 10$ and $M = 36$ are selected to build the FQS as shown in Figure 3.5. The partition of the qualitative states, qp in the translation and orientation components are constructed as

$$\begin{cases} qp_l^n | qp_l^n \in [0, l_{n_1}, l_{n_2}, \cdots, l_{n_{(N-1)}, l_N}] \\ qp_\theta^m | qp_\theta^m \in [0, \theta_{m_1}, \theta_{m_2}, \cdots, \theta_{m(M-1)}, 2\pi] \end{cases} \tag{3.4}$$

where

$$\begin{aligned} qp_l^n &= \tfrac{l_N}{N}, \quad qp_\theta^m = \tfrac{2\pi}{M}, \\ 0 &\le qp_l^1 \le qp_l^2 \le \cdots \le qp_l^{N-1} \le l_N, \\ 0 &\le qp_\theta^1 \le qp_\theta^2 \le \cdots \le qp_\theta^{M-1} \le 2\pi. \end{aligned} \tag{3.5}$$

Furthermore, Poisson annotated human silhouette is normalized within the boundary with

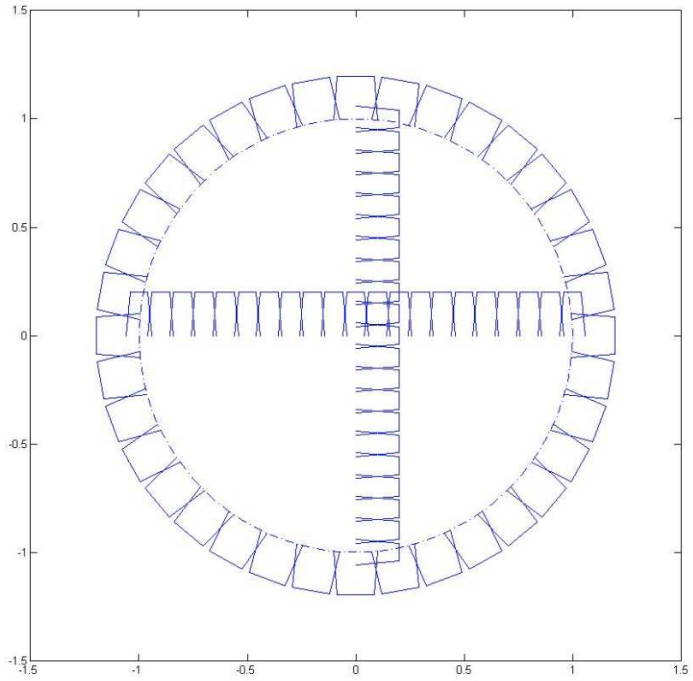


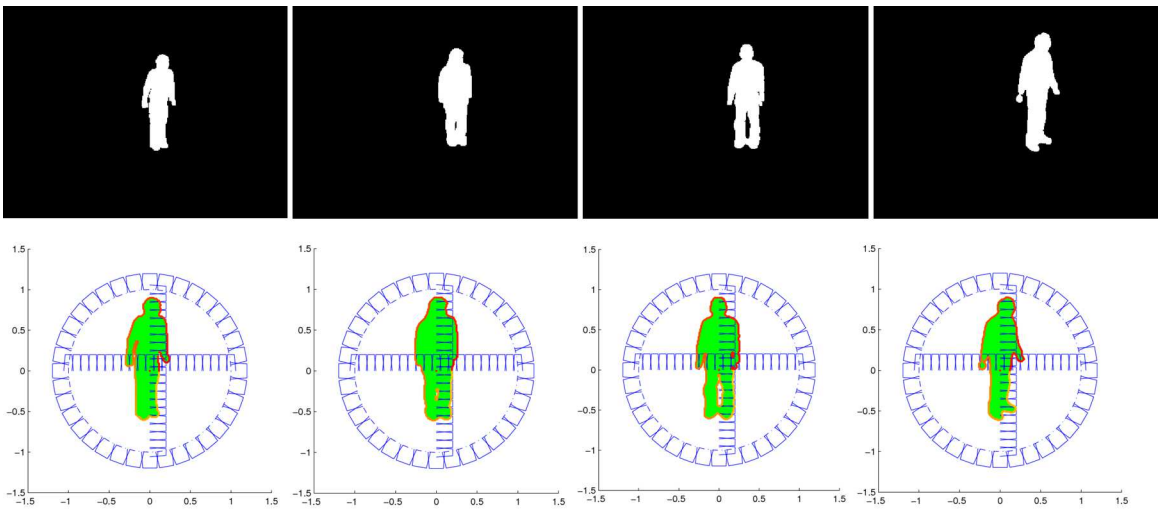
Figure 3.5: Fuzzy quantity space with resolution $N = 10$ and $M = 36$.

[-1 1] with respect to each qualitative state,

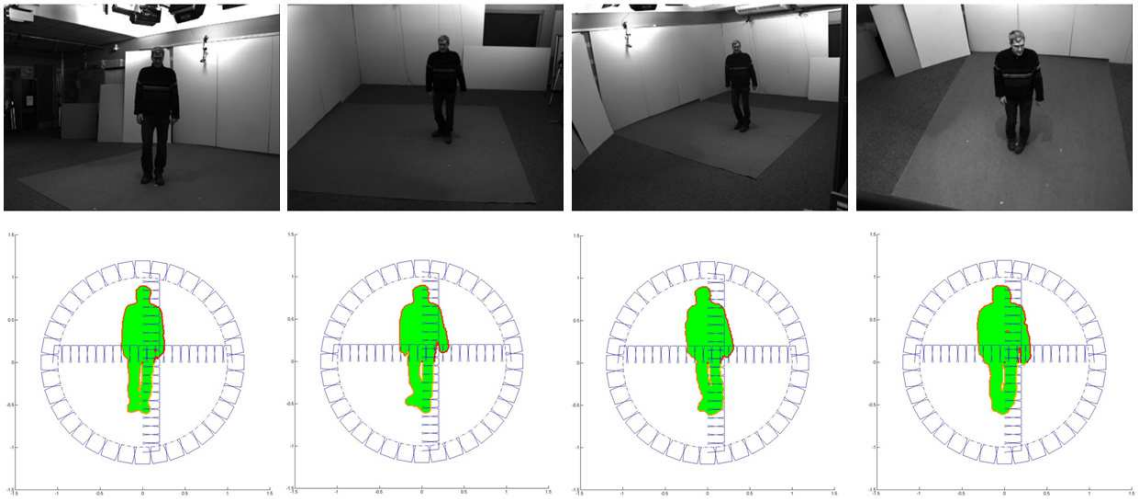
$$\begin{cases} QS(qp_l) = qp_l | qp_l \in \left[\frac{ql_{n_1}}{ql}, \frac{ql_{n_2}}{ql}, \cdots, \frac{ql_{n_{N-1}}}{ql}, 1 \right] \\ QS(qp_\theta) = qp_\theta | qp_\theta \in \left[\frac{q\theta_{m_1}}{2\pi}, \frac{q\theta_{m_2}}{2\pi}, \cdots, \frac{q\theta_{m_{M-1}}}{2\pi}, 1 \right] \end{cases} \tag{3.6}$$

where $x-y$ translation states qp_l are normalized by the body length ql and the orientation states are normalized into 2π.

The outcome of this step is the fuzzy qualitative human model with each component of the body is bounded within the Q^o and Q^t that generalized over the body size, body anatomy and camera positions (φ angle). Some examples of the output are visualized in Figure 3.6. The usefulness of doing this are; (1) It is possible to precisely locates the body parts with fuzzy qualitative states and invariant to the aforementioned uncertainties (body size, body anatomy and camera φ angle) , (2) It can be use to extract the feature in fuzzy qualitative manner as to proposed FQHC.



(a) The corresponding FQ-PHM for human silhouette of different size



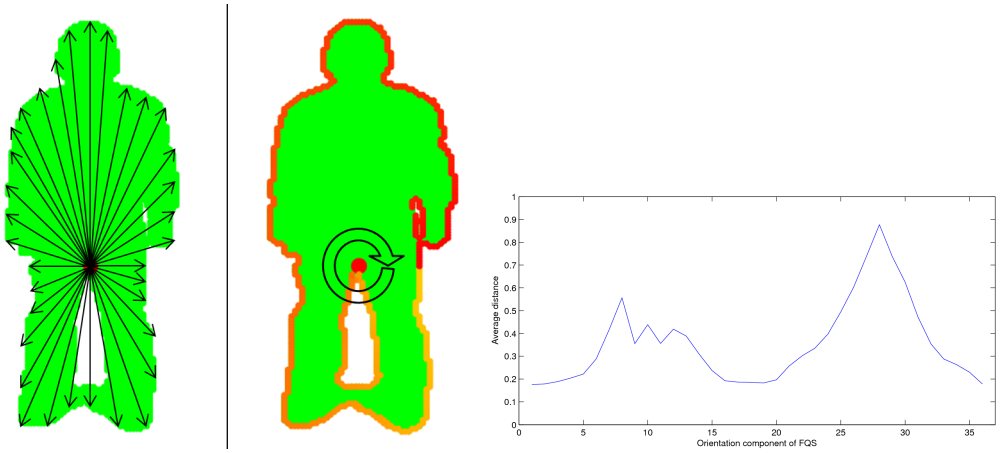
(b) The corresponding FQ-PHM for human image of different φ angle

Figure 3.6: One can notice that the size and the position of the body parts are almost similar for all the human subjects once they are being normalized onto the FQS with **r** as the origin. Thus, it is a human model that generalized over the human size and φ angle.

3.2.2 Fuzzy Qualitative Human Contour

Many years of study on human vision in various domains including cognitive science, neuropsychology, and neurophysiology showed a consensus among researchers that a human recognizes an object based on its appearance such as contour, texture, and colour information (Mel, 1997). Numbers of researches in computer vision are inspired by this finding with the notable work in human detection where the HOG descriptor is introduced by Dalal & Triggs (2005). This is because HOG is capable of representing human appearance and shape of an image which is described by the distribution of intensity gradients

or edge directions. Specifically, more weightage is on the human parts that show sufficient displacement such as shoulder to distinguish human from other objects. However, without generalization over the body size and body anatomy, HOG is found to be not so effective in viewpoint learning as shown in the experiment results (section 3.2.3). As a solution, FQHC is extracted from the proposed FQ-PHM. As an overview, FQHC is extracted by utilizing the distance measure from **r** towards human edges as demonstrated in Figure 3.7.



(a) (b)

Figure 3.7: (a)In the left image, the distance from the ref-point to the outer edge is computed. The distance is organized according to clockwise direction as shown in the right image. (b)The example of the human contour descriptor by averaging the distance in each orientation states of the FQS, $QS_o(\theta_m)$.

First, a set of edge pixels, e_j of the human body is obtained for every qualitative states in Q^o and denoted as $E_{QS_o(\theta_m)}$

$$E_{QS_o(\theta_m)} = \{e_1, e_2, \cdots, e_J\} \in QS_o(\theta_m) \tag{3.7}$$

Note that, the edge pixels can be easily obtained by extracting the outer pixels of the silhouette image. In this work, the FQHC is constructed by computing the average of the distances from the **r** towards the sets of edge pixels that are bounded in the corresponding

fuzzy qualitative orientation state, QS_o,

$$\bar{E}_{QS_o(\theta_m)} = \frac{1}{\left|E_{QS_o(\theta_m)}\right|} \sum_{j=1}^{J} \left\| e_j - r \right\|^2 \tag{3.8}$$

then, a descriptor for the human contour designated as d is constructed by concatenate the $\bar{E}_{QS_o(\theta_m)}$ for all qualitative states,

$$d = \{\bar{E}_{QS_o(\theta_1)}, \bar{E}_{QS_o(\theta_2)}, \cdots, \bar{E}_{QS_o(\theta_M)}\} \tag{3.9}$$

following the clockwise direction as shown in the right image of Figure 3.7(a). The dimension of d is determined by the number of the orientation component, M in the Q^o.

Algorithm 1 FUZZY QUALITATIVE HUMAN CONTOUR EXTRACTION

Require: An input image
  Step 1: Silhouette extraction. Perform silhouette extraction to obtain binary representation of human body, S.
  Step 2: Apply Poisson solution. Apply (3.1) towards the human silhouette to obtain reference point, $\mathbf{r}$.
  Step 3: Normalize onto the FQS. Normalize the Poison annotated human silhouette S' into the FQS with the range of $[-1\ 1]$ with $\mathbf{r}$ as the origin (3.6).
  Step 4: Extract human contour descriptor.
  for all $QS_o(\theta_m)$ such that $1 \leq m \leq M$ **do**
    for all e_j such that $1 \leq j \leq J$ **do**
      Compute average of the distances, $\bar{E}_{QS_o(\theta_m)}$ from $\mathbf{r}$ to e_j
      as (3.8)
    end for
  end for
  return FQHC descriptor, d

There are reasons to use distance averaging within bounded qualitative state. First is to create a vector descriptor with fix dimension. By comparing it with directly use all the edge pixels from the human silhouette, the feature vector can be in various dimensionalities due to the different number of edge pixels for different human subject and the feature extraction process can be infeasible. Secondly, the averaging of the edge pixels within the qualitative state can smoothen the inconsistency of some edge pixels due to the noises. This is because the borderline edge pixels can belongs to two qualitative states at

the same time, and thus the crossover edge pixels in both qualitative states can alleviate the abrupt change in the human contour extraction. The extracted FQHC can then use to perform view estimation in the view specific action recognition framework and helps in the construction of VSAM. For ease understanding, the steps for the extraction of FQHC are summarized in Algorithm 1.

3.2.3 Robustness of Fuzzy Qualitative Human Contour

In order to validate the robustness of the proposed FQHC, the clustering algorithm is applied to evaluate the performance of FQHC in distinguishing between several human viewpoints. At the same time, the performance is compared with the HOG.

3.2.3 (a) Predefined viewpoints

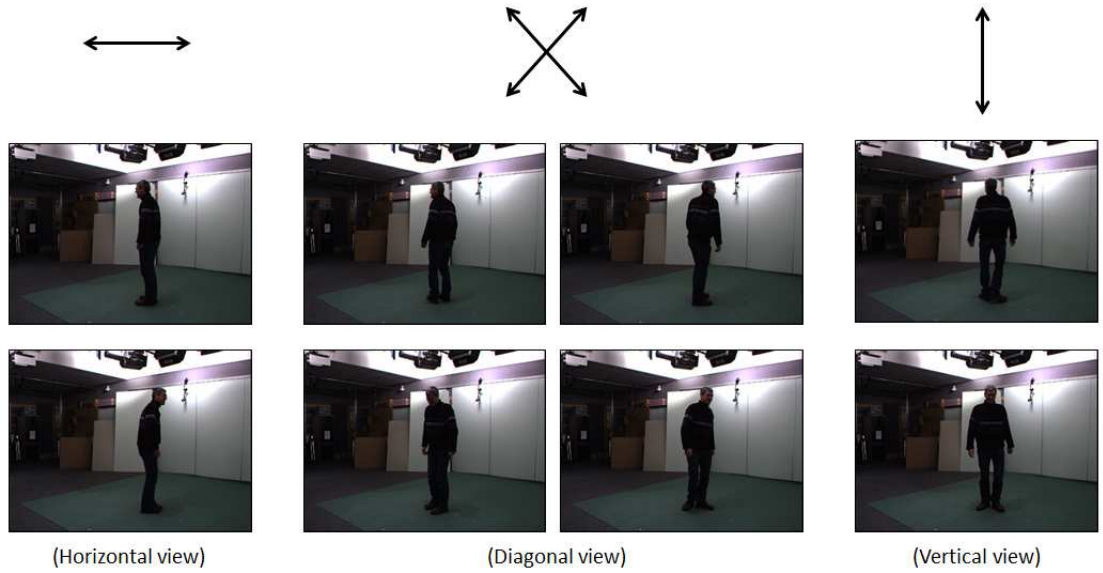


Figure 3.8: Definition of viewpoints, from left to right, 'horizontal view, v_1', 'diagonal view, v_2' and 'vertical view, v_3'.

In this thesis, the scope is narrowed to three predefined dominant viewpoints which are the most common viewpoints that will be encountered from a video camera. They are vertical, v_1, diagonal, v_2, and horizontal, v_3, views as depicted in Figure 3.8. These dominant viewpoints are composed of a set of different atomic viewpoints and designated

as $v_1 = \{1,5\}$; $v_2 = \{2,4,6,8\}$; $v_3 = \{3,7\}$ with refer to Figure 3.9. According to Rogez et al. (2014), these viewpoints are sufficient for to analyse human actions from different angles.

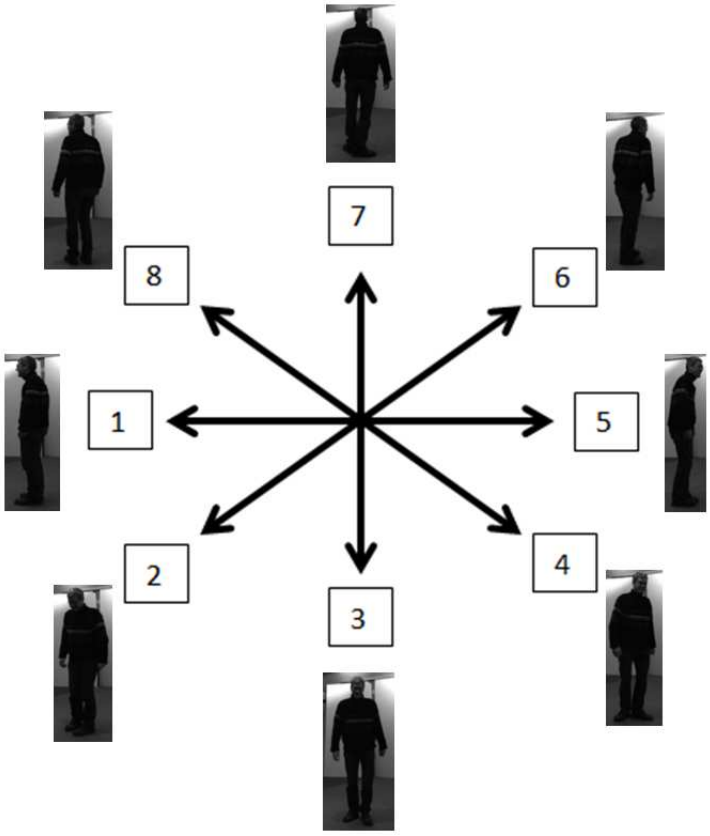


Figure 3.9: Definition of of atomic viewpoints from 1 to 8.

3.2.3 (b) Dataset

In the experiments, the INRIA Xmas Motion Acquisition Sequences (IXMAS) multiview dataset (Weinland et al., 2007) for view-invariant human action recognition is used. The dataset consists of 13 daily-live motions performed, each three times, by 11 actors and captured by five cameras. In this testing, first three frames of every motion video in IXMAS dataset are retrieved for viewpoint clustering purpose. This is because in the video sequence, the subject are still remain at the initial standing position at the beginning of each video frames except for the "get up" action as the subject is initially sitting on the ground. In that case, the last three frames were used.

3.2.3 (c) *Ground truth*

IXMAS dataset is taken from multiple cameras environment where the cameras are calibrated at the positions of approximately 45° gap in Θ angle between each camera. This means that each of the camera viewpoint is capable of providing the information of respective viewpoints in v_1, v_2, and v_3. For example, if the subject captured by camera 1 is in v_1, camera 2 will be v_2 of the subject, camera 3 will be v_3 of the subject and finally camera 4 will be categorized back to v_1 again. With this, the ground truth are be obtained easily by human inspection and will be used for evaluation purpose.

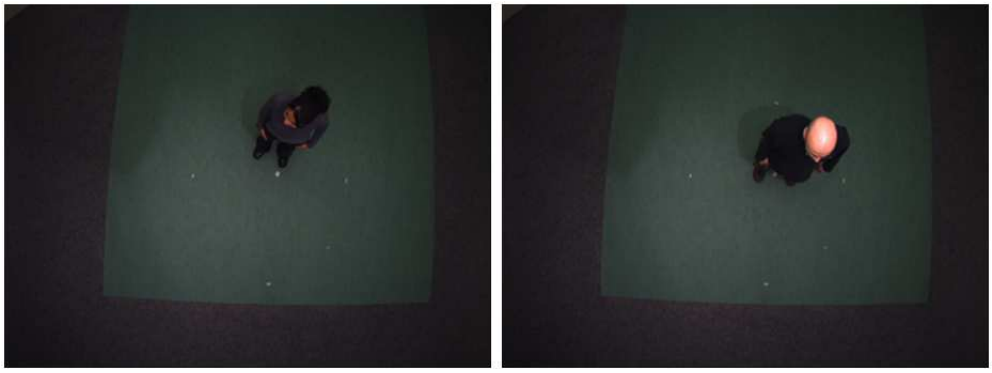


Figure 3.10: Examples of top view, v_4.

Note that special case which is the top view camera (camera 5) in the IXMAS dataset is included in this testing only for robustness test and denoted as v_4 (Figure 3.10). However, v_4 is not included in the consideration for viewpoint estimation in the overall proposed framework as this viewpoint is beyond the scope of this thesis. The proposed framework will fail to recognize an action as the human model obtain from v_4 as it is deviated too much from a normal human pose. This is because the φ angle is too high and only limited part of the body is visible from the image frame as shown in Figure 3.10. The ground truth for v_4 is directly adopted from the original labeling (Weinland et al., 2006).

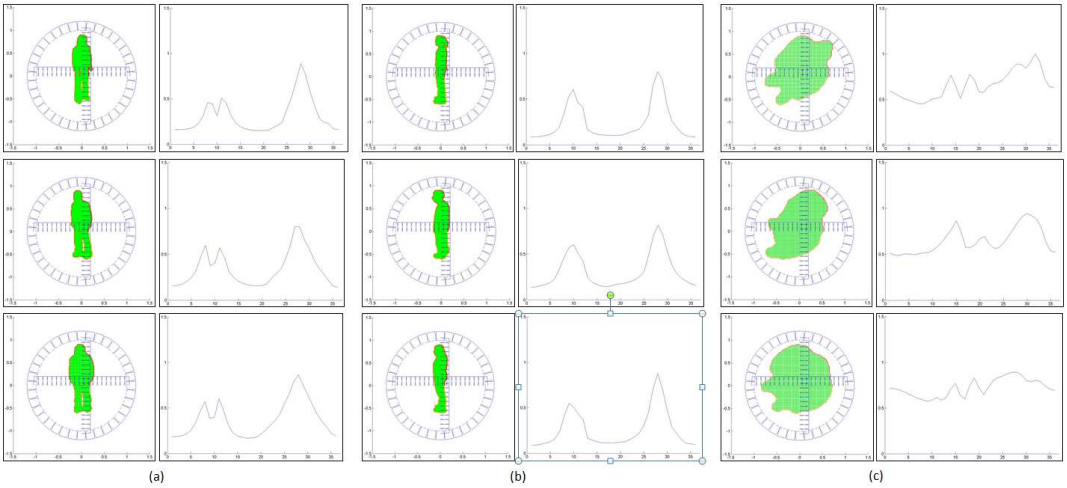


Figure 3.11: Examples of fuzzy qualitative human contour descriptors for different viewpoints.

3.2.3 (d) Clustering of FQHC

In order to validate the robustness of the extracted FQHC in distinguishing different human viewpoints, clustering algorithm is employed to learn the FQHC and automatically group them into the respective viewpoint cluster. Based on the observation in Figure 3.11, one can notice that the human contour descriptors are different between the viewpoints. From this observation, it is presumable that FQHC possess the capability to differentiate different human viewpoints, and thus clustering algorithm is used to test their discriminative strength.

In this validation, each cluster means the different viewpoints. In a simplified manner, the cluster is denoted as the set of viewpoints that defined previously which are v_1 to v_4. Therefore, the number of cluster here is $K = 4$ corresponding to each type of viewpoints $V = \{v_1, v_2, \cdots, v_K\}$. With the collection of the FQHC descriptors extracted from all the training samples $d = \{d_1, d_2, \cdots, d_T\}$, they are input into the clustering algorithms and the outcomes are expected to be similar to Figure 3.12 where each of d is correctly assigned to their respective viewpoint cluster.

The main objective of this testing is to evaluate the capability of the FQHC descriptors which are extracted from the FQ-PHM in distinguishing one viewpoint from another.

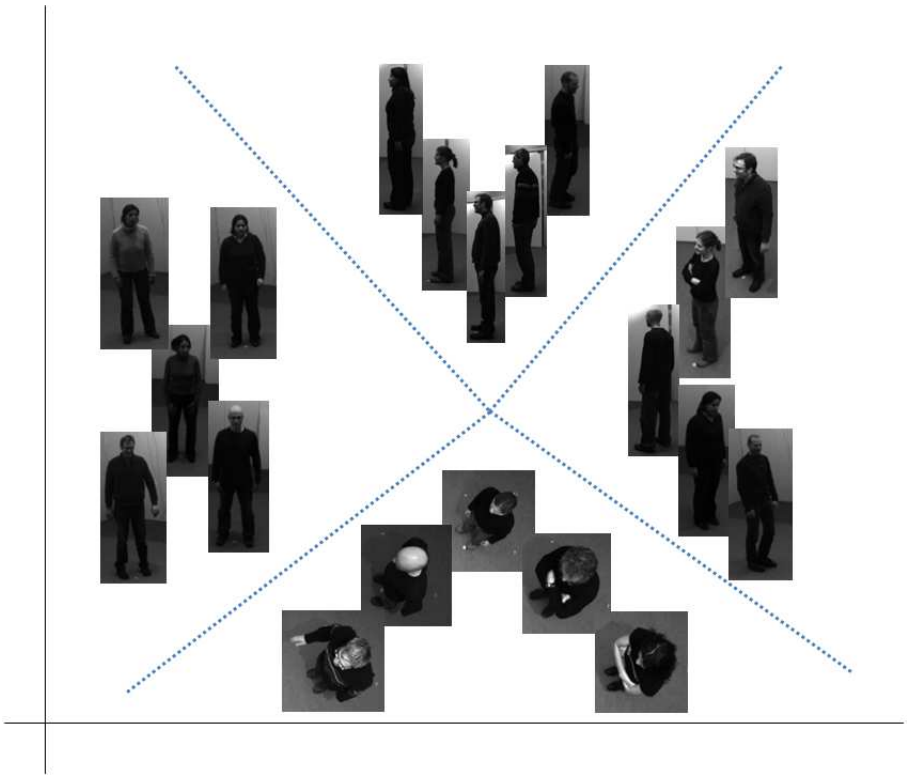


Figure 3.12: Expecting outcome for the viewpoint clustering.

This is an important criteria as it will be the prerequisite for the viewpoint estimation module in the proposed view specific action recognition framework. The performance of FQHC is also compared with the HOG descriptor (Dalal & Triggs, 2005) and the results are being evaluated in terms of precision and recall using K-means (KM) and Fuzzy c-means (FCM) clustering algorithms (Xu et al., 2005). The average testing results of the 20 trials are reported in Table 3.1.

Table 3.1: Precision (Ps) and Recall (Rc) for the clustering results.

	v_1		v_2		v_3		v_4	
	Ps	Rc	Ps	Rc	Ps	Rc	Ps	Rc
FQHC_KM	**0.92**	0.74	0.57	0.57	0.58	0.72	**1.00**	0.99
FQHC_FCM	0.83	**0.87**	**0.65**	**0.58**	**0.66**	**0.72**	**1.00**	0.99
HOG_KM	0.67	0.82	0.54	0.36	0.64	0.58	0.91	**1.00**
HOG_FCM	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail

According to Table 3.1, the precision and recall for v_1 and v_4 achieved high precision and recall but v_2 and v_3 showed fair results due to the confusion between the diagonal and vertical views as demonstrated in Fig. 3.13(a). The human contour descriptors for v_2 and

v_3 are similar to each other and this is acceptable as in the real environment, human has the difficulty to distinguish them too.

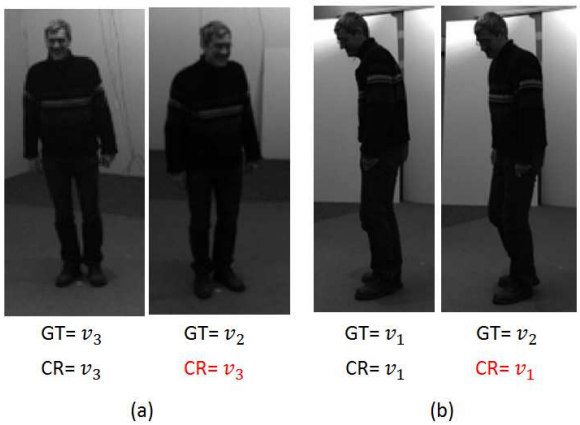


Figure 3.13: The examples of the viewpoint confusion with its ground truth denoted as GT and the computational result denoted as CR. One can notice that the right image in (a), the CR is conflict with GT where the computer incorrectly group it as v_3. While in (b), the right image is incorrectly grouped as v_1. In despite, this is acceptable due to the ambiguity abounded in the processing.

Apart from this, due to the randomness of initialization in clustering algorithm, it may lead to undesirable clustering outcome. For example, the "fail" cases in Table 3.1 occur because these descriptors are unable to form the expected clusters to represent each viewpoint in 20 trials. However, this can be a good evaluation criterion to determine the discriminative strength of a descriptor. In common practise, the lower the error rate of the clustering performance implies the better the discriminative strength of the descriptors. Table 3.2 shows the error rates of the testing.

Table 3.2: Error rate of the clustering.

	Error rate
FQ-PHM_KM	**0.2**
FQ-PHM_FCM	0.3
HOG_KM	0.8
HOG_FCM	1.0

Human contour descriptor extracted from the FQ-PHM performed well with KM and FCM with low error rate but HOG receive high error rate in KM and even failed to perform clustering with FCM as shown in Table 3.1 and 3.2. From these results, one can conclude

that, the proposed FQHC has better reliability to represent a viewpoint compared to the HOG.

3.3 View Specific Action Model (VSAM)

Another important component in the proposed view specific action recognition framework is the VSAM. It can be learned through desire machine learning technique with appropriate motion features but subject to the specific viewpoints as defined in section 3.2.3 (a). It is obviously that the prior information of the subject viewpoint is the priority during the training of the VSAM, manually human annotation can be done by visual inspection to obtain these prior information but it is a tedious job and impractical. An alternative approach is to apply the viewpoint estimation module which utilized the proposed FQHC. The comparison between the performance of using the VSAM generated from human annotation and viewpoint estimation algorithm is discussed in the experiment.

3.4 Experiments and Discussions

In this section, the experiments are conducted to evaluate the performance of the proposed viewpoint estimation algorithm and also the feasibility of action recognition using the view specific action recognition framework. IXMAS dataset (Weinland et al., 2006) is again used in these experiments.

3.4.1 Viewpoint Estimation

The previous validation in section 3.2.3 (d) had showed the robustness of FQHC in representing the viewpoints but in a fixed camera position. As an extension, this experiment is conducted to observe the effectiveness of using FQHC to perform viewpoint estimation on different camera positions. In more specific, the camera positions are denoted as Cam

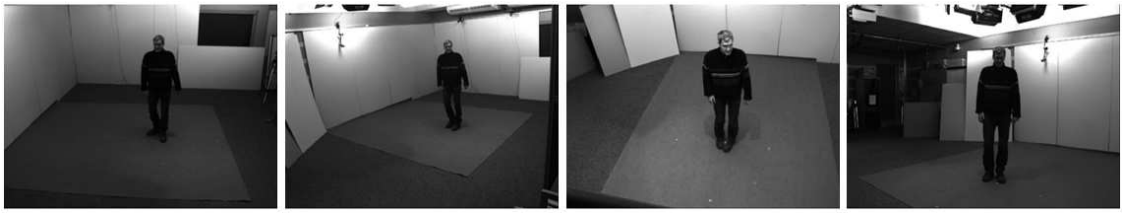


Figure 3.14: Image from left to right representing Cam 1, Cam 2, Cam 3, and Cam 4 respectively with all these camera are set up at different φ angle.

1 to Cam 4 as depicted in Figure 3.14. These cameras (i.e, Cam 1 to Cam 4) respectively represents different φ angle but each of them covered v_1 to v_3 in capturing the human action. In viewpoint estimation, the first three frames of the video sequence are chosen for the purpose as the subjects are all in the initial standing position. Note that, viewpoint estimation is just normal classification tasks based on the learned viewpoint clusters in this context.

Table 3.3: Accuracy of view estimation (for check watch action).

Feature	Cam 1	Cam 2	Cam 3	Cam 4	Average
HOG	0.67	0.67	**0.61**	0.53	0.62
FQHC	**0.75**	**0.81**	0.58	**0.78**	**0.73**

Based on table 3.3, it is observable that the average performance of FQHC in viewpoint estimation is better than HOG. The effectiveness of FQ-PHM in normalizing the φ angle has greatly enhanced the viewpoint estimation performance in Cam 1, Cam 2, and Cam 4. However, the accuracy for Cam 3 is low, but still comparable with HOG at only 3% difference. This is because the intensity of φ angle in Cam 3 is higher compared to the other Cam that may potentially cause huge distortion in the human modelling process. Besides that, the confusion matrix of the viewpoint recognition output using the FQHC to estimate v_1, v_2, and v_3 is illustrated in Figure 3.15.

From the confusion matrix, one can notice that v_1 is confuse with v_2 but almost distinguished itself with v_3 which is reasonable in the sense that v_1 and v_3 are practically two very different viewpoints at a huge gap. On the other hand, v_2 is having vast confusion

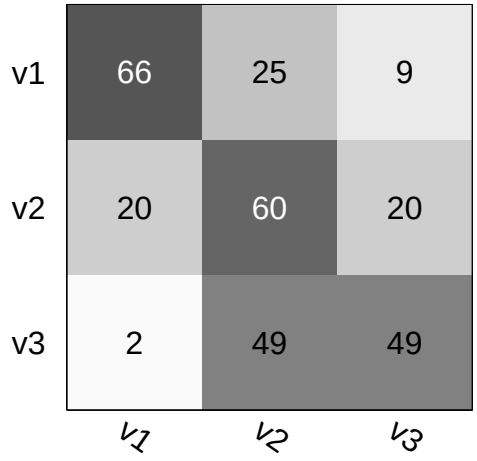


Figure 3.15: Confusion matrix between v_1, v_2, and v_3.

between v_1 and v_3 as v_2 is act as the intermediate stage for the transition from v_1 to v_3. Thus, ambiguous situation such as Figure 3.13 could happen between the slight changes during the transition of v_1 or v_3 into v_2 and thus yield such confusions. From this experiment, one can concludes that, a binary classifier may not be so effective in this case and a sophisticated classifier that is able to model this ambiguity could be a better solution.

3.4.2 Action Recognition

The final objective of the view specific action recognition framework is to recognize an action with independent to the viewpoints. The experiments in this section discuss the effectiveness of the proposed framework from viewpoint detection to action recognition using the VSAM. Spatio-temporal bag of features (Laptev et al., 2008) is employed to extract the motion features with respect to the three viewpoints (v_1, v_3 and v_2) to build the VSAM which will be used in the experiments.

3.4.2 (a) Comparison with human annotation on viewpoint estimation

As mentioned earlier, the viewpoint of an subject can be manually annotated by human. However, this is a tedious job and thus becomes impractical for HMA system. Thus, automated viewpoint estimation algorithm is proposed. The performance of utilizing

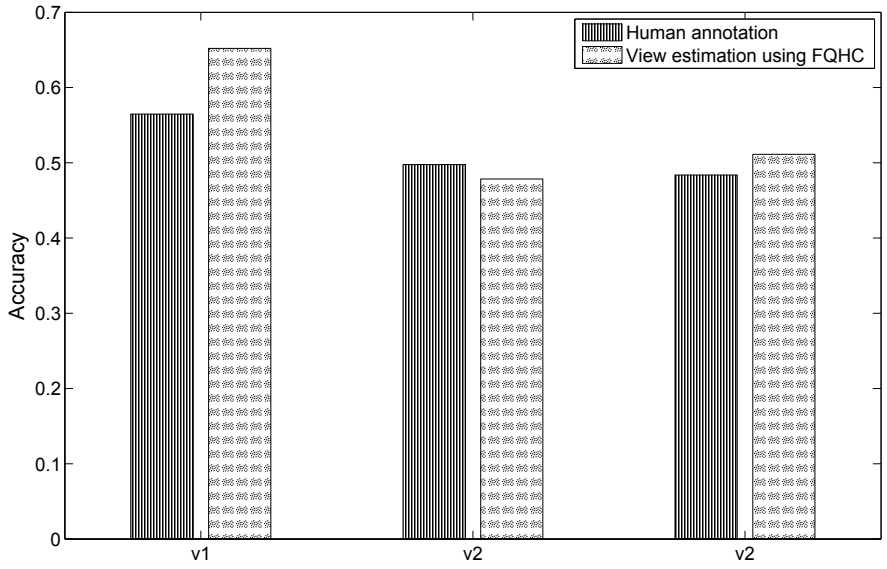


Figure 3.16: Comparison of human action recognition rate by using the view specific action model trained from human annotated viewpoints and view estimation algorithm with fuzzy qualitative human contour.

FQHC in viewpoint estimation algorithm is compared to the human annotated viewpoints (by directly look up from the ground truth) in generating the VSAM. The performances are evaluated through the action recognition task in specific viewpoints.

Based on Figure 3.16, the overall performance in action recognition using the proposed viewpoint estimation algorithm in generating VSAM is comparable with the VSAM generated by human annotation. This had proved that the viewpoint estimation algorithm is capable of generating the decision which is close to the human decision and thus, it is feasible be used in the proposed framework.

3.4.2 (b) Effectiveness of view specific action recognition framework

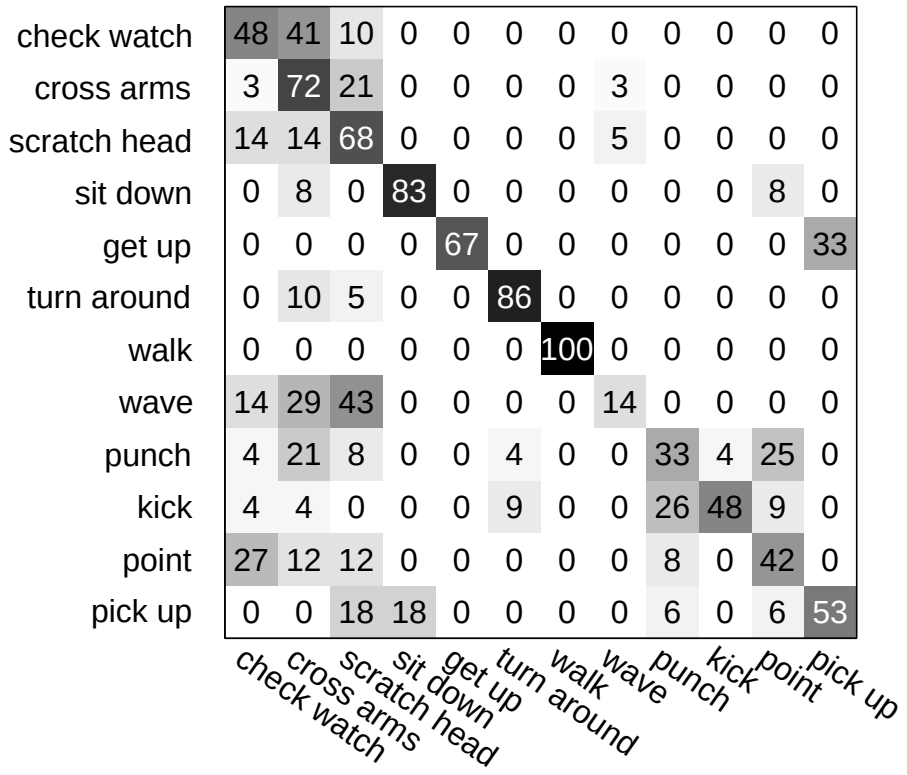
In order to justify the effectiveness of the view specific action recognition framework, action recognition is performed and the results for each specific viewpoint are showed in Figure 3.17. The results convey the message that some of the actions can be recognized better in certain viewpoint. For examples, there are huge different in the recognition rate

for “cross arm”, “scratch head”, and “point” actions in v_1 compared to the v_2 and v_3. This is reasonable as these three actions have significant difference when observe from v_1. In v_2 and v_3 the characteristic of these actions could be similar and become confusing. Such situation will affect the action recognition result. Similarly for the “wave” action, it is found to be better recognize in v_3 compared to the other viewpoints.

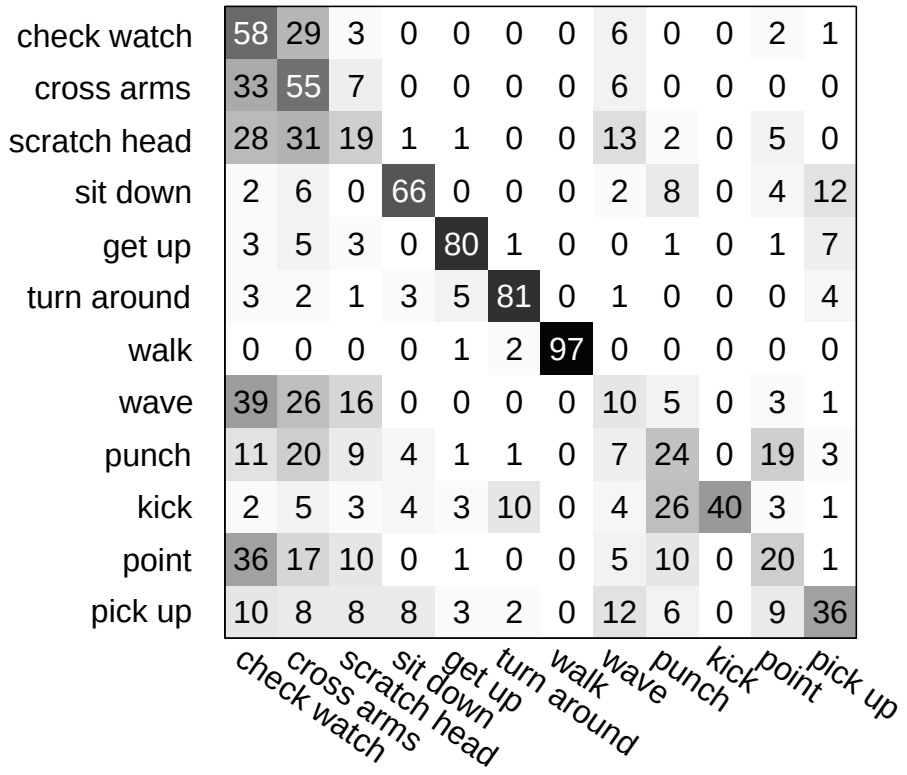
On the other hand, the actions that involved the displacement together with the change of viewpoints will remain high recognition rate across v_1 to v_3 such as “turn around” and “walk” actions as these actions are similar throughout all viewpoints. Nonetheless, although the view specific action recognition framework can eliminate some of the confusion in certain actions, the ambiguity of certain actions still exist as reflected by the confusion matrix in Figure 3.17. For instances, from the overall performance as depicted in Figure 3.17(d), “check watch”, “cross arm”, “scratch head”, “wave”, “punch”, and “point” actions are still confused among each others. As a reminder that, the effectiveness of the proposed framework is also depend on the chosen feature. The features that are good in characterizing the action between each viewpoints will indirectly enhance the performance of the framework.

3.5 Summary

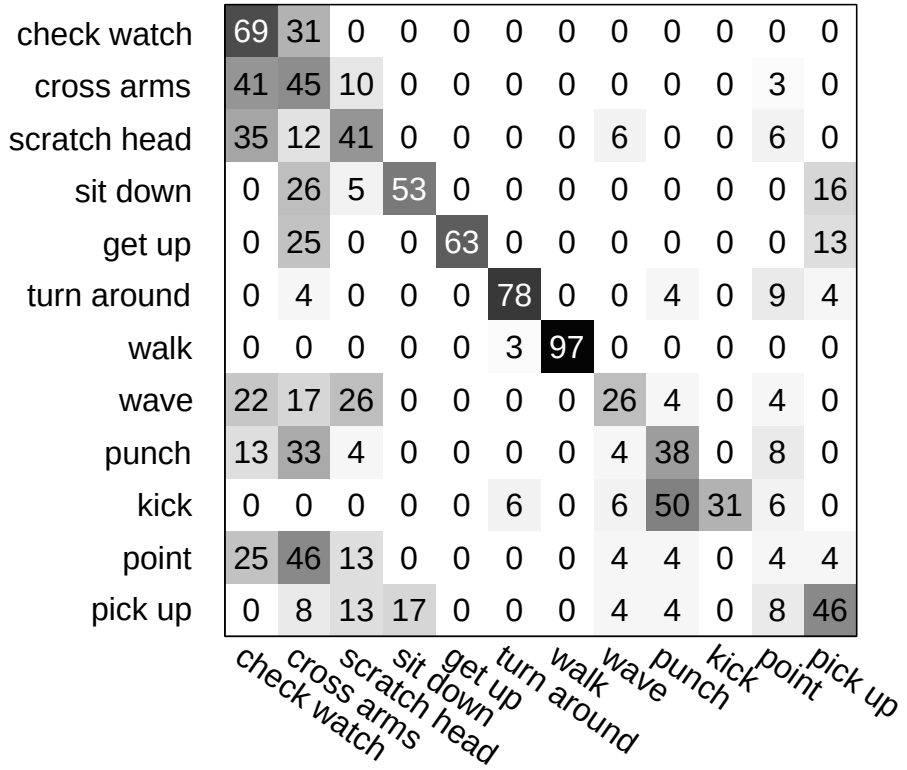
In this chapter, the view specific action recognition framework is introduced. In view estimation module, the FQ-PHM is proposed with its advantage in generating a human model that generalized over the human body size, body anatomy and camera positions. To the extend, the FQHC can be extracted from the FQ-PHM and is verified as a better human contour descriptor compare to HOG to perform viewpoint estimation. Besides that, the overall result in action recognition from different viewpoints showed the effectiveness of view specific action recognition framework but ambiguity still existed in the final result. As a solution, FQRC is introduced in the next chapter to overcome this.



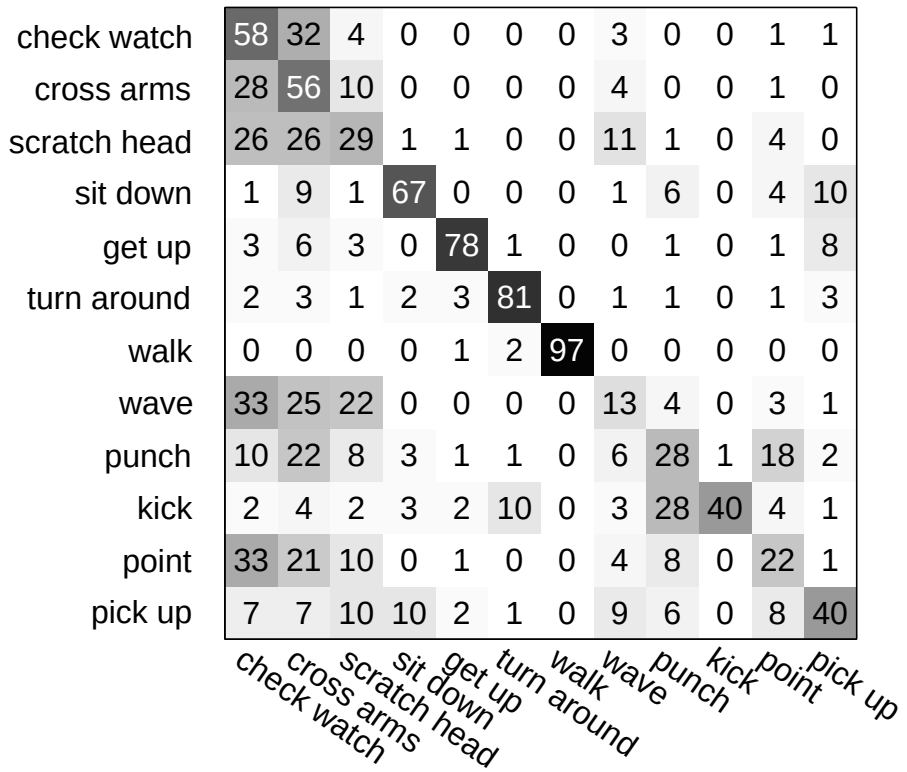
(a) Accuracy on v_1



(b) Accuracy on v_2



(c) Accuracy on v_3



(d) Average accuracy on all views

Figure 3.17: Comparison between action recognition rate from different viewpoints. Higher grayscale intensity means higher recognition rate towards the respective action in the confusion matrix.

CHAPTER 4: FUZZY QUALITATIVE RANK CLASSIFIER

4.1 Introduction

One of the biggest challenges in real world decision making process is to cope with the uncertainty which causes the ambiguity in decision making process. How do humans deal with this growing confusion? In computer vision, this is an important and yet difficult image understanding problem due to their variability, confusion, and uncertainty in classification tasks. As a recall that action recognition and viewpoint estimation are suffered from the ambiguous issue as referring to Figure 1.6 and 3.13 respectively. The human action can be ambiguous in the way that it confused with other actions that have similar factors. While the viewpoint estimation confused itself with another during the slight transition between viewpoints. Both cases are very hard to be distinguished even by human visual inspection.

Besides that, as mentioned in the literature, a surge of interest has sparked in activity recognition recently that takes into account the existence of scene context (Ikizler-Cinbis & Sclaroff, 2010; Marszalek et al., 2009) to enhance the HMA system in view independent manner. It provides prudent information to infer an action with the hypothesis where certain activity occurs with high chances only at certain scenes such as swimming at the coast, walking at the city, and climbing at the mountain scene. Nonetheless, scene images itself can be ambiguous too, for example in Figure 4.1, the Figure 4.1(b) is a Coast scene or a Mountain scene? And thus, instead of HMA, Scene understanding act as the intermediate to study in this context and devise a feasible solution that is capable of modelling these ambiguous cases.

An online survey was conducted towards the public to validate the finding on the subjective human decision making due to the ambiguous case. The survey includes the

(a) Coast (b) ? (c) Mountain

Figure 4.1: Example of ambiguous scene between Coast and Mountain.

people from different ages and background on a set of scene dataset. Interestingly, the outcome of the survey showed that different people tend to give different answer for the same scene image. This has reflected that the scene images can be ambiguous and the conventional crisp or binary classifier is not a good option to deal with this. Technically, the ambiguity in decision making is denoted as "**non-mutually exclusive**" case in this thesis as these images couldn't distinguish itself from another, instead, they share common characteristics. In this chapter, the objective is to deal with the non-mutually exclusive scenario in decision making. As a result, FQRC is proposed to model the non-mutually exclusive case by generating the FQTM and perform inference in multi-label ranking manner.

4.2 Online Survey

Psychological and metaphysical (Forguson & Gopnik, 1998) proved that there is an influence of human factors (background, experience, age, etc.) in decision making. In here, the objective is to show that the research in ambiguous case is subjective to human decision and thus these input images are indeed non-mutually exclusive. For this purpose, an online survey was created with a fair number of scene images, randomly chosen from the OSR scene dataset (Oliva & Torralba, 2001). The online survey was made available for a month and participated by a group of people in the range of 12 to 60 years old from

different backgrounds and countries. Their task is selecting a class that best reflects the given scene accordingly without prior knowledge of what the ground truth is.

Some examples of the results from the survey are shown in Figure 4.2. For the complete survey result, interested reader is encouraged to refer to the corresponding web-page[1]. Based on the Figure 4.2, one can clearly notice that there is a variation of an answer (scene class) for each scene image. For instance, in Figure 4.2(a), although the favorite selection is "Highway" class, the second choice which is "Insidecity" class still occupies noticeable distribution. In qualitative point of view, this observation is valid as the scene image comprises of many buildings to form the city view. Similarly in Figure 4.2(h), the dominant choice is "Forest" class while the second choice of "Mountain" class is still valid.

Nevertheless, one should not overpass the minority choices. For example, in Figure 4.2(g), the dominant selection is a "Mountain" class. However, there are minority participants who selected "Coast", "Opencountry" and "Forest", respectively. Even though these choices are minority, the selections are still valid as it could be noticed that similar appearance between those selected scenes.

[1] http://web.fsktm.um.edu.my/∼cschan/project2.htm

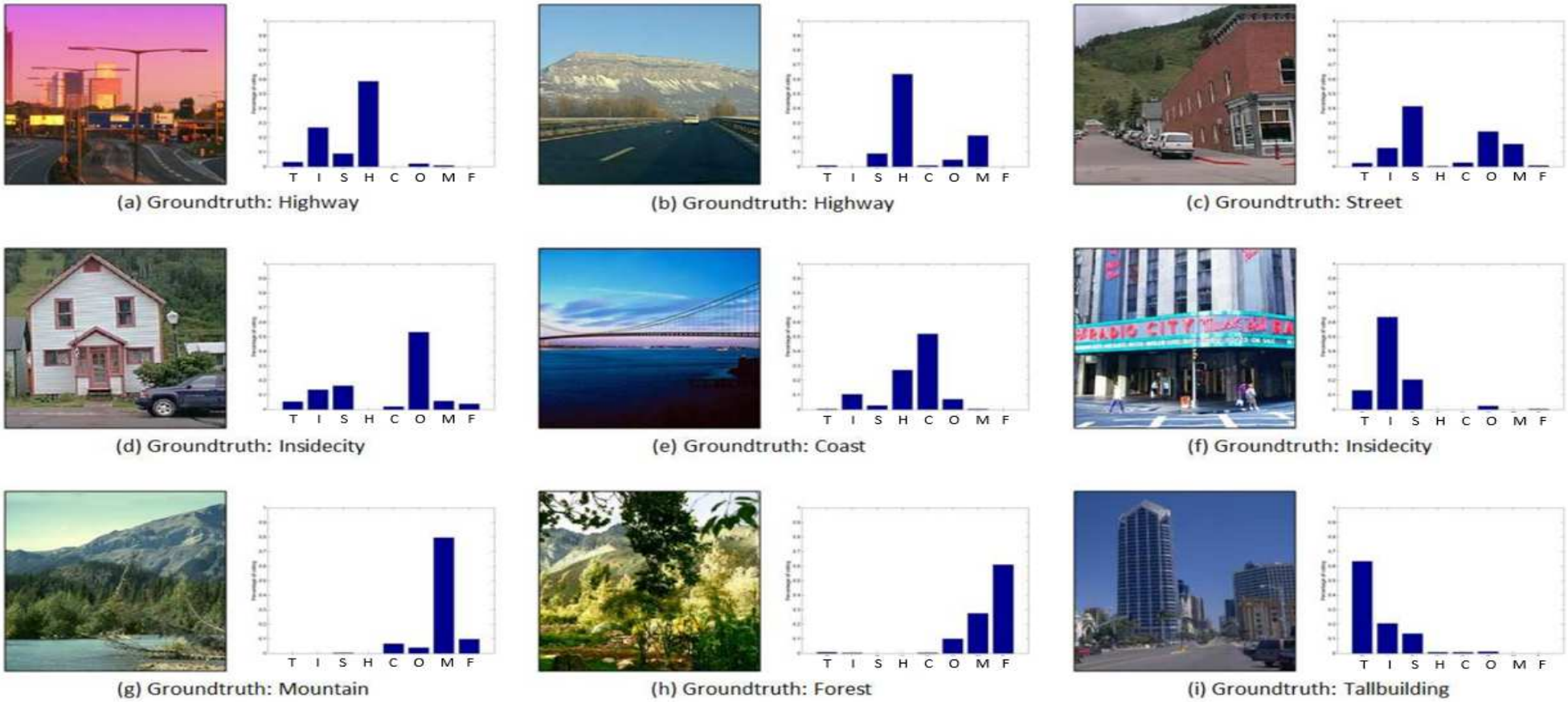


Figure 4.2: Examples of the online survey results. It is validated that scene images are indeed non-mutually exclusive (from left to right, the bar on each histogram represents the distribution of “Tallbuilding, T”, “Insidecity, I”, “Street, S”, “Highway, H”, “Coast, C”, “Opencounty, O”, “Mountain, M” and “Forest, F” accordingly).

Besides that, one could observe that the best result from the histogram of Figure 4.2(a,b,c,e,f,g,h,i) agreed with the ground truth with an exception case Figure 4.2(c). In particular, the image seems to be "Opencountry" more than "Insidecity". This is a very interesting outcome to show that human are bias in identifying a scene image. In summary, the survey shown that assuming scene images are mutually exclusive. With this, simplifying the classification problems (uncertainty, complexity, volatility and ambiguity) to a binary classification task is impractical as it does not reflect how human reasoning is performed in reality. This is similar to the ambiguous case in HMA such as the action recognition and viewpoint estimation problems as discussed in Chapter 3.

4.3 Motivation of Study Non-mutually Exclusive Case in Classification

In general, the task of a classifier (denoted as a function f) is to find a way, which, based on the observations, assigns a sample, $\mathbf{x} \in \mathscr{X}$ to a specified class label, $\mathbf{y} \in (\mathscr{Y} \subseteq \{1,2,\ldots,K\})$, where $\mathscr{X}$ is the input space, $\mathscr{Y}$ is the output space and K is the number of classes label. The task is to estimate a function $(f \in \mathscr{F}) : \mathbf{x} \rightarrow \mathbf{y}$, where $\mathscr{F}$ is the function space. A function f is independent and identically distributed, generated using the input-output pairs according to an unknown distribution $P(\mathbf{x},\mathbf{y})$ so that f can classify unseen samples $(\mathbf{x},\mathbf{y})$,

$$(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_\mathbf{N},\mathbf{y}_\mathbf{N}) \in (\mathscr{X} \times \mathscr{Y})^\mathbf{N} \tag{4.1}$$

The best function f, which one can obtain is the one that minimizes the bound of error represented by a risk function (4.2). However, one must note that, the risk $R(f)$ is unable to directly computed since the probability of $P(\mathbf{x},\mathbf{y})$ is unknown.

$$R(f) = \int loss(f(\mathbf{x}),\mathbf{y})P(\mathbf{x},\mathbf{y}) \tag{4.2}$$

In a non-mutually exclusive case, (4.2) is much difficult to achieve since the respective images are non-mutually exclusive due to the inconsistent of human decision, where different people tend to provide different answers. Theoretically, the importance of the non-mutually exclusive data can be derived from the inequality Chernoff bound (Chernoff, 1952):

$$P\left\{\left|\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}\mathbf{x}_i - E[\mathbf{x}]\right| \geq \varepsilon\right\} \leq 2\exp(-2\mathbf{N}\varepsilon^2) \tag{4.3}$$

This theorem states that the probability of sample mean differ by more than ε from the expected mean is bounded by the exponential that depends on the number of samples $\mathbf{N}$. Note that if more data is available, the probability of deviation error will converge to zero. However, this is not true because of uniform convergence of function space $\mathscr{F}$ (von Luxburg & Schölkopf, 2008). Using the risk function (4.2) one can represent the inequality (4.3) as follows,

$$P\left\{\left|R_{emp}(f) - R(f)\right| \geq \varepsilon\right\} \leq 2\exp(-2\mathbf{N}\varepsilon^2) \tag{4.4}$$

where $R_{emp}(f)$ and $R(f)$ are the empirical and actual risk, respectively. Inequality (4.4) shows that for a certain function f it is highly probable that the empirical error provides good estimates of the actual risk. Luxburg and Scholkopf von Luxburg & Schölkopf (2008) stated that the empirical risk $R_{emp}(f)$ can be inaccurate when $\mathbf{N} \to \infty$ since Chernoff bound only holds for a fixed function f which does not depend on the training data. But in contrary, f does depend on training data. Therefore, they came up with the uniform convergence and obtained the following inequality:

$$P\left\{\sup_{f\in\mathscr{F}}\left|R_{emp}(f) - R(f)\right| \geq \varepsilon\right\} \leq 2\exp(-2\mathbf{N}\varepsilon^2) \tag{4.5}$$

Suppose to have finitely g functions, $\mathscr{F} = \{f_1, f_2, \ldots, f_g\}$ and $\mathscr{C}^i = \left|R_{emp}(f_i) - R(f_i)\right| \geq \varepsilon$, then using the union bound, (4.5) can be represented as:

$$\begin{aligned} & P\left\{ \sup_{f \in \mathscr{F}} \left|R_{emp}(f) - R(f)\right| \geq \varepsilon \right\} \\ & = P(\mathscr{C}^1 \vee \mathscr{C}^2 \vee \cdots \vee \mathscr{C}^g) \\ & = \sum_{i=1}^{g} P(\mathscr{C}^i) - \left\{ \mathscr{D}^2 + \mathscr{D}^3 + \cdots + \mathscr{D}^g \right\} \\ & \leq \underbrace{2g \exp(-2\mathbf{N}\varepsilon^2)}_{\text{1st term}} - \underbrace{\text{bound}(\mathscr{D}^2 + \mathscr{D}^3 + \cdots + \mathscr{D}^g)}_{\text{2nd term}} \end{aligned} \tag{4.6}$$

where $\mathscr{D}^i$ is the sum of the probabilities of every combination of i event, e.g, $\mathscr{D}^g = P(\mathscr{C}^1 \wedge \mathscr{C}^2 \wedge \cdots \wedge \mathscr{C}^g)$. This leads to a bound which states that the probability that empirical risk is close to the actual risk is upper bounded by two terms. The first term is the error bound because of the mutually exclusive data and the second term is due to the non-mutually exclusive data. Most of the conventional classification methods, however, only utilize the mutually exclusive part. In contrast, the proposed methods - FQRC models both the mutually and non-mutually exclusive parts.

4.4 Implementation of Fuzzy Qualitative Rank Classifier (FQRC)

The aim of the FQRC is to model the non-mutually exclusive data which result in a trained model namely the Fuzzy Qualitative Trained Model (FQTM) in the training step. This model is then used as the classifier to infer the testing samples and result in a multi-label ranking output instead of crisp or binary classification result. To begin with, the FQRC utilized the FQS to build the Two Dimensional Fuzzy Qualitative State (2D-FQstate). With this, it learns the feature distribution in fuzzy qualitative manner that capture the characteristic of non-mutually exclusive.

4.4.1 2D Fuzzy Qualitative State

A 2D-FQstate is denoted as $QST^{(\mathbf{i},\mathbf{j})}$ which can be composed from two qualitative states with one along the x translation component Q_x^t and another one along the y translation component Q_y^t in the FQS. x and y axis can respectively represents the scale of feature or attribute values in the training. Figure 4.3 shows an example of a 2D-FQstate.

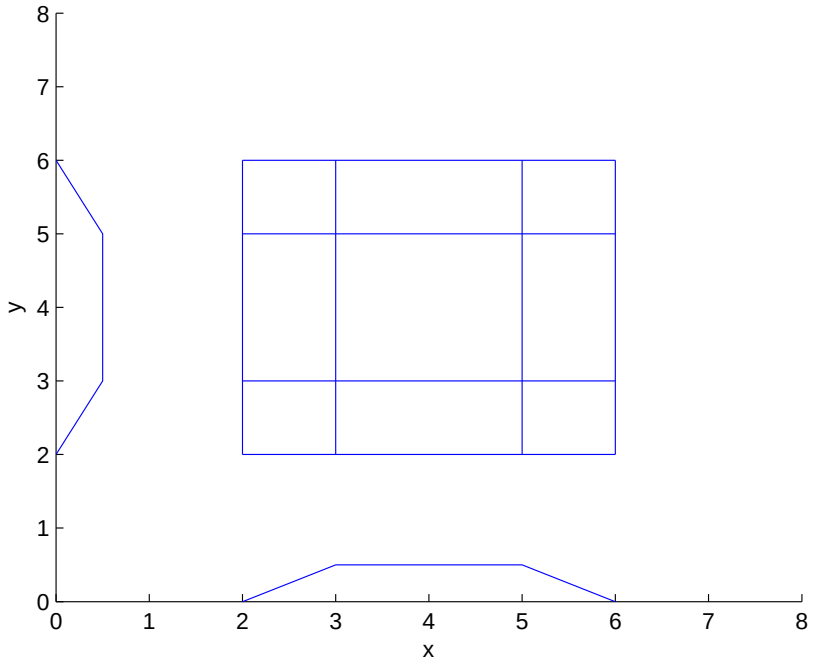


Figure 4.3: An illustration of 2D-FQstate in a fuzzy quantity space.

4.4.2 Training

In order to build a FQTM from a set of training data, $T = \{T_x, T_y\}$, the feature values of the training data T_x and T_y must be first normalized into the range of [-1 1]. Secondly, these normalized training data, T' are mapped into the FQS in order to build the FQTM ($T' \mapsto FQS$). In this context, let's assume that a total of $\mathbf{I} \times \mathbf{J}$ 2D-FQstate are built in the FQS. The FQTM can be represented as:

$$\text{FQTM} = \{QST^{(1,1)}, QST^{(1,3)}, \cdots, QST^{(2,2)}, \cdots, QST^{(\mathbf{I},\mathbf{J})}\} \tag{4.7}$$

In training the model, a weight function w is defined as:

$$w^k = \frac{L_k}{\sum_{k=1}^{K} L_k} \tag{4.8}$$

where L_k is the occurrence number of T' of a particular subject class, k in a 2D-FQstate. Therefore, in each 2D-FQstates in the trained model, there is a weight corresponding to each class,

$$QST^{(\mathbf{i},\mathbf{j})} = \{w^1, w^2, \cdots, w^K\} \tag{4.9}$$

where K is the total number of class in the classification task. For example, if $K = 3$, each 2D-FQstate in the FQTM will be represented as $QST^{(i,j)} = \{w^1, w^2, w^3\}$ and $\sum w = 1$. The advantage of this approach is that the final output of FQTM is capable of modelling the non-mutually exclusive data. For illustration purpose, a simple FQTM with mutually-exclusive class, $K = 3$ is shown in Figure 4.4.

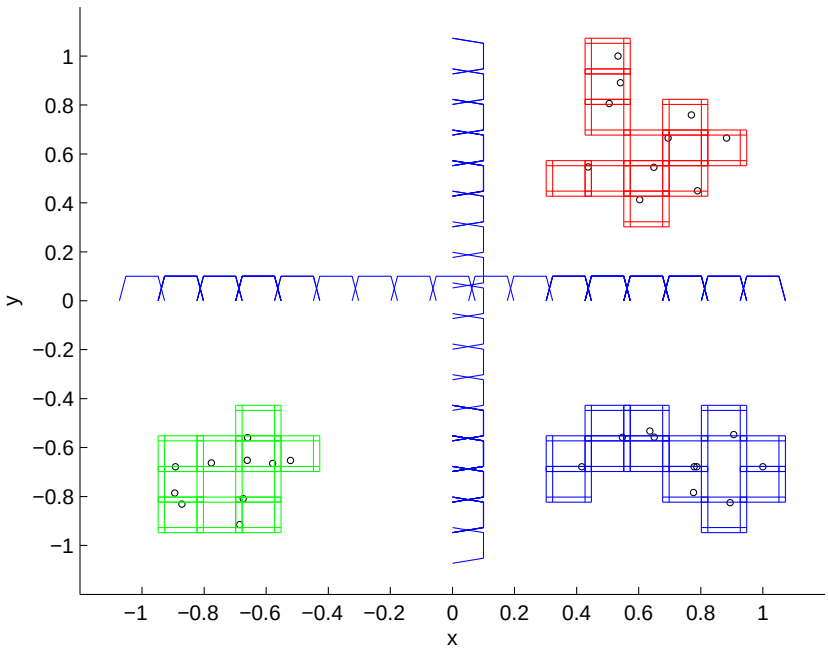


Figure 4.4: An example of fuzzy qualitative trained model with $K = 3$.

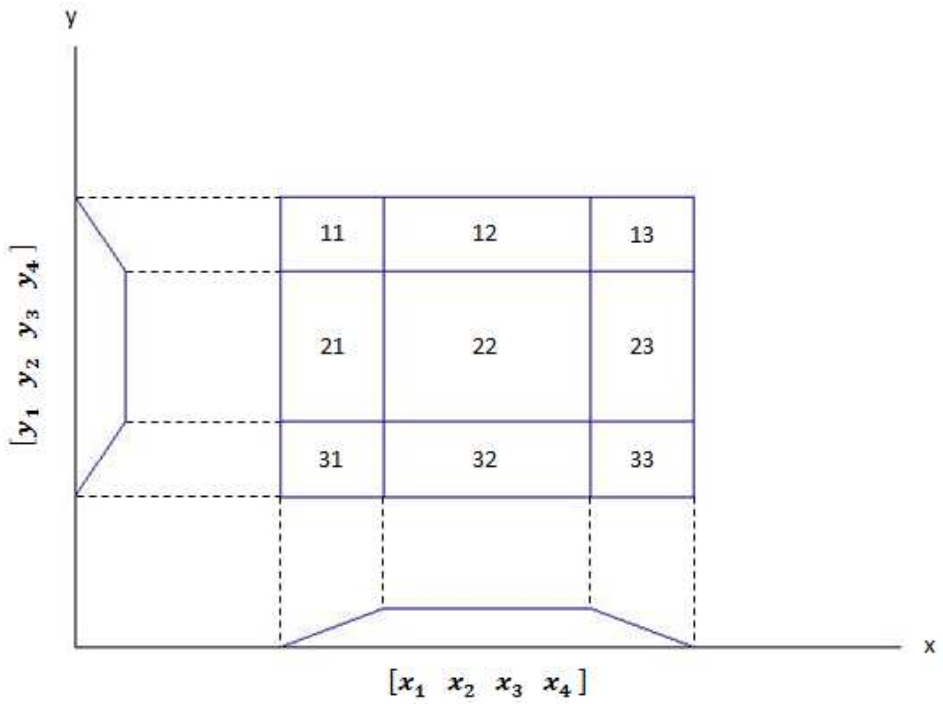


Figure 4.5: Fuzzy Qualitative Partition.

4.4.3 Inference

In the classification stage, let's denote $\mathbf{d} = (\mathbf{d}_x, \mathbf{d}_y)$ as a normalized testing data represented with two feature values, $\mathbf{d}_x$ and $\mathbf{d}_x$. In order to choose the most likely 2D-FQstate that $\mathbf{d}$ belongs to, Fuzzy Qualitative Partition (FQP) is introduced in the inference process. FQP consists of nine partitions derived from the 2D-FQstate (Table 4.2) where each partition has different degree of membership, μ, with the calculation is as Table 4.1.

Table 4.1: Notation of the FQP.

	$x_1 \leqslant \mathbf{d}_x \leqslant x_2$	$x_2 \leqslant \mathbf{d}_x \leqslant x_3$	$x_3 \leqslant \mathbf{d}_x \leqslant x_4$
$y_3 \leqslant \mathbf{d}_y \leqslant y_4$	$\mathbf{P}^{(11)}$	$\mathbf{P}^{(12)}$	$\mathbf{P}^{(13)}$
$y_2 \leqslant \mathbf{d}_y \leqslant y_3$	$\mathbf{P}^{(21)}$	$\mathbf{P}^{(22)}$	$\mathbf{P}^{(23)}$
$y_1 \leqslant \mathbf{d}_y \leqslant y_2$	$\mathbf{P}^{(31)}$	$\mathbf{P}^{(32)}$	$\mathbf{P}^{(33)}$

For example, the FQP gives the intuition of: $\mathbf{P}^{(22)}$ denote the FQP where both the degree of membership of the x and y axis are 1. $\mathbf{P}^{(12)}, \mathbf{P}^{(21)}, \mathbf{P}^{(32)}, \mathbf{P}^{(23)}$ denote the FQP in

which either degree of membership of the x or y axis is 1. $\mathbf{P}^{(11)}, \mathbf{P}^{(13)}, \mathbf{P}^{(31)}, \mathbf{P}^{(33)}$ denote the FQP in which neither degree of membership of the x or y axis is 1.

Table 4.2: The μ calculation in FQP.

	$x_1 \leqslant \mathbf{d}_x \leqslant x_2$	$x_2 \leqslant \mathbf{d}_x \leqslant x_3$	$x_3 \leqslant \mathbf{d}_x \leqslant x_4$
$y_3 \leqslant \mathbf{d}_y \leqslant y_4$	$\frac{\mathbf{d}_x - x_1}{x_2 - x_1} \times \frac{y_4 - \mathbf{d}_y}{y_4 - y_3}$	$1 \times \frac{y_4 - \mathbf{d}_y}{y_4 - y_3}$	$\frac{x_4 - \mathbf{d}_x}{x_4 - x_3} \times \frac{y_4 - \mathbf{d}_y}{y_4 - y_3}$
$y_2 \leqslant \mathbf{d}_y \leqslant y_3$	$\frac{\mathbf{d}_x - x_1}{x_2 - x_1} \times 1$	1×1	$\frac{x_4 - \mathbf{d}_x}{x_4 - x_3} \times 1$
$y_1 \leqslant \mathbf{d}_y \leqslant y_2$	$\frac{\mathbf{d}_x - x_1}{x_2 - x_1} \times \frac{\mathbf{d}_y - y_1}{y_2 - y_1}$	$1 \times \frac{\mathbf{d}_y - y_1}{y_2 - y_1}$	$\frac{x_4 - \mathbf{d}_x}{x_4 - x_3} \times \frac{\mathbf{d}_y - y_1}{y_2 - y_1}$

However, there are cases where $\mathbf{d}$ will fall into more than one 2D-FQstates, this is denoted as $l > 1$ where $l = \{1, 2, 4\}$. This will happen when $\mathbf{d}$ falls into the FQP as below:

- $\mathbf{d}$ belongs to two 2D-FQstates, $l = 2$ when it falls into $\mathbf{P}^{(12)}, \mathbf{P}^{(21)}, \mathbf{P}^{(32)}$, and $\mathbf{P}^{(23)}$.
- $\mathbf{d}$ belongs to four 2D-FQstates, $l = 4$ when it falls into $\mathbf{P}^{(11)}, \mathbf{P}^{(13)}, \mathbf{P}^{(31)}$, and $\mathbf{P}^{(33)}$.

In order to choose the most possible 2D-FQstate where $\mathbf{d}$ belongs, a degree of membership for each 2D-FQstate corresponds to $\mathbf{d}$, $\mu_{\mathbf{d}}$ is calculated based on Table 4.2. From the calculation of $\mu_{\mathbf{d}}$, the 2D-FQstate that holds the highest degree of membership towards $\mathbf{d}$, $QST^{\mathbf{d}}$ will be selected,

$$QST^{\mathbf{d}} = max\{\mu_d^{QST^1}, \mu_d^{QST^2}, \cdots, \mu_d^{QST^l}\} \tag{4.10}$$

In the end, the corresponding weights, w are output as the ranking result to $\mathbf{d}$.

$$QST^{\mathbf{d}} = \{w^1, w^2, w^3\} \tag{4.11}$$

For example, (4.11) shows that **d** is holding the weights, w^1 that belongs to Class 1, w^2 that belongs to Class 2, and w^3 which belongs to Class 3. This is the advantage of the proposed approach that any possible class that **d** could belongs to is taking into account. This is in contrast with the conventional crisp or binary classification solutions where they assumed that one sample can only classified into one class.

4.5 Experiments and Discussions

In order to test the effectiveness and the robustness of the proposed FQRC, scene understanding is applied in advance instead of HMA as the ground truth is available from the survey. The Outdoor Scene Recognition (OSR) Dataset (Oliva & Torralba, 2001) is used as it is the most popular scene dataset and features (GIST) are provided. A total of four classes of the scenes are used throughout the experiments which are "Insidecity", "Coast", "Opencountry", and "Forest". Examples of those scenes are shown in Figure 4.6. These four classes of the scenes are chosen in the experiments because each of them has their own unique characteristics that corresponds to the degree of "Openness" (exposure of the open space) and the degree of "Naturalness" (coverage of natural substances) (Oliva & Torralba, 2001). For example, the coast scenes have high value of Openness while the forest scenes have low value of Openness. Figure 4.7 shows the original distribution of the four classes of the scenes that correspond to the degree of the attributes (also known as features). The attributes (i.e.: degree of "Openness" and degree of "Naturalness") introduced by Oliva & Torralba (2001) are called the spatial envelope properties. The score of each scene image corresponds to each attribute are then further processed by Parikh & Grauman (2011) which also will be used in this experiments. The source of these attribute values are publicly available at http://ttic.uchicago.edu/ dparikh/relative.html.

In this experiment, the Insidecity scenes and Opencountry scenes from the OSR scene dataset are chosen for the evaluation purpose. This is because from these two scene

(a) Examples of the Insidecity scene

(b) Examples of the Coast scene

(c) Examples of the Opencountry scene

(d) Examples of the Forest scene

Figure 4.6: Examples of the scenes from four classes (Oliva & Torralba, 2001).

classes, some of the scene images have possess the characteristic of other classes that caused the ambiguous in classification. Thus they are more suitable to be tested in the experiment to meet the objective. 'Leave-one-out' method is used for the classification task where each of the scene images will be classified into the four scene classes. Figure 4.8 illustrates the "Insidecity" scenes and Table 4.3 presents the classification results. Similarly, Figure 4.9 illustrates the "Opencountry" scenes and Table 4.4 presents the classification results.

From the results (Table 4.3-4.4), it shows the effectiveness and robustness of the pro-

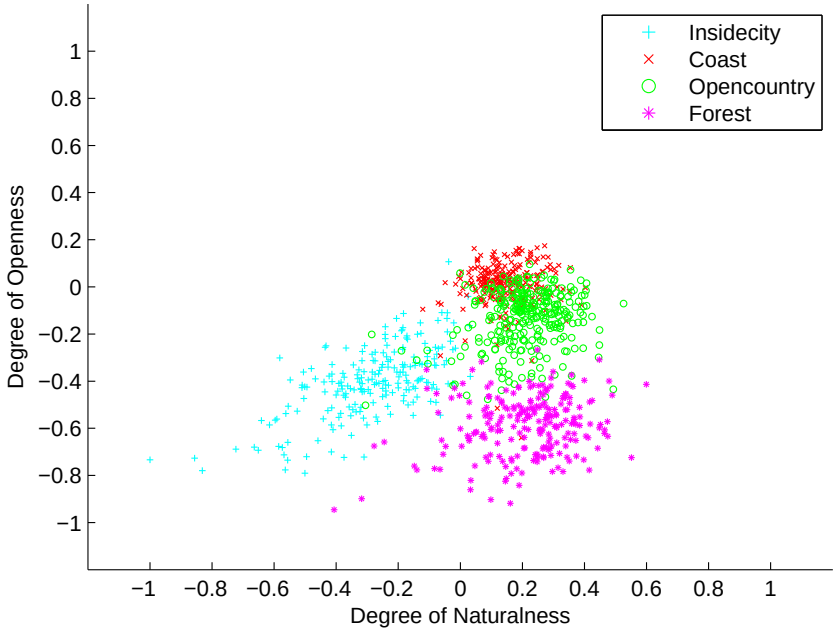
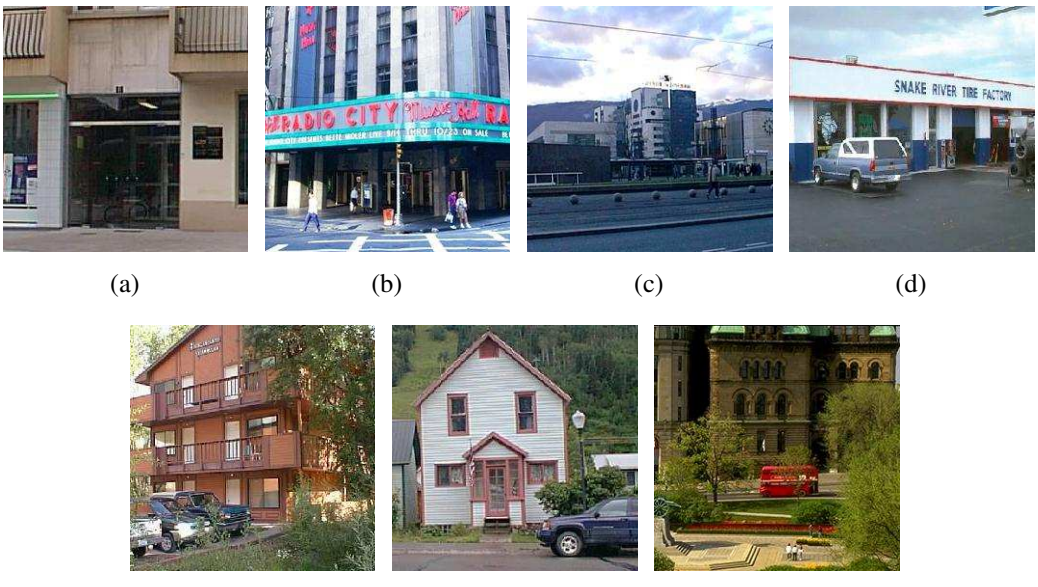


Figure 4.7: The distribution of four classes of scenes correspond to the degree of the attributes. One can notice that some of the scene images are crossover in term of the attribute distribution. This means that these scene images are not mutually-exclusive and potentially ambiguous to the other scene images.



(a) (b) (c) (d)

(e) (f) (g)

Figure 4.8: Examples of Insidecity annotated scenes (Oliva & Torralba, 2001).

posed approach. For example, the proposed method confidently classified both the Figure 4.8(a), 4.8(b) and Figure 4.9(a) as “Insidecity” or “Opencountry” class respectively with $w = 1$. This is because Figure 4.8(a) and 4.8(b) have low degree of Openness and low degree of Naturalness which are the characteristics of “Insidecity” scenes. Figure 4.8(c)

Table 4.3: Inference outputs for the Insidecity scenes.

Scene	Weight, w			
(Figure)	Insidecity	Coast	Opencountry	Forest
4.8(a)	1	0	0	0
4.8(b)	1	0	0	0
4.8(c)	0.7273	0.2727	0	0
4.8(d)	0.7273	0.2727	0	0
4.8(e)	0.1250	0	0.1250	0.7500
4.8(f)	0.8235	0	0	0.1765
4.8(g)	0.8235	0	0	0.1765

and 4.8(d) being classified as the combination of "Insidecity" class and also "Coast" class because they have the characteristics of "Coast" scenes which are high degree of Openness and high degree of Naturalness. Figure 4.9(b) to 4.9(d) show the combination of "Coast" class and "Opencountry" class, respectively. On the other hand, Figure 4.8(e) to 4.8(g) hold the degrees that belong to insidecity class and also "Forest" class because of the low degree of Openness and the high degree of Naturalness is detected from those scenes and these are the characteristics of the "Forest" scenes. However, they do not hold the degree to the "Coast" class because their degree of Openness does not reach the level as a "Coast" scenes.

4.5.1 System Accuracy

This section is to test the accuracy of the proposed approach at classifying the scene images. The ground truth for this is provided by the (OSR) Dataset Oliva & Torralba (2001). The results are based on the average outcome from 20 iterations with 70% of training data and 30% of testing data. In addition, the effectiveness of using different resolution of FQS are tested with $N = \{4, 8, 12, 16\}$. Figure 4.10 shows the examples of the FQTM built from different resolutions of the FQS and the corresponding results are

(a) (b) (c) (d)

(e) (f) (g)

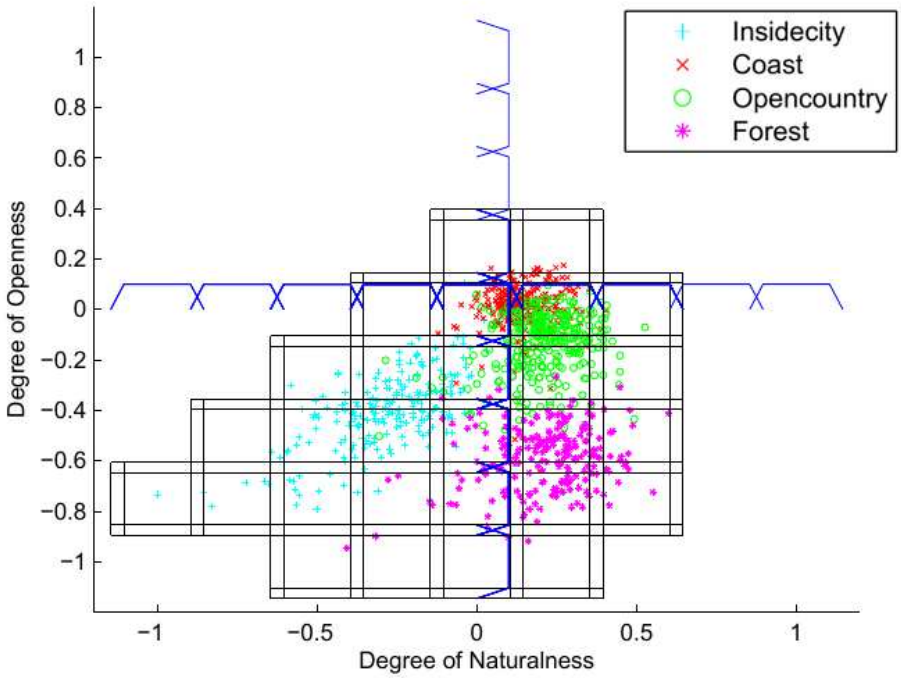
Figure 4.9: Examples of Opencountry annotated scenes (Oliva & Torralba, 2001).

Table 4.4: Inference outputs for the Opencountry scenes.

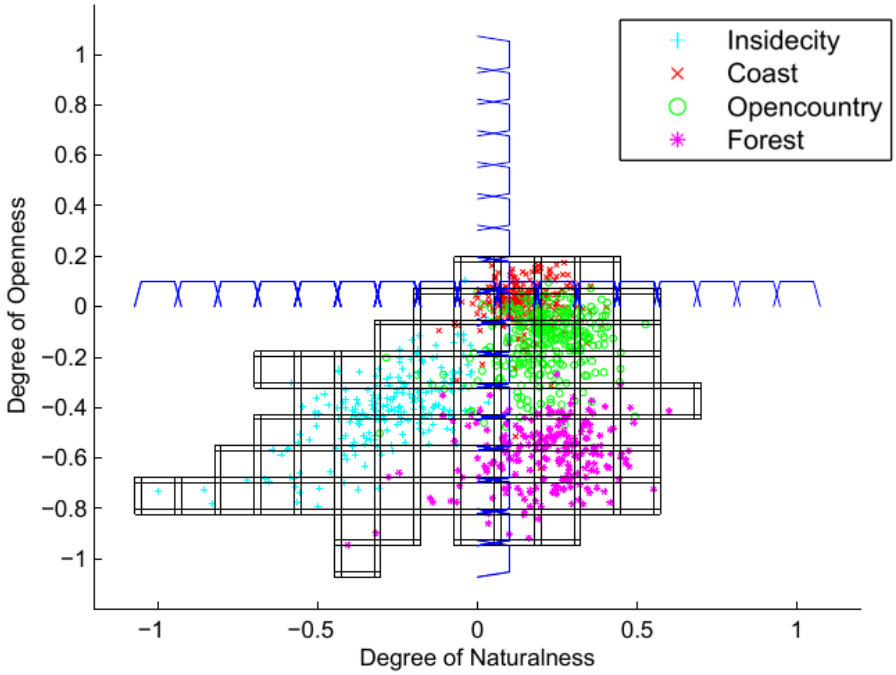
Scene	Weight, w			
(Figure)	Insidecity	Coast	Opencountry	Forest
4.9(a)	0	0	1	0
4.9(b)	0	0.1111	0.8889	0
4.9(c)	0	0.0435	0.9130	0.0435
4.9(d)	0	0.3387	0.6613	0
4.9(e)	0	0	0.6471	0.3529
4.9(f)	0	0	0.0233	0.9767
4.9(g)	0	0.0435	0.3913	0.5652

shown in Figure 4.11.

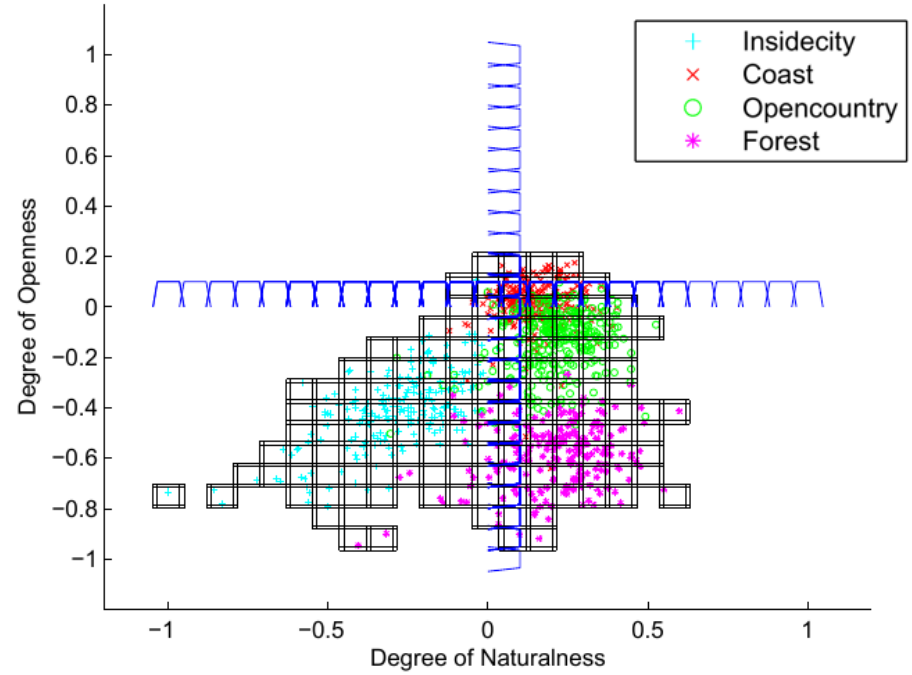
From the results, one can observe that, in general, the proposed approach has achieved stable accuracy for different FQS resolutions. The average accuracy (%) is 80.5 ± 2.5 and it is found that $N = 8$ holds the best accuracy. The poorest result is when $N = 4$ where our proposed approach results in the confusion between "Opencountry" and "Coast" scenes. This is because these two scenes are quite similar to each other and thus many crossover


Insidecity
Coast
Opencountry
Forest
Degree of Openness
Degree of Naturalness
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1

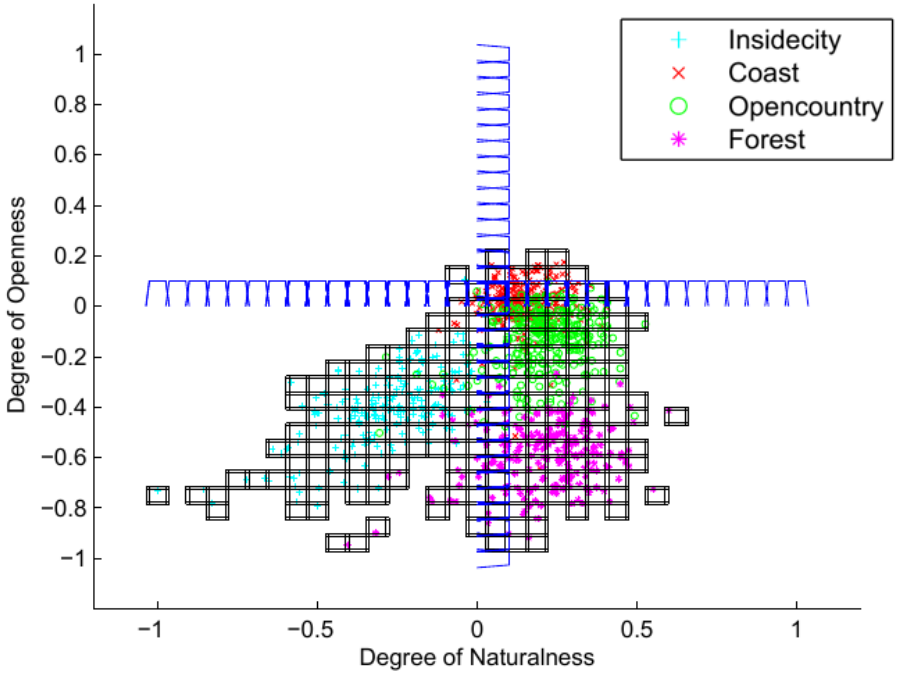
(a) $N = 4$


Insidecity
Coast
Opencountry
Forest
Degree of Openness
Degree of Naturalness
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1

(b) $N = 8$



(c) $N = 12$



(d) $N = 16$

Figure 4.10: Examples of fuzzy qualitative trained model with different N.

	Insidecity	Coast	Opencountry	Forest
Insidecity	0.81	0.03	0.10	0.05
Coast	0.00	0.94	0.06	0.00
Opencountry	0.03	0.51	0.41	0.05
Forest	0.01	0.00	0.03	0.95

(a) $N = 4$

	Insidecity	Coast	Opencountry	Forest
Insidecity	0.89	0.02	0.05	0.02
Coast	0.01	0.76	0.21	0.01
Opencountry	0.04	0.17	0.73	0.05
Forest	0.02	0.00	0.03	0.92

(b) $N = 8$

	Insidecity	Coast	Opencountry	Forest
Insidecity	0.88	0.02	0.03	0.01
Coast	0.01	0.75	0.21	0.01
Opencountry	0.03	0.15	0.74	0.04
Forest	0.01	0.00	0.03	0.91

(c) $N = 12$

	Insidecity	Coast	Opencountry	Forest
Insidecity	0.83	0.01	0.03	0.01
Coast	0.01	0.77	0.18	0.01
Opencountry	0.02	0.19	0.72	0.04
Forest	0.01	0.00	0.03	0.87

(d) $N = 16$

Figure 4.11: Confusion matrix of crisp classification results for different N.

data in the FQTM. Too large the resolution will lead to ineffective classification and thus big confusion happened with another scene classes.

4.5.2 Comparison with K-nearest Neighbour

The performance of the proposed method using different portion of training data (70%, 50%, and 30%) is tested and compared to crisp classifier, K-nearest Neighbour (KNN). For fair comparison, the crisp classification result of the FQRC is obtain by selecting the scene class that has the highest weight, w from the ranking result. Based on Table 4.5, first of all it shows that the results are inline with KNN and this proves that the proposed method is capable of performing crisp classification. However, FQRC is better than KNN in terms that it does not assume that scene classes are mutually exclusive. Secondly, the classification results does not effected much by the size of training data as the accuracy different by using 30% and 70% of the training data is only $\pm$4%.

Table 4.5: Comparison with KNN based on different % of training data.

Training data	Accuracy for FQS (%)				Accuracy for KNN (%)			
(%)	Insidecity	Coast	Opencountry	Forest	Insidecity	Coast	Opencountry	Forest
70	0.89	0.76	0.73	0.92	0.89	0.70	0.65	0.92
50	0.88	0.77	0.69	0.91	0.91	0.70	0.67	0.91
30	0.87	0.77	0.69	0.89	0.89	0.70	0.70	0.91

4.6 Summary

In this chapter, the online survey has validated that scene images are non-mutually exclusive and the conventional crisp classification methods might not work effectively on this ambiguous case. The proposed FQRC has been discussed and showed its effectiveness in

modelling the non-mutually exclusive data, particularly in scene images. However, there are two limitations in the proposed methods: 1) Choosing the optimal N in constructing the FQS by trial and error is not practical as it is a tedious and time consuming job; and 2) It is unable to support multi-dimension classification which means that the current method only able to perform classification with maximum two feature dimensions. A more sophisticated FQRC method is proposed in the next chapter to overcome these limitations namely the DFQRC.

CHAPTER 5: DYNAMIC FUZZY QUALITATIVE RANK CLASSIFIER

5.1 Introduction

As mentioned in chapter 4 where there are situations with the decisions are ambiguous and these phenomena are denoted as non-mutually exclusive case in the thesis. HMA and scene understanding are in this category and conventional crisp or binary classification methods are less effective in modelling these ambiguous cases. This is because, the crisp or binary classifier tends to ignore or overlook the possible class that also described the ambiguous sample. This is deviated from how human reasoning is done and it might deteriorate the overall system performance.

This notion became popular among researchers where instead of single label, multi-label classification framework is proposed. The notable pioneer works are by Boutell et al. (2004); M.-L. Zhang & Zhou (2007) in scene understanding. However, these approaches are not practical due to: firstly, the work requires human intervention to manually annotate the multi-label training data which is a tedious job. Secondly it leads to a large number of classes with the sparse number of sample (Tsoumakas & Katakis, 2007) which the annotated image's classes are potentially bias to inconsistently human decision (Forguson & Gopnik, 1998). Thus, FQRC is proposed in chapter 4 which is capable of modelling the non-mutually exclusive data that addresses the above shortcomings.

However, FQRC requires to find the appropriate resolution (N) to build the FQS. In specific, the model parameters are chosen manually based on prior information and in a trial-and-error manner. This is a heuristic and time consuming approach. Besides that, the FQRC do not support multi-dimension classification tasks. In order to cope with these, DFQRC is proposed. It is capable to learn the fuzzy tuple adaptively with the training data to build the FQTM. Furthermore, it is proved in the experiments that DFQRC is

more effective and efficient for inference process.

5.2 Implementation of Dynamic Fuzzy Qualitative Rank Classifier (DFQRC)

Similarly to FQRC, the aim of the DFQRC is to model the non-mutually exclusive data. However, in this extension, the training step is done without heuristic methods or trial and error. Instead, the algorithm learns the 4-tuple fuzzy number (Shen & Leitch, 1993) from the training data adaptively which contributes to achieve a more sophisticated FQTM that is capable of inferring the multi-label ranking output which is close to how human makes decision.

5.2.1 Learning 4-tuple Fuzzy Number

According to Chan & Liu (2009); Chan et al. (2008, 2007); Liu et al. (2008a,b, 2009); Shen & Leitch (1993), 4-tuple fuzzy number is a better qualitative representation as the representation has high resolution and good flexibility. In this work, the objective is to dynamically learn the best composition of 4-tuple fuzzy number in the FQS that best represent the FQTM from the training data. This is to avoid trial and errors in choosing the suitable resolution of FQS in the FQRC as presented in previous chapter. In this work, the learning of the 4-tuple number is proposed by utilizing the histogram approach. As for the learning outcome, the dominant region of the 4-tuple fuzzy number indicates the mutually exclusive part, while the intersection between 4-tuple fuzzy number indicates the non-mutually exclusive part, as shown in Figure 5.1.

Let's denote the 4-tuple fuzzy number here as $\mathbf{m} = [a\ b\ \alpha\ \beta]$ with the condition $a < b$ and $ab > 0$. The final output of FQTM will be $J \times K$ matrix containing 4-tuple fuzzy number for each feature, j and class, k as in (5.1). Those 4-tuple fuzzy number are represented in the form as $\mathbf{m}_{jk} = [a\ b\ \alpha\ \beta]_{jk}$. One can notice that in the FQRC, the FQTM is utilizing multiple 4-tuple numbers in Q_x^t and Q_y^t that representing the 2D-

FQstate to build the FQTM. This is contradict to DFQRC where the learning module output only one 4-tuple fuzzy number for each feature. Such representation allows to perform multi-dimension classification compare to the previous FQRC where it is only limited to two dimensions. Furthermore this representation is opposed to Boutell et al. (2004); M.-L. Zhang & Zhou (2007) in scene understanding where human intervention in manually annotates the training data is not required. Here, the training data is modeled as (5.1).

$$\text{FQTM} = \begin{bmatrix} \mathbf{m}_{11} & \mathbf{m}_{12} & \cdots & \mathbf{m}_{1K} \\ \mathbf{m}_{21} & \mathbf{m}_{22} & \cdots & \mathbf{m}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m}_{J1} & \mathbf{m}_{J2} & \cdots & \mathbf{m}_{JK} \end{bmatrix} \tag{5.1}$$

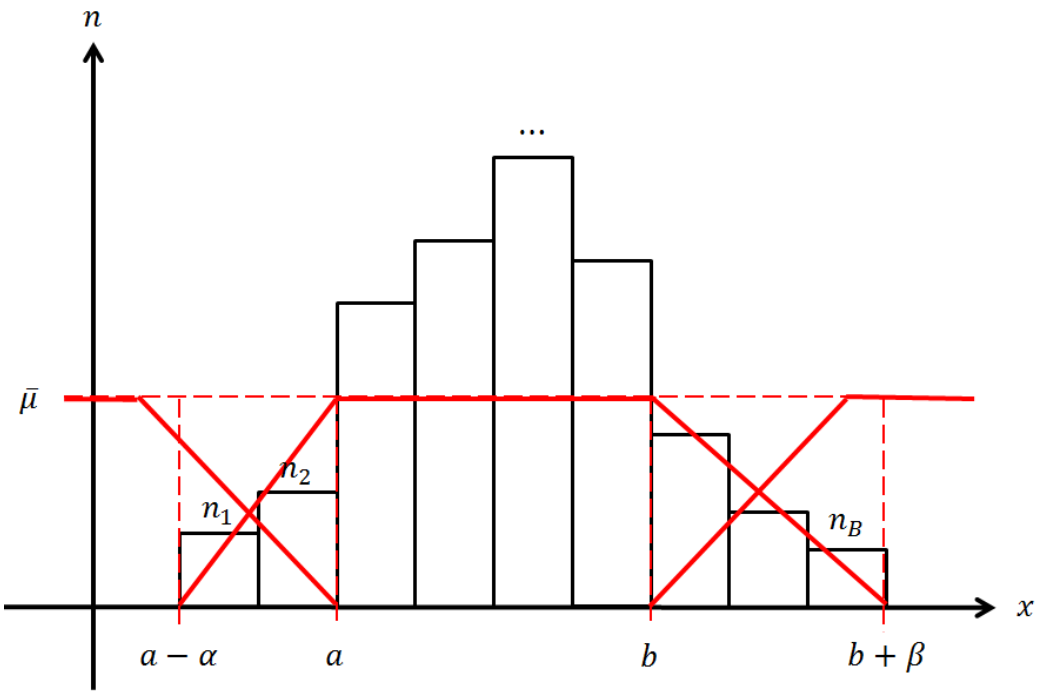


Figure 5.1: Parametric representation of a histogram, x is the feature value, $\mathfrak{n}$ denotes the occurrence of training data from its respective bin $\mathfrak{n}_1, \mathfrak{n}_2, \ldots, \mathfrak{n}_B$. a and b represent the lower and upper bound of $\bar{\mu}$, while $a-\alpha$ and $b+\beta$ represent the minimum and maximum of x value. The dominant region (mutually exclusive) is the area of $[a,b]$. The intersection area (non-mutually exclusive) is the areas of $[a-\alpha,a]$ and $[b,b+\beta]$.

The representation in (5.1) is to conserve the appropriate membership function, **m**, of

each respective feature (row) for each class (column). In order to learn the 4-tuple fuzzy number, histogram representation was chosen. As illustrated in Figure 5.1, The histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is the frequency density over the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data.

More specifically, a histogram is a function that counts the number of observations, $\mathfrak{n}$, that fall into each of the disjoint categories (known as bins), whereas the graph of a histogram is merely one way to represent a histogram. Thus, if let $\mathfrak{N}$ be the total number of observations and B be the number of bins, then $\mathfrak{N} = \sum_{i=1}^{B} \mathfrak{n}_i$. In the proposed method, for every feature and class label, $\mathbf{x}_{jk} = \{x_i^{jk}\}_{i=1}^{\mathfrak{N}}$, a histogram is created to obtain the $\mathbf{m}_{jk}$.

The histogram is utilized in representing the occurrence of the training data to the corresponding feature values with a desire bin width. There is no "best" number of bins, and different bin sizes would reveal different features of the data. There are a few theoreticians have attempted to determine an optimal number of bins (Dalal & Triggs, 2005; Shimazaki & Shinomoto, 2007; Wand, 1997), but these methods generally make strong assumptions about the shape of distribution. Depending on the actual data distribution and the goals of analysis, different bin number may be appropriate. An experiment is usually needed for this purpose. To find the bin width, $\mathfrak{v}$,

$$\left\lceil \mathfrak{v} = \frac{\wedge x - \vee x}{B} \right\rceil \tag{5.2}$$

where $\lceil \bullet \rceil$ indicates the ceiling function and $B = 50$ is the total number of bins chosen empirically in this framework. The occurrence of the training data is counted in each bin and yield a feature vector of $\mathfrak{N} = \{\mathfrak{n}_1, \mathfrak{n}_2, \cdots, \mathfrak{n}_B\}$. With this, the dominant region, $\bar{\mu}$ can

be located by,

$$\bar{\mu} = \frac{\sum_{i=1}^{B} \mathfrak{n}_i}{b} \tag{5.3}$$

where b denoted the total number of bin which satisfy $\mathfrak{n} > 0$.

The dominant region (mutually exclusive) is defined as the region where the distribution of training data is higher than $\bar{\mu}$. This region is marked with the membership value equals to 1. By referring to Figure 5.1, the parameters of a and b of $\mathbf{m}$ can be determined as the lower and upper bound of the area that possess membership value equals to 1. The intersection region (non-mutually exclusive) $a - \alpha$ and $b + \beta$ can be determined as the lower and upper bound of the area that possess membership value equals to 0 respectively. Algorithm 2 summarizes the learning process with a set of training image, $\mathbf{I}$ with K classes.

Algorithm 2 LEARNING FQTM

Require: A training dataset
Step 1: Grouping images Group every image to its respective class label, $\mathbf{I} \rightarrow \{\mathbf{I}^k\}_{k=1}^{K}$.

Step 2: Acquiring the feature values for all $\mathbf{I}^k$, perform preprocessing to obtain $\mathbf{x}_k$ where J features are acquired. Then compute $\mathbf{x}_{jk} = \{x_i^{jk}\}_{i=1}^{\mathbf{N}}$.
Step 3: Learning Model
for all j such that $1 \leq j \leq J$ **do**
  for all k such that $1 \leq k \leq K$ **do**
    Build a histogram of $\mathbf{x}_{jk}$
    Compute $\bar{\mu}$ with (5.3)
    Obtain $\mathbf{m}_{jk} = [a\ b\ \alpha\ \beta]_{jk}$ based on $\bar{\mu}$
  end for
end for
return FQTM

5.2.2 Inference

The goal here is to relax the mutually-exclusive assumption on the training data and classify a testing sample into their possibility classes and therefore, one testing sample can belong to multiple classes. This is unlike the conventional classification method or fuzzy

inference engine that the defuzzification step eventually derives a crisp decision and as well it is different from the previous FQRC in the way to perform inference.

Given a testing sample and its respective feature values x, the membership value μ of feature j belong to class k can be approximated by (5.4).

$$\mu_{jk}(x_j) = \begin{cases} 0, & x_j < a - \alpha \\ \alpha^{-1}\left(x_j - a + \alpha\right), & a - \alpha \leqslant x_j < a \\ 1, & a \leqslant x_j \leqslant b \\ \beta^{-1}\left(b + \beta - x_j\right), & b < x_j \leqslant b + \beta \\ 0, & x_j > b + \beta \end{cases} \tag{5.4}$$

where the parameter a, b, α, and β are retrieved from $\mathbf{m}_{jk}$ of the FQTM. The product, ρ_k of the membership values of all the features for each class, k is computed using

$$\rho_k = \prod_{j=1}^{J} \mu_{jk}(x_j) \tag{5.5}$$

Finally, the ρ_k is normalized and denote as r_k,

$$r_k = \frac{\rho_k}{\sum \rho} = \frac{\prod_{j=1}^{J} \mu_{jk}(x_j)}{Z} \tag{5.6}$$

where $Z = \sum \rho$ act as the normalizer. The intuition to use the product of membership values of all the features for each class, ρ_k is to calculate the confident value of them. This is the core to relate the inference mechanism closer to the principle of human reasoning and relax the non-mutually exclusive cases. If the feature of a testing data is dominantly belonged to a certain class, k (which means the membership value of that particular attribute, $\mu_{jk} = 1$), and similarly for the other features, at the end of the ρ_k computation, the testing sample that belongs to that particular class is a definite because the product

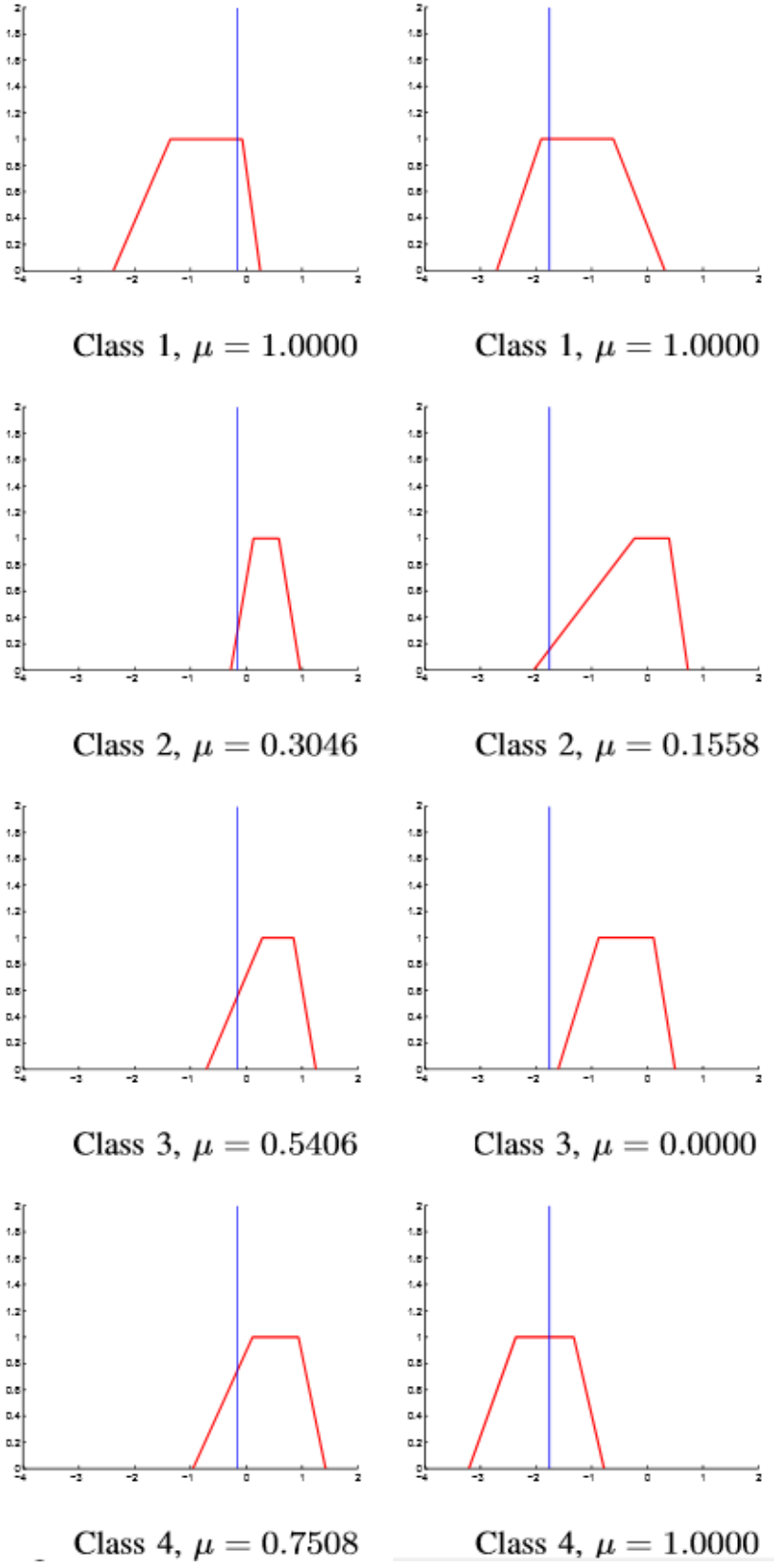
between values 1 is equal to 1. On the other hand, if the uncertainties for the feature (membership value of the feature $\mu_{jk} < 1$) are cumulated, the confident value decreases. In mathematical view, the products between values of less than 1 will eventually produce smaller value.

5.2.3 Example

An example on a scene image is presented here using the proposed DFQRC. Figure 5.2 shows an example walk-through of the inference process with a testing image, s (Figure 5.5(g)) and a learnt FQTM. Let's denote the feature values of the testing image as "Naturalness", $x_1 = -0.1545$ and "Openness", $x_2 = -1.7597$, respectively. For simplicity, only two features are used in this example but not limited to. This is because the proposed method can support multi-dimension classification task. By employing the learnt FQTM, ρ_k is computed as to (5.5) and r_k as (5.6).

In the inference process, $r_1 = 0.5561$, $r_2 = 0.0264$, $r_3 = 0.0000$ and $r_4 = 0.4175$ are obtained respectively. Each of these values represents that the scene s has the confident value r_1 belongs to "Insidecity", r_2 belongs to Coast, r_3 belongs to "Opencountry", and r_4 belongs to Forest where $\sum r = 1$. Based on human perspective, this result is reasonable as in the scene image, there are characteristics of "Incidecity" and "Forest". For examples, there are buildings, vehicles, as well as trees. Therefore, in the inference process, high degree of memberships of the features values from both classes is observed and thus infer a high value for r_1 and r_4. While, on the other hand, it possesses almost zero for r_2 and zero for r_3 because of low or zero value determined from the respective attribute values.

As discussed, most state-of-the-art approaches assumed that action class (Aggarwal & Ryoo, 2011; Ji & Liu, 2010; Moeslund & Granum, 2001; Poppe, 2010) and scene images (Bosch et al., 2006; Fei-Fei & Perona, 2005; Oliva & Torralba, 2001; Vogel & Schiele, 2007) are mutually-exclusive. Therefore, different strategies to build a sophis-



(a) Degree of Naturalness (b) Degree of Openness

Figure 5.2: The degree of membership, μ, of the attributes ('Natural' on the left, 'Open' on the right) for the respective classes.

ticated binary classifier (inference engine) were proposed in those state-of-the-art approaches. As opposed to these solutions, this work argued that there are cases either in action recognition or scene understanding could be non-mutually exclusive. Hence, DFQRC provides the learning scheme and inference engine that contributed in such a way that the training model captured the non-mutually exclusive characteristic of the data and the multi-label or ranking interpretation replaces the binary decision. Nevertheless, a comprehensive study of the intuition of using the 4-tuple membership function in the proposed DFQRC to solve the non-mutually exclusive problem is provided in Appendix 1.

5.3 Experiments and Discussions

Before applying to HMA, the performance of the propose DFQRC is evaluated with two public scene image datasets - the Outdoor Scene Recognition (OSR) dataset (Oliva & Torralba, 2001) and the Multi-Label Scene (MLS) dataset (Boutell et al., 2004; M.-L. Zhang & Zhou, 2007) with the reason that multi-label groundtruth is available partly for the OSR dataset from the online survey while the MLS dataset were manually annotated by three human observers. These are necessary to test the effectiveness of the proposed DFQRC in multi-label and ranking classification tasks.

The OSR dataset contains 2688 colour scene images, 256x256 pixels from a total of 8 outdoor scene classes ("Tallbuilding, T", "Insidecity, I", "Street, S", "Highway, H", "Coast, C", "Opencounty, O", "Mountain, M" and "Forest, F"). Figure 4.6 illustrates example of the OSR dataset and is publicly available[1]. In the meantime, MLS dataset contains a total of 2407 scene images with 15 (6 base + 9 multi-label) classes.

In the feature extraction stage for the OSR dataset, six different features which are also called attributes have been employed to represent the scene images. The six at-

[1]http://people.csail.mit.edu/torralba/code/spatialenvelope

tributes as introduced in Parikh & Grauman (2011) are the measurement on naturalness, openness, perspective, large objects, diagonal plane and close-depth of the scene images. Note that, an alternative representation such as other feature extraction methods can be employed as the front-end instead of the attributes. Since the focus in this study is the introduction of fuzzy qualitative approach to perform classification, any existing feature representation for images can be employed as the input to the system. In the meantime, for MLS dataset, the precomputed 294 dimensions feature vectors, R^{294} are employed as proposed by Boutell et al. (2004); M.-L. Zhang & Zhou (2007). Finally, in OSR dataset, 'leave-one-out' mechanism is used in the experiment. While for the MLS dataset, the distribution of training and testing data is according to the setting in (Boutell et al., 2004; M.-L. Zhang & Zhou, 2007).

Overall, the experiments are divided into five sections where each of them is tested on different perspectives of the proposed DFQRC. The bin number, B of the histogram is empirically set as 50.

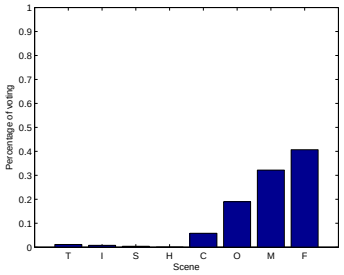
5.3.1 Effectiveness

This experiment is to show the correctness of the propose DFQRC in handling non-mutually exclusive data and the inconsistency of human decision making process. Let's denote Y_d as the set of result value for scenery image d from the survey and W_d be the set of predicted label from the DFQRC. The results are compared in the following aspects:

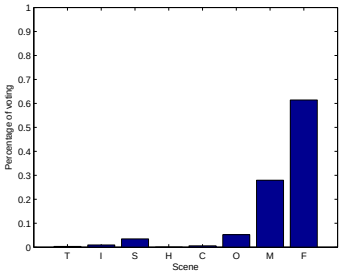
5.3.1 (a) Qualitative Observation

The corresponding results from the online survey and DFQRC are illustrated in Figure 5.3. Based on the figure, one can notice that the outcomes from both solutions are almost similar in terms of the ranking and the voting distributions. For instance, in Figure 5.3(d), majority of the participants have chosen "Tallbuilding" (84.2%) and follow by "Insid-

(a) Scene image 1

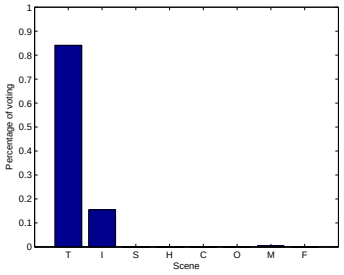


(b) Result of online survey, Y_d

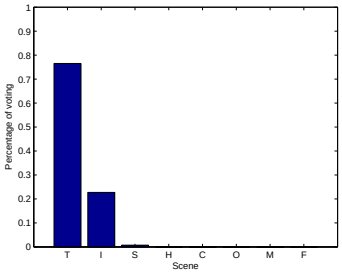


(c) Result of FQRC, W_d

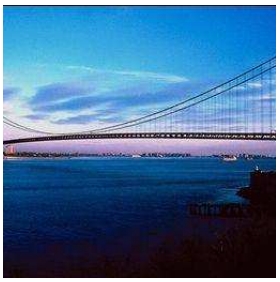
(d) Scene image 2



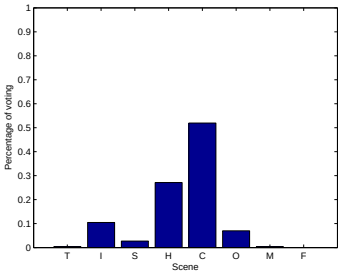
(e) Result of online survey, Y_d



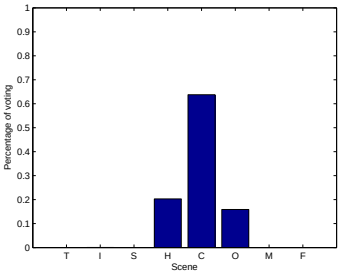
(f) Result of FQRC, W_d



(g) Scene image 3



(h) Result of online survey, Y_d



(i) Result of FQRC, W_d

Figure 5.3: Examples of the comparison between the results of online survey and FQRC ('Tallbuilding, T', 'Insidecity, I', 'Street, S', 'Highway, H', 'Coast, C', 'Opencounty, O', 'Mountain, M' and 'Forest, F'). These results had shown that our proposed approach is very close to the human reasoning in scene understanding.

ecity" (15.4%). This is close to the reading computed from FQRC where "Tallbuilding" is 76% and "Insidecity" hold 22.7%.

However, it is reasonable that to obtain exactly the same values to the online survey results is almost impossible due to the subjective human decision. Surprisingly, the ranking of the distribution in the online survey is very close to the result computed from DFQRC. For example, in Figure 5.3(d), by considering only the 'hit' labels for both results ("Tallbuilding, C" and "Insidecity, I"), the order of the distribution for DFQRC computed result is T more than I which is similar to the survey results although the values are

not exactly the same.

Based on this observation, a preliminary conclusion can be drawn that the proposed approach is able to emulate human reasoning in classifying scene images. To further validate this, quantitative evaluation is conducted in the following context.

5.3.1 (b) Quantitative Evaluation

In order to show that DFQRC is able to model how human makes decisions, a quantitative evaluation is performed by using several evaluation criteria which are α-Evaluation, Cosine similarity measure, and error rate calculation.

α-Evaluation. Evaluation of multi-label classification results is more complicated compared to that of binary classification because a result can be fully correct, partly correct, or fully incorrect. By using the example given by Boutell et al. (2004), let's assume a set of classes c_1, c_2, c_3 and c_4. By taking an example of testing sample with its ground truth that belongs to classes c_1 and c_2, the different output results can be interpreted as below:

- c_1, c_2 (fully correct),
- c_1 (partly correct),
- c_1, c_3 (partly correct),
- c_1, c_3, c_4 (partly correct),
- c_3, c_4, (fully incorrect)

Herein, to measure the degree of correctness of those possible results with their proposed α-Evaluation. The score is predicted by the following formula:

$$score(W_d^b) = \left(1 - \frac{|\beta M_d + \gamma Q_d|}{|Y_d^b \cup W_d^b|}\right)^{\alpha} \tag{5.7}$$

where Y_d^b is the set of ground truth labels for the image sample $d \in D$ in binary form ($Y_d > 0$) and W_d^b is the set of prediction labels from the DFQRC in binary form ($W_d > 0$). Also, $M_d = Y_d^b - W_d^b$ (missed labels) and $Q_d = W_d^b - Y_d^b$ (false positive labels). In here, α, β and γ are constraint parameters as explained in Boutell et al. (2004). In the evaluation, $\alpha = 0.5, \beta = 1$ and $\gamma = 1$ are selected and the accuracy rate of D is computed with,

$$accuracy_D = \frac{1}{|D|} \sum_{d \in D} score(W_d^b) \tag{5.8}$$

where higher accuracy reflects better reliability of the DFQRC because the 'hit' label (i.e: label with distribution more than zero) is almost similar to the survey results.

Cosine similarity measure. Cosine similarity measure is use to investigate the similarity of the histogram obtained from the survey and the DFQRC, respectively, by matching the pattern of the distributions. First, the cosine distance (5.9) of the histogram distributions of each scene image is computed.

$$distance(W_d) = cos\Theta = \frac{Y_d \cdot W_d}{\|Y_d\| \, \|W_d\|} \tag{5.9}$$

Then, the average value of the similarity value for D is computed as (5.10) to evaluate the overall performance.

$$similarity_D = \frac{1}{|D|} \sum_{d \in D} (1 - distance(W_d)) \tag{5.10}$$

where larger value of $similarity_D$ indicates higher similarity.

Error rate calculation. In this evaluation criteria, how much the computed result from the DFQRC is deviated from the survey results is investigated. To begin with, the error vector by subtracting both of the histogram distributions is obtained,

$$err(W_d) = |W_d - Y_d| \tag{5.11}$$

Then, the mean and standard deviation of the error vector is computed to observe the range of error as shown in Figure 5.4.

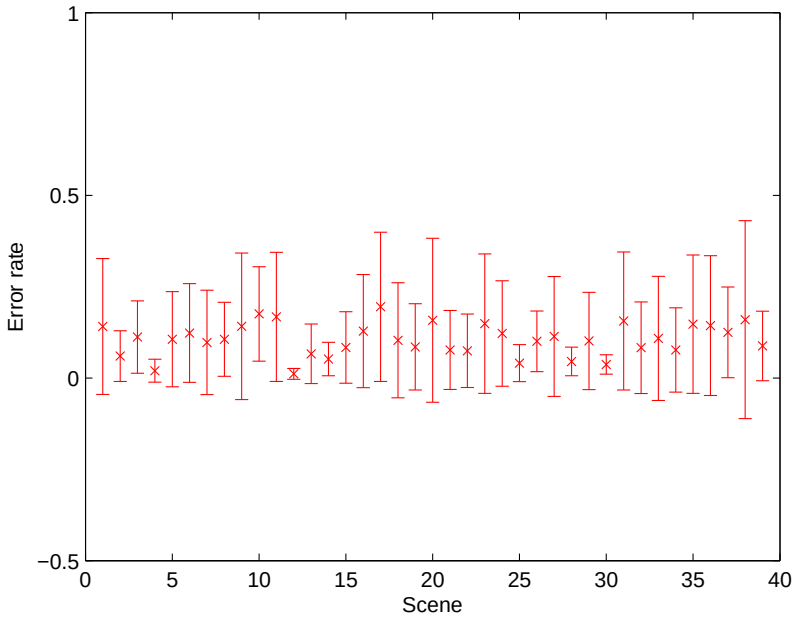


Figure 5.4: Error bar of DFQRC results compared to the online survey results for each scene image.

For the overall judgment in error rate, the average standard deviation of the error values obtained from the scene images is computed. Smaller value indicates less deviation of the DFQRC results from the online survey results.

All the three evaluation criteria are tested by comparing with the online survey results (with and without α-cut) using the proposed DFQRC. The results are shown in Table 5.1.

Table 5.1: Quantitative Evaluation of DFQRC compared to online survey results.

Scene	α-evaluation (accuracy)	similarity	error (Average)
Without α-cut	0.75	0.72	0.13
With α-cut (1%)	0.79	0.72	0.13

From the results in Table 5.1, one could observe acceptable output from these three evaluation criteria. The accuracy is above 70%, which indicates that the computational results using the DFQRC is close to human reasoning in decision making where the 'hit' label is highly matched with the answer from the survey. The high similarity from the results shows that the proposed approach is able to provide an outcome similar to a human decision in terms of voting distribution and ranking.

Based on the qualitative and quantitative results, it is clarified that the scene images are non-mutually exclusive and the state-of-the-art approaches that used binary classifier to deduce an unknown image to a specific class is not practical. Besides that, DFQRC has proven its effectiveness as a remedy for this situation based on the evaluation by utilizing the online survey results.

5.3.2 Feasibility

In this experiment, the feasibility of the proposed DFQRC is clarified in terms of the capability in performing multi-label, multi-class, multi-dimension and ranking classification tasks. The explanation for each of the capabilities is as below:

- **Multi-label** - the classification outputs are associated with a set of labels
- **Multi-class** - the classifier that supports more than two classes, $K > 2$ in a single classification task
- **Multi-dimension** - the classifier that supports more than two features, $J > 2$ in a single classification task

- **Ranking** - higher interpretation of the classification results by reordering the inference outcome.

Table 5.2: Comparison of the DFQRC with the other classifiers in terms of scene understanding.

Classifier	Multi-label	Multi-class	Multi-dimension	Ranking
KNN	-	✓	-	-
SVM	-	-	✓	-
Platt et al. (2000)	-	✓	✓	-
Boutell et al. (2004)	✓	✓	✓	-
FQRC	✓	✓	✓	✓

Table 5.2 shows how the FQRC distinguishes itself from the other classifiers and each of the capabilities have been clarified with the succeeding experiments in the following sections.

5.3.2 (a) DFQRC with 2 attributes and 4 scene classes (Multi-label & Multi-class)

From the comparison results show in Table 5.3, it can be observed that one drawback of the FQRC is it provides similar results on certain images, which is absurd as all the corresponding images are so different from each other and imply that each of the images has its own value of attributes, which should be different from other images. DFQRC, in contrast, is able to model this behavior and provides an output that is closer to human thinking and decision. Apart from that, the confident values inferred from DFQRC are more reasonable compared to FQRC. For example, in Figure 5.5(e), from human perspective of view, one will consider that the confident level of this image belonged to “Insidecity” is higher than the “Forest”. Such improvement is endowed by the proposed 4-tuple fuzzy membership learning algorithm.

(a) (b) (c) (d) (e) (f) (g)

Figure 5.5: Example of images of Insidecity.

Table 5.3: Inference output with two attributes and four classes for the scene images in Figure 5.5.

Scene (Figure)	DFQRC				FQRC			
	Insidecity	Coast	Opencountry	Forest	Insidecity	Coast	Opencountry	Forest
5.5(a)	0.9280	0	0	0.0720	1	0	0	0
5.5(b)	1	0	0	0	1	0	0	0
5.5(c)	0.5068	0.1587	0.3344	0	0.7273	0.2727	0	0
5.5(d)	0.6845	0	0.3155	0	0.7273	0.2727	0	0
5.5(e)	0.5296	0.0483	0	0.4221	0.1250	0	0.1250	0.7500
5.5(f)	0.5872	0.0146	0.007	0.3911	0.8235	0	0	0.1765
5.5(g)	0.5561	0.0264	0	0.4175	0.8235	0	0	0.1765

5.3.2 (b) DFQRC with 6 attributes and 4 scene classes (Multi-dimension)

In this testing, the proposed DFQRC shows its strength in performing multi-dimensional classification compare to FQRC where 6 attributes instead of 2 are employed to perform the classification tasks. The 6 attributes are the score values of ‘Naturalness’, ‘Openness’, ‘Perspective’, ‘Size-Large’, ‘Diagonal-Plane’, and ‘Depth-Close’, respectively. Using the similar testing images as in Figure 5.5, the classification results from the DFQRC are shown in Table 5.4.

Table 5.4: Inference output with six attributes and four classes for the scene images in Figure 5.5.

Scene (Figure)	Insidecity	Coast	Opencountry	Forest
5.5(a)	1	0	0	0
5.5(b)	1	0	0	0
5.5(c)	0.6722	0.1001	0.2277	0
5.5(d)	0.9179	0	0.0821	0
5.5(e)	0.5188	0	0	0.4812
5.5(f)	0.8411	0	0.0013	0.1575
5.5(g)	0.5936	0	0	0.4064

By comparing the result between Table 5.3 and 5.4, it can be observed that the result using six attributes are more reasonable than two attributes, especially in Figure 5.5(a), 5.5(e), 5.5(f), and 5.5(g), respectively. In the case of Figure 5.5(e), with using six attributes instead of two, the result improved in the way that the noise was eliminated which is the "Coast" class that should never been an option for this particular image. However, the values of confident of Figure 5.5(e) in "Insidecity" and "Forest" have change significantly. In spite of, the confident level of "Insidecity" is still more than "Forest" which matched to the subjective judgment.

Slight changes in these results were incurred as a resultant from the additional of the number of attributes into the classification framework. In fact, more attributes tend to increase the uniqueness of one class from another and this has indirectly increased the discriminative strength of the classifier. However, it is almost impossible to find the optimum attributes (or features) that are best to distinguish one class from another classes especially in non-mutually exclusive cases. Furthermore, using excessive attributes in the algorithm will increase the computational cost. Therefore, the proposed DFQRC considers a more generative way that provides a good tradeoff between the multi-dimensional classification capability and the performance of the classification task.

5.3.2 (c) DFQRC in ranking (Ranking ability)

The goal of this experiment is to show the effectiveness of the proposed DFQRC in higher interpretation such as the ranking interpretation by classifying the possibility of an unknown image into the eight learned scene classes with the correct ordering.

Table 5.5 shows the sub-sample results using randomly selected scene images from the "Insidecity" class. The visual appearances of these images are illustrated in Figure 5.5. Herein, it is noticeable that the DFQRC is able to correctly classify each image which has the possibility (confident value, r_k) in "Insidecity" class. This is true as the bench-

Table 5.5: Inference output with 6 attributes and 8 classes of Figure 5.5.

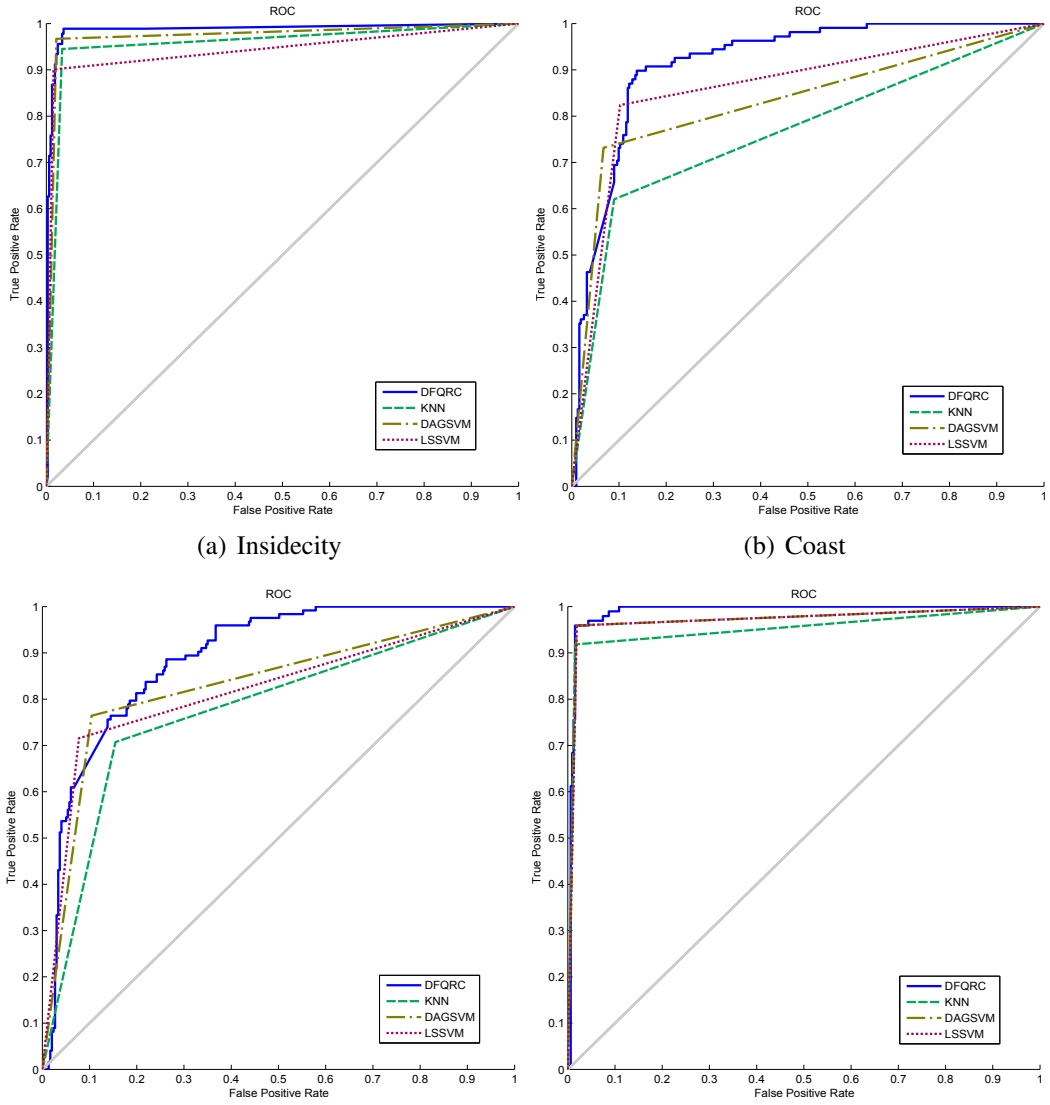
Scene (Figure)	Tallbuilding	Insidecity	Street	Highway	Coast	Opencountry	Mountain	Forest
5.5(a)	0.4562	0.4562	0.0876	0	0	0	0	0
5.5(b)	0.7644	0.2356	0	0	0	0	0	0
5.5(c)	0	0.3339	0.0308	0.4725	0.0497	0.1131	0	0
5.5(d)	0	0.5880	0.0499	0.3094	0	0.0526	0	0
5.5(e)	0.0726	0.2631	0.4202	0	0	0	0	0.2440
5.5(f)	0.1412	0.3456	0.4361	0	0	0.0005	0.0119	0.0647
5.5(g)	0.0811	0.2826	0.4183	0	0	0	0.0245	0.1935

marking for these sub-sample images is selected from the "Insidecity" class. Nonetheless, this approach also discovered that each of these images can have possibility belongs to other classes. For instance, it is discovered that Figure 5.5(a) has the possibility as "Tallbuilding" and "Street" class.

5.3.3 Comparison to State-of-the-art Binary Classifiers in Single Label Classification Task

One of the strengths of the proposed DFQRC is, it provides the feasibility to perform single-label classification task like the other binary classifiers with comparable result. To verify this, DFQRC is tested against the state-of-the-art binary classifiers such as KNN, Directed Acyclic Graph SVM (DAGSVM) (Platt et al., 2000), and Fuzzy least squares SVM (LSSVM) (Tsujinishi & Abe, 2003). In the DFQRC, max aggregation method ($\max(r)$) is employed to obtain the class with maximum confident value as the binary classification results.

For simplicity, the classification task is conducted with two attributes and four classes for all classifiers. In the configuration of each classifier in the comparison, conventional KNN is used with the empirical chosen parameter $K = 5$. As for DAGSVM (Platt et al., 2000) and LSSVM (Tsujinishi & Abe, 2003), DAGSVM runs with RBF as kernel and margin parameter, $C = 100$ using SMO training while LSSVM is implemented based on the linear SVM with $C = 2000$ and incorporates with the least square solution.



(a) Insidecity (b) Coast

(c) Opencountry (d) Forest

Figure 5.6: ROC comparison between DFQRC and the other binary classifiers.

The $F-score$ (Figure 5.7) is calculated to show the accuracy of the classification task by comparing our DFQRC and three other classifiers. In information retrieval literatures, the $F-score$ is often used for evaluating this quantity:

$$F-score = \frac{2\Psi\eta}{\rho+\eta}. \tag{5.12}$$

The recall, η and the precision, Ψ measure the configuration errors between the ground truth and the classification result. For a good inference quality, both the recall and precision should have high values. The ROC graphs show in Figure 5.6 is to evaluate the

sensitivity of the classifiers while Figure 5.7 illustrates the F-score for each classification task. From both figures, it can be observed that the proposed method is comparable with the KNN, DAGSVM, and LSSVM. In overall, DFQRC outperforms other binary classifiers but is slightly inefficient as compared to DAGSVM.

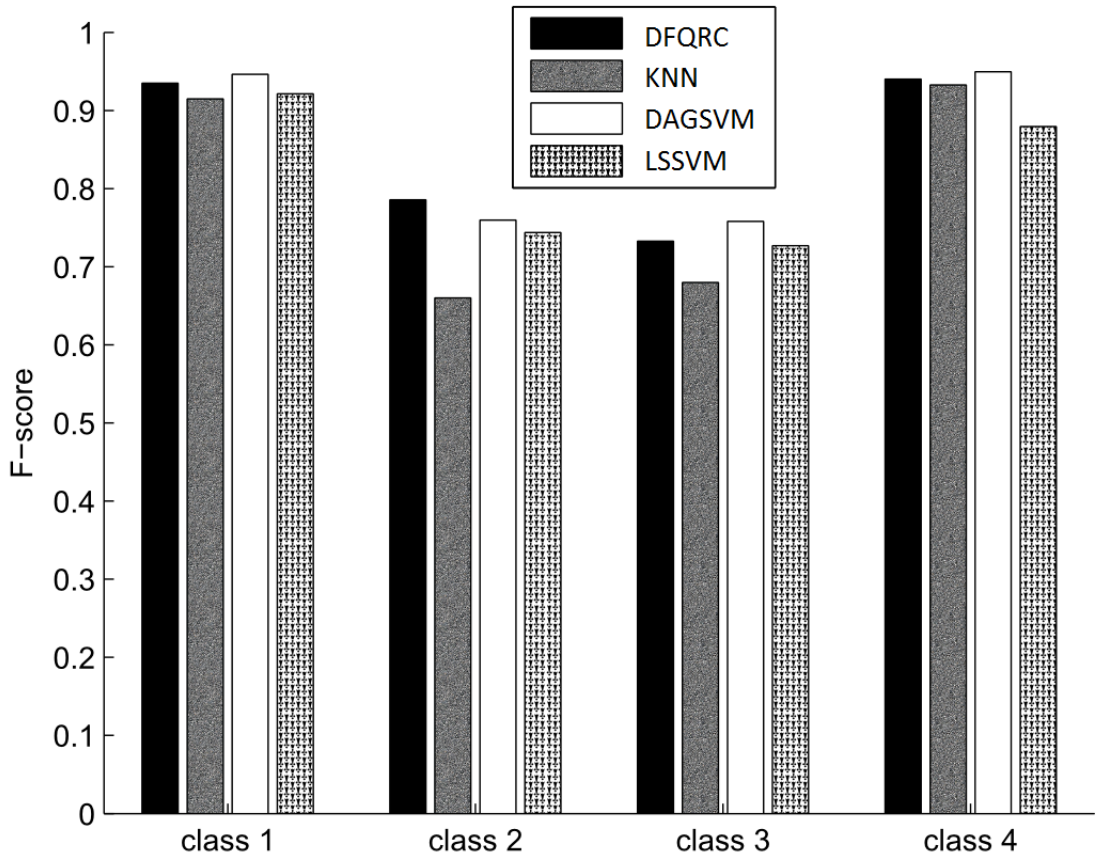


Figure 5.7: Comparison of F-score between the classifiers. Class 1 (Insidecity), Class 2 (Coast), Class 3 (Opencountry), and Class 4 (Forest).

One of the main reasons is DAGSVM used an efficient data structure to express the decision node in the graph, and an improved decision algorithm is used to find the class of each test sample and thus makes the decision more accurate compared to other binary classifiers. In short, DAGSVM is a discriminative classifier that was implemented and trained to distinguish distinctly amongst the data where there is no crossover tolerance in the data distribution. This is in contrary to the DFQRC as a generative classifier to relief the ignorance of non-mutually exclusive data. This is the reason why DAGSVM should be better in compared to DFQRC as a binary classifier. However, here, in this context,

the objective is to show that one of the strengths of DFQRC is the capability to perform single-label classification task while playing the role of ranking classifier, which yields comparable results with the other state-of-the-art binary classifiers.

5.3.4 Comparison to State-of-the-art Multi-label Classifiers

In order to show the effectiveness and efficiency of the proposed method in multi-label classification task, in this experiment, DFQRC is compared with the state-of-the-art multi-label scene classification approaches (Boutell et al., 2004; M.-L. Zhang & Zhou, 2007). This comparison is performed with MLS dataset. The comparison is done on two aspects: computational complexity and accuracy.

5.3.4 (a) Computational complexity

First, the complexity of the DFQRC compared to the approaches proposed by Boutell et al. (2004) and M.-L. Zhang & Zhou (2007) by using Big O notation is conducted with the results presented in Table 5.6. In this context, N denotes the number of classes, M is the number of features, and T is the number of data. The complexity of M.-L. Zhang & Zhou (2007) approach in the training phase consists of three parts; prior, conditional probability, and the main function of training, while Boutell et al. (2004) required to train a classifier for every base class. These greatly increase the computational cost compare to the DFQRC.

Table 5.6: Complexity of DFQRC compared to the state-of-the-arts.

Method	Part	
	Training Phase	Testing Phase
M.-L. Zhang & Zhou (2007)	$O(N)+O(T)+(O(3TN)+O(N))$	$O(2N)$
Boutell et al. (2004)	$O(NT^3)$	$O(N)$
DFQRC	$O(NM)$	$O(NM)$

In order to verify the complexity of these methods, the computational time comparison is done with the results are showed in Table 5.7. From the result, it is noticeable that, DFQRC used the shortest time to train the model which is almost six times faster than M.-L. Zhang & Zhou (2007) and 227 times faster than Boutell et al. (2004). However, the inference takes a longer time compared to both methods. This is because DFQRC retrieves the fuzzy membership values by considering all the 4-tuple membership functions that corresponds to all features for every class. This also means that with a reduction in terms of the number of features, it is possible to obtain faster computational speed. The computational time for the testing is done by using all testing data, so it is acceptable as one testing data can be processed with an average of 3 milliseconds. Nonetheless, M.-L. Zhang & Zhou (2007) suffered from finding the optimal number of nearest neighbor involved in the classification step. This had directly affected the performance of the classification.

Table 5.7: Computational time of DFQRC compared to Boutell et al. (2004) and M.-L. Zhang & Zhou (2007) on MLS dataset.

Method	Computational time (s)		
	Training	Testing	Overall
M.-L. Zhang & Zhou (2007)	0.9363	0.5662	**1.5025**
Boutell et al. (2004)	37.8859	**0.3725**	38.2584
DFQRC	**0.1666**	3.9479	4.1145

5.3.4 (b) Accuracy

For fair comparison, instead of employing all the scene data from the MLS scene dataset, only the multi-label class scene data is selected for this testing. It means that, those testing data that are categorized as base class in Boutell et al. (2004) according to the ground truth were eliminated and only the test data in multi-label class were used. This explains why

the results are different from Boutell et al. (2004). Again, it should be pointed out that the intention of this work is focused on the multi-label scene classification.

Table 5.8: α-Evaluation of DFQRC compared to M.-L. Zhang & Zhou (2007) and Boutell et al. (2004).

Method	α-evaluation			
	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
M.-L. Zhang & Zhou (2007)	1	0.54	0.39	0.20
Boutell et al. (2004)	1	0.69	0.49	0.27
DFQRC	**1**	**0.69**	**0.54**	**0.37**

Based on Boutell et al. (2004), α is the forgiveness rate which determines how much to forgive the errors made in predicting labels. Small value of α is more aggressive (tend to forgive error) while a high value is conservative (penalizing error strictly). In relation to the multi-label classification, $\alpha = \infty$ with a score = 1 occurs only when the prediction is fully correct (all hit and no missed) or 0 otherwise. On the other hand, when $\alpha = 0$, the score will be always $= 1$ unless the answer is fully incorrect (all missed). From Table 5.8, one can observe that the DFQRC outperforms the two other methods with better accuracy in the α-evaluation.

In summary, through a series of comprehensive experiments on different perspectives, the proposed DFQRC has performed well in all the evaluations. For examples, the result obtained from DFQRC is comparable to the human decision as compared to the online survey in scene understanding that presented in Chapter 4. In conjunction, DFQRC has proved its capability in performing single label, multi-label, multi-class, multi-dimension, and ranking classification tasks. In addition, it outperformed the state-of-the-arts single label and multi-label approaches. After all the validation that have been done towards the proposed DFQRC, it is convincing that DFQRC is capable of modelling the ambiguous cases. It is then further applied in action recognition to test its effectiveness.

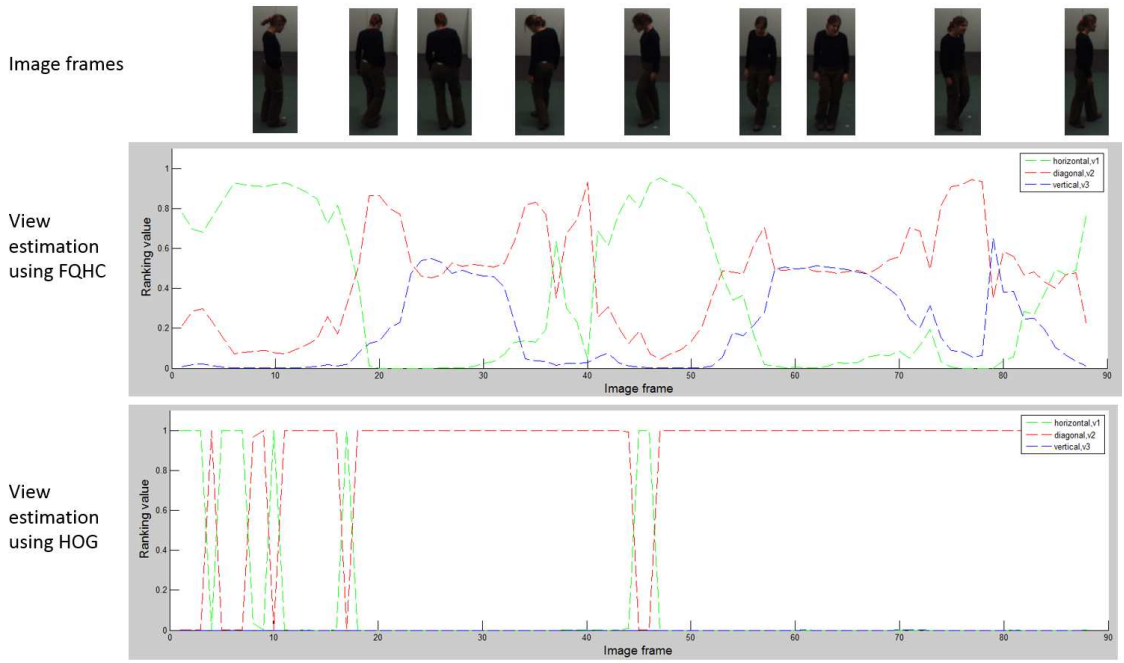
5.4 Application in Human Motion Analysis

This testing is to evaluate the performance of DFQRC in addressing the ambiguity of view estimation and action recognition in terms of multi-label and ranking classification.

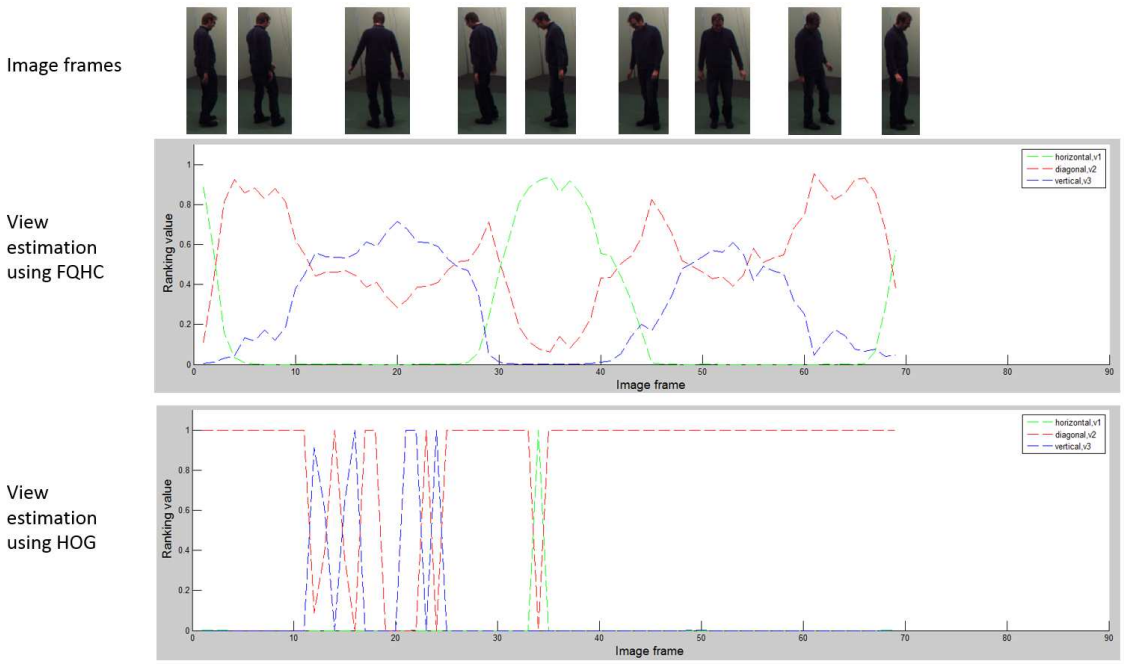
5.4.1 View Estimation

In view estimation, human viewpoint can be ambiguous where they tend to be confused between other viewpoints (i.e: v_1, v_2, and v_3 as defined in Chapter 3), especially during the transition from one viewpoint to another. In this testing, DFQRC is use to model the viewpoints of the "turn around" action in IXMAS dataset. Some qualitative results are illustrated in Figure 5.8 with the corresponding turning graphs. The first row shows the original image of the person at the corresponding frame, while the second row shows the view estimation results using the proposed FQHC and compared to the usage of HOG at third row.

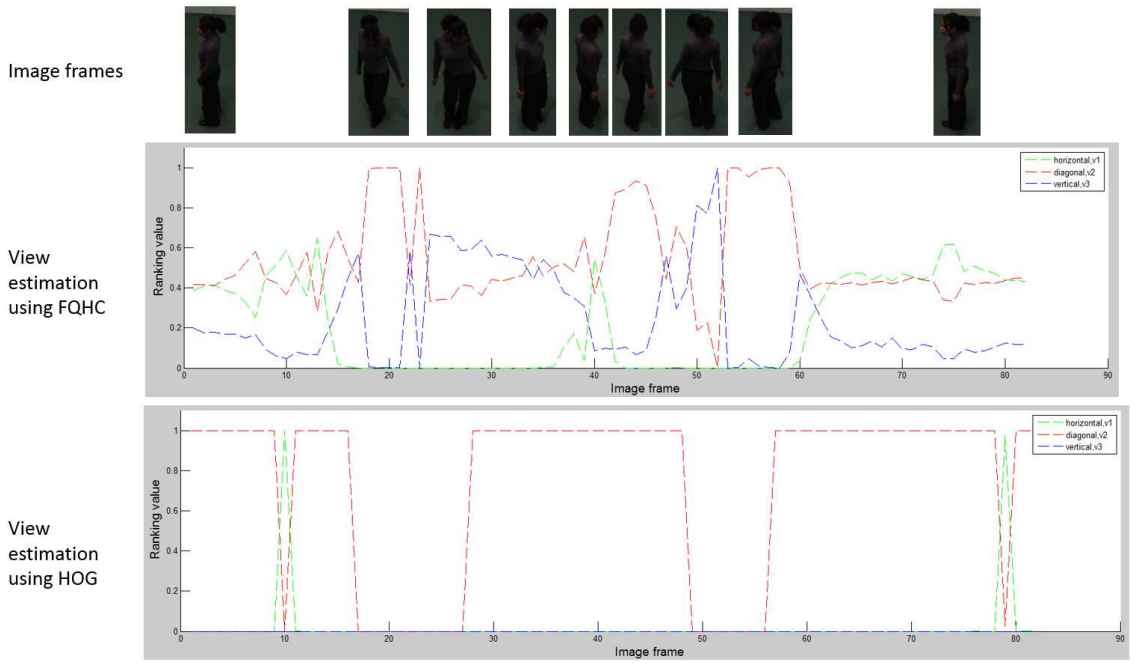
From Figure 5.8, instead of crisp classification result, the ranking corresponds to the three viewpoints (v_1, v_2, and v_3) are given. Different camera positions in terms of φ angle are tested. Although there are noises (inconsistency in turning pattern), the proposed method is still able to model the turning activity especially the transition from one viewpoint to another. These transitions proved that uncertainty exist in view estimation and that is why the previous crisp classification could not perform well. By using DFQRC, the person's viewpoint can be learned and infer as ranking outputs. By comparing FQHC and HOG features in modelling the turning activity, again, FQHC achieve more reasonable results where the viewpoints are correctly inferred using FQHC while HOG missed out most of the viewpoints as refer to Figure 5.8.



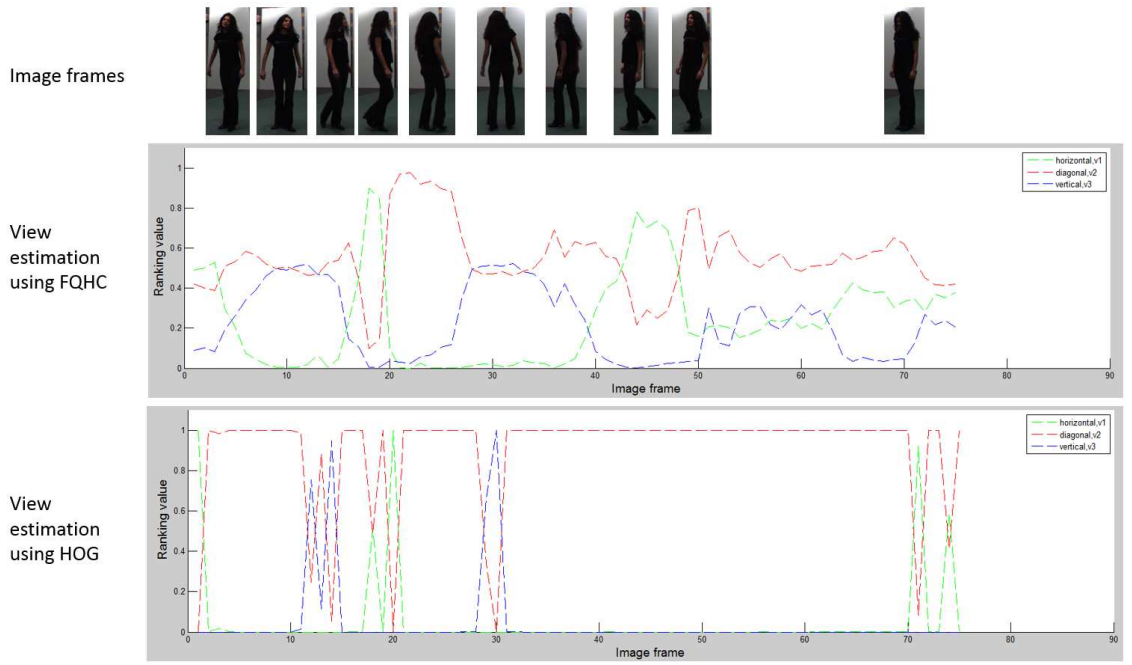
(a) Camera position 1



(b) Camera position 2



(c) Camera position 3



(d) Camera position 4

Figure 5.8: Comparison between FQHC (second row) and HOG (third row) in viewpoint estimation using DFQRC.

5.4.2 Action Recognition

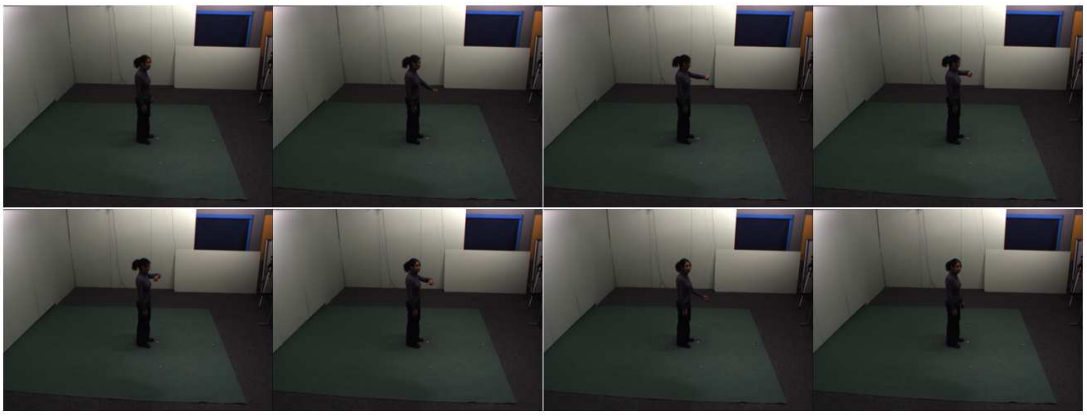
The aim of the thesis is to build an action recognition system that is robust over the aforementioned uncertainties. In the process pipeline, the uncertainties on different sizes and viewpoints (φ angle) were addressed with the view specific action recognition framework that utilized the FQ-PHM. However, ambiguity is still exist in the final action recognition result which means that some of the action are non-mutually exclusive such as walking, jogging, and running which cannot be deduce with crisp or binary classification method. Instead, DFQRC is used to model the non-mutually case and generate ranking and multi-label output. The proposed DFQRC do not ignore any of the possibility of the classes that a testing subject could belongs to. Putting them together, the results of action recognition in ranking interpretation and multi-label classification output is conducted as follow.

5.4.2 (a) Ranking interpretation

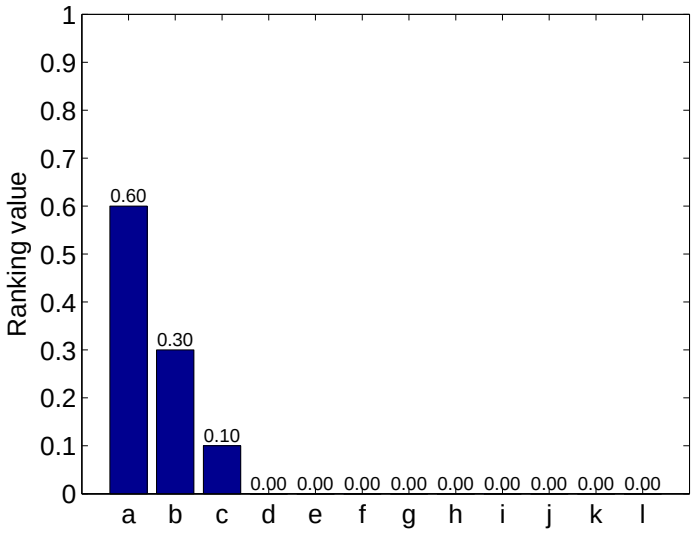
In the ranking interpretation, both the uncertainties of viewpoint and action in classification process are taken into account. The former is the ambiguity of the subject viewpoints in a frame while the latter is the ambiguity of the subject action. In this testing, the ranking of the viewpoints as previous testing is denoted as $\mathbf{W}$; while the action ranking is obtained by utilizing the output from the spatio-temporal bag of features and denoted as $\mathbf{A}$. By taking both information, the correlation $\mathbb{R}$ is defined as:

$$\mathbb{R} = \frac{\mathbf{W}_{v_1} \cdot \mathbf{A}_{v_1} + \mathbf{W}_{v_2} \cdot \mathbf{A}_{v_2} + \mathbf{W}_{v_3} \cdot \mathbf{A}_{v_3}}{\mathbf{Z}}. \tag{5.13}$$

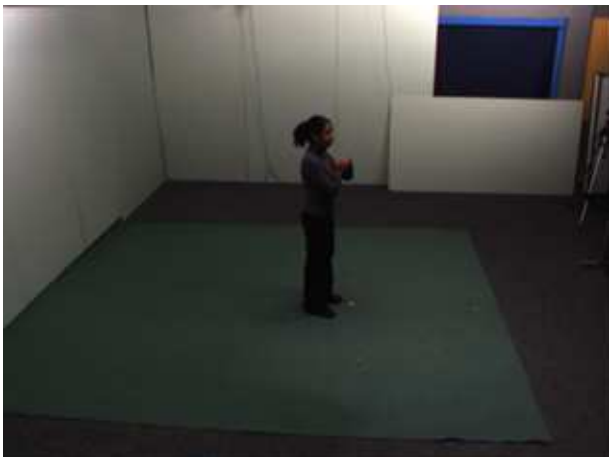
where $\mathbf{Z} = \sum(\mathbf{W}_{v_1} \cdot \mathbf{A}_{v_1} + \mathbf{W}_{v_2} \cdot \mathbf{A}_{v_2} + \mathbf{W}_{v_3} \cdot \mathbf{A}_{v_3})$. With this, the final result to infer the action of the subject is obtained in a ranking manner. Figure 5.9 to 5.12 show the examples of ranking output from IXMAS dataset. The action are denoted as “a - check watch, b - cross arms, c - scratch head, d - sit down, e - get up, f - turn around, g - walk, h - wave,



(a) Ground truth = Check watch action



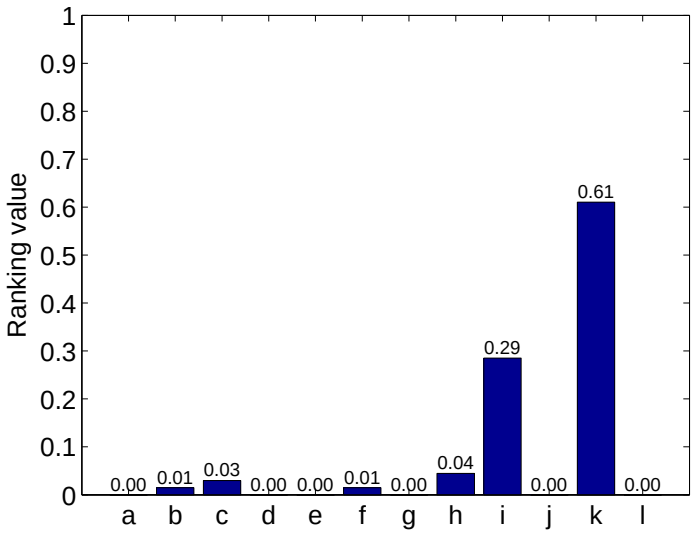
(b) Multi-label (viewpoints)



(c) Major confusion: Cross arm action

Figure 5.9: Ranking result for “Check watch” action at Cam 1.

(a) Ground truth = Punch action

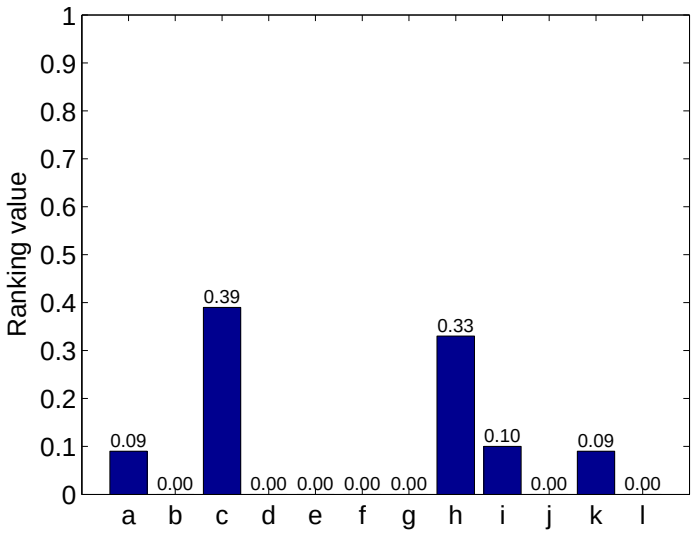


(b) Corresponding ranking

(c) Major confusion: Point action

Figure 5.10: Ranking result for “Punch” action at Cam 2.

(a) Ground truth = Wave action

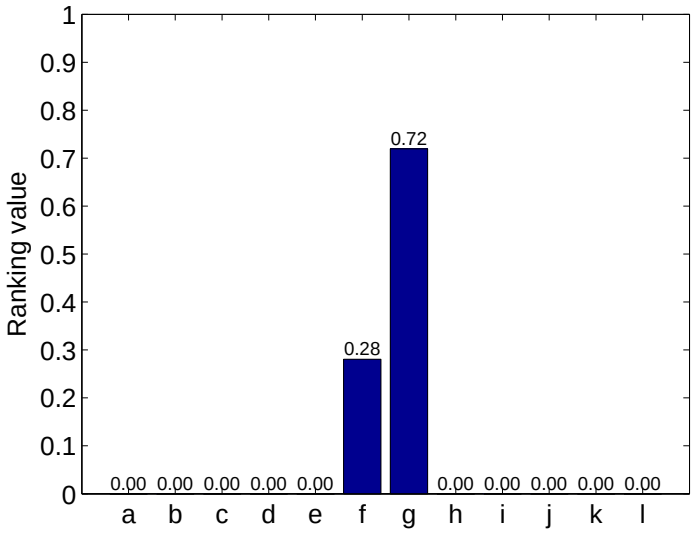


(b) Corresponding ranking

(c) Major confusion: Scratch head action

Figure 5.11: Ranking result for “Wave” action at Cam 3.

(a) Ground truth = Walking



(b) Corresponding ranking

(c) Major confusion: Turning activity

Figure 5.12: Ranking result for “Walking” action at Cam 4.

i - punch, j - kick, k - point, l - pick up".

In each example, the ranking result represents the action of a subjects with different values of confident in all 12 actions in IXMAS dataset (a to l). First, it is noticeable from the examples where the action with highest confident value is matched with the ground truth. Secondly, the confusion between the actions can be noticed from the ranking result. For instance, the "check watch" action is confuse with "cross arm" action in Figure 5.9. This is reasonable, as from human inspection, both actions are very similar, this is ambiguous to human as well if decision need to be made based on the image frame in Figure 5.9(c). This implies that the feature extracted for both actions could be similar and thus confusion in classification could happen.

Apart from these, note that Figure 5.9 to 5.12 show the examples taken from different camera positions (φ and Θ angles). Nonetheless, the system is resulting in reasonable ranking outputs as these uncertainties have been cope with the proposed view specific action recognition framework.

5.4.2 (b) Multi-label action recognition

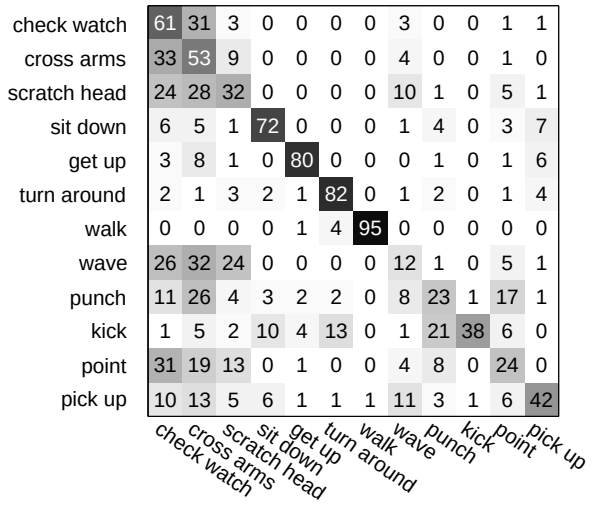
The usage of multi-label classification is to compromise the uncertainty in action recognition pipeline, for instance the viewpoint and action ambiguity. In multi-label classification, the result is inferred base on the "hit or miss" concept. In specific, to obtain multi-label classification result, the ranking output computed by DFQRC is declared as "hit" when the ranking value for the possible class that matched with the ground truth is more than zero, and "miss" when it is equal to zero. In mathematics expression, it can be obtained with the summation of the dot product between the ranking result Y_d and the

ground truth GT.

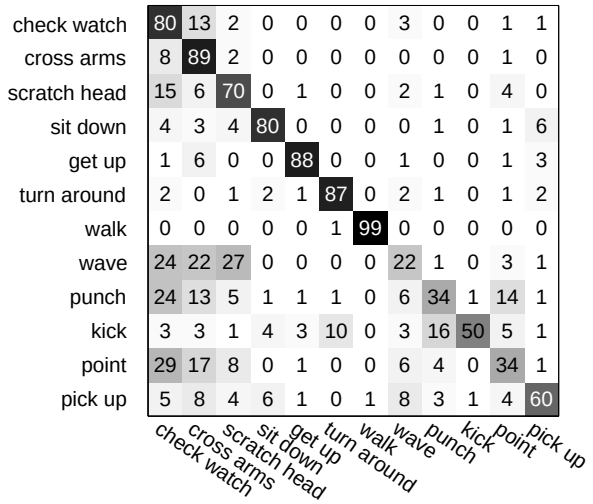
$$\text{result} = \begin{cases} \text{hit}, & \sum Y_d \cdot GT > 0 \\ \text{miss}, & \sum Y_d \cdot GT = 0 \end{cases} \tag{5.14}$$

As a simple examples, let's assume that $Y_d = [0.2\ 0\ 0\ 0.8\ 0]$ and the $GT = [0\ 1\ 0\ 0\ 0]$ with respect to five classes, the result will be "miss" as the $\sum Y_d \cdot GT = 0$. On the other hand if $Y_d = [0.2\ 0.8\ 0\ 0\ 0]$, the result will be "hit" as $\sum Y_d \cdot GT = 0.8$ which is more than 0. The overall action recognition result by using multi-label on viewpoints, and multi-label on viewpoints with action are showed in Figure 5.13 with comparison to the result that uses single label classification result. One can observe that the multi-label classification result with only viewpoints, and the multi-label classification with viewpoints + action are gradually improving the recognition accuracy compared to the single label classification method. This proved that the uncertainties in each criteria (viewpoint and action) is affecting the overall action recognition performance due to the confusion or ambiguity.

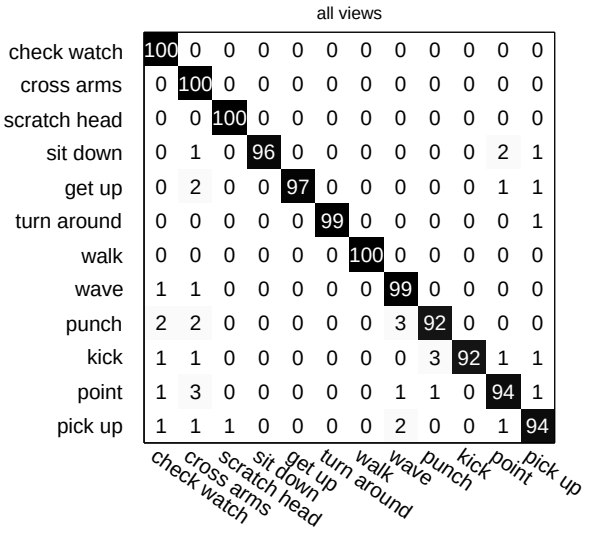
These results conveyed an important message where the human viewpoint ans action could be non-mutually exclusive due to the ambiguity that abounded. Instead, multi-label and ranking interpretation are more appropriate to infer such cases instead of binary classification result. The reasoning behind is it does not ignore any of the possible class that are feasible to label the testing subject. This is contradict with the other state-of-the-art methods (Ahmad & Lee, 2006; Anderson, Luke, et al., 2009b; Ashraf et al., 2013; Lewandowski, Makris, & Nebel, 2010; Weinland et al., 2007, 2006; Yilma & Shah, 2005) where a crisp or binary answer is mandatory.



(a) Single label classification



(b) Multi-label in viewpoints



(c) Multi-label in viewpoints and actions

Figure 5.13: Comparison between action recognition rate using binary classification and multi-label classifications.

5.5 Summary

In this chapter, DFQRC is proposed as the extension of FQRC that is able to learn the 4-tuple fuzzy number adaptively from the training data. From the experiments, surprisingly DFQRC is capable of producing the ranking results that is similar to human decision. Apart from that, DFQRC is equipped with the ability to perform multi-label, multi-class, multi-dimension and ranking classification tasks. In addition, it outperformed the state-of-the-art methods in multi-label scene classification task. Last but not least, it showed the effectiveness to apply in the HMA, where in specific, the viewpoints estimation and action recognition task with reasonable outcomes.

CHAPTER 6: CONCLUSIONS

6.1 Summary

This thesis was set out to explore the uncertainties abounded in the HMA system that hinder the effort to build a practical HMA system which is feasible to deploy in real world environment. So far, most studies have been focused on algorithms that are limited with some constraints and assumptions which are impractical solutions as they might be over-fitted the constrained pre-collected dataset. Specifically, the thesis is driven towards solving three uncertainties in HMA which are the human size variation, viewpoint variation, and classification ambiguity in conjunction with the two main contributions which are the view specific action recognition framework and the fuzzy qualitative rank classifier.

With reflect to the objective (section 1.3), an extensive literature review has been done to study the feasibility of fuzzy approaches in addressing the uncertainties in HMA system. From the study, the fuzzy qualitative reasoning had been identified with better capability to model the uncertainties compared to the others and thus is chosen to implement the proposed solutions in this thesis.

A part from this, view specific action recognition had also proved its capability in achieving promising performance in view independent HMA. In the framework, the view estimation module that comprise of the proposed FQ-PHM and the FQHC have been evaluated and outperform the state-of-the-art human contour descriptor HOG in the robustness test and HMA performance. It serve the purposes that the human model is invariant to size, body anatomy, and camera positions which is vital to the proposed framework. In addition, the proposed view specific action recognition framework is a view invariant human action recognition framework that uses only single camera which is an advantage

over the state-of-the-art works that required multiple cameras. Another important finding in this research is, some actions are better recognized in certain viewpoints. This is providentially benefit to the vision-based HMA system where features investigation towards each viewpoint on specific action could be a possible extension to enhance the performance in view invariant HMA framework.

The second argument in the thesis is regarding the classification ambiguity where the final classification results might be confused with other possible classes. These scenarios are not so effective to be interpreted with crisp or binary answers as both might be correct. A validity test with online survey had proven the existence of such cases and it is designated as "non-mutually exclusive" case in the thesis. It is important to raise the awareness in the research community regarding this very important, but largely neglected issue. This thesis had studied the ambiguous case in HMA such as the confusions in view estimation and action recognition and as well as in the scene understanding. This is because scene context itself could be an extra cue in human action interpretation and regretfully to be ignored. As a solution, FQRC had been proposed to address the ambiguous results with first model the uncertainties in the early stage to construct the FQTM and output the multi-label ranking result at the inference stage. In addition, DFQRC was proposed to overcome the infeasibility of FQRC by introducing the adaptive model in FQTM learning. The effectiveness and efficiency of FQRC and DFQRC had been tested extensively and theirs capability in addressing the ambiguous cases had been confirmed and outperformed the state-of-the-art methods. Most importantly, it provides promising results in view estimation, action recognition, and scene understanding.

6.2 Limitations

Due to the narrowed scope of this thesis, the current works is limited to the human actions that are available in the testing datasets. The core of this thesis is to research on the

methods that are better in addressing the aforementioned uncertainties. And thus, current testing dataset is sufficient to achieve the objective. This indicates that more works can be done to realize the proposed framework into the real-world applications with some of them are listed in the future works.

6.3 Future Works

Although the results presented here have demonstrated the effectiveness of the proposed approaches in dealing with different uncertainties, the works provide basis for further research in several areas.

6.3.1 Expand the actions

The current framework is constrained to limited actions in the testing dataset. This is because the scope of this thesis is mainly focus on the validation of the proposed methods and evaluate the performance of the framework. Many works need to be done in expanding the action bank especially in creating the VSAM to support more actions. Such extension is application based where it helps in the realization of the proposed framework to works in the real-world environment.

6.3.2 Early Event Detection

Apart from that, fuzzy qualitative approaches being successful in handling the uncertainties in various HMA applications as highlighted in chapter 2, can be very well explored to be potentially applied in highly complex HMA applications such as human activity forecasting (Kitani et al., 2012) and early detection of crimes (Hoai & De la Torre, 2012; Ryoo, 2011). There do not exist literature on the fuzzy capability in handling the uncertainties arising in such scenarios, which have high quotient of importance as they are focusing on forecasting an event or early detecting crimes from happening. Therefore,

even the minutest level of uncertainty is required to be taken care of for reliable decision making. Fuzzy qualitative reasoning with its capability in handling the uncertain situations can substantially benefit in performing these complex tasks.

6.3.3 Human Activity Recognition in Still Images

Another interesting area to be explored as part of the future works is the recognition of human activities using still image. The work has received much attention in the recent past in the computer vision community (Delaitre et al., 2010; Desai et al., 2010; Gupta et al., 2009; Maji et al., 2011; Prest et al., 2012; W. Yang et al., 2010; Yao & Fei-Fei, 2010). In this research topic, most of the works considered it to be same as an image classification problem. Lately, several researchers are trying to obtain a thorough understanding of the human poses, the objects, and the interactions between them in a still images to infer the activities. For example, Yao & Fei-Fei (2012) proposed a method to recognize the human-object interactions in still images by explicitly modelling the mutual context between the human poses and the objects, so that each can facilitate the recognition of the other. Their mutual context model outperform the state-of-the-art in object detection, human pose estimation, as well as the recognition of human-object interaction activities. However limited information that can be extracted from the still image may induce the ambiguity in the classification task. This can be a potential area to explore by utilizing the proposed method in handling the uncertainties, incomplete data or vague information in regards with the human-object interactions, or human-scene context in still images.

REFERENCES

Acampora, G., Foggia, P., Saggese, A., & Vento, M. (2012). Combining neural networks and fuzzy systems for human behavior understanding. In *Ieee ninth international conference on advanced video and signal-based surveillance* (pp. 88–93).

Aggarwal, J. K., & Cai, Q. (1997). Human motion analysis: A review. In *Ieee nonrigid and articulated motion workshop* (pp. 90–102).

Aggarwal, J. K., Cai, Q., Liao, W., & Sabata, B. (1994). Articulated and elastic non-rigid motion: A review. In *Ieee workshop on motion of non-rigid and articulated objects* (pp. 2–14).

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, *43*(3), 16.

Ahmad, M., & Lee, S.-W. (2006). Hmm-based human action recognition using multiview image sequences. In *International conference on pattern recognition* (Vol. 1, p. 263 -266).

Al-Jarrah, O., & Halawani, A. (2001). Recognition of gestures in arabic sign language using neuro-fuzzy systems. *Artificial Intelligence*, *133*(1), 117–138.

Anderson, D., Keller, J. M., Skubic, M., Chen, X., & He, Z. (2006). Recognizing falls from silhouettes. In *International conference of the ieee engineering in medicine and biology society* (pp. 6388–6391).

Anderson, D., Luke, R. H., Keller, J. M., & Skubic, M. (2008). Extension of a soft-computing framework for activity analysis from linguistic summarizations of video. In *Ieee international conference on fuzzy systems* (pp. 1404–1410).

Anderson, D., Luke, R. H., Keller, J. M., Skubic, M., Rantz, M., & Aud, M. (2009a). Linguistic summarization of video for fall detection using voxel person and fuzzy logic. *Computer Vision and Image Understanding*, *113*(1), 80–89.

Anderson, D., Luke, R. H., Keller, J. M., Skubic, M., Rantz, M. J., & Aud, M. A. (2009b). Modeling human activity from voxel person using fuzzy logic. *IEEE Transactions on Fuzzy Systems*, *17*(1), 39–49.

Anderson, D., Luke III, R. H., Stone, E. E., & Keller, J. M. (2009). Fuzzy voxel object. In *World congress of the international fuzzy systems association / conference of the european society for fuzzy logic and technology* (pp. 282–287).

Angelov, P., & Filev, D. (2005). Simpl_ets: a simplified method for learning evolving takagi-sugeno fuzzy models. In *Ieee international conference on fuzzy systems* (pp. 1068–1073).

Angelov, P., Ramezani, R., & Zhou, X. (2008). Autonomous novelty detection and object tracking in video streams using evolving clustering and takagi-sugeno type neuro-fuzzy system. In *Ieee international joint conference on neural networks* (pp. 1456–1463).

Angelov, P. P., & Filev, D. P. (2004). An approach to online identification of takagi-sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(1), 484–498.

Ashraf, N., Shen, Y., Cao, X., & Foroosh, H. (2013). View invariant action recognition using weighted< i> fundamental ratios</i>. *Computer Vision and Image Understanding*.

Balcilar, M., & Sonmez, A. C. (2013). Region based fuzzy background subtraction using choquet integral. In *Adaptive and natural computing algorithms* (pp. 287–296). Springer.

Beauchemin, S. S., & Barron, J. L. (1995). The computation of optical flow. *ACM Computing Surveys*, *27*(3), 433–466.

Bhattacharyya, S., Dutta, P., & Maulik, U. (2007). Binary object extraction using bi-directional self-organizing neural network (bdsonn) architecture with fuzzy context sensitive thresholding. *Pattern Analysis and Applications*, *10*(4), 345–360.

Bhattacharyya, S., & Maulik, U. (2013). Target tracking using fuzzy hostility induced segmentation of optical flow field. In *Soft computing for image and multimedia data processing* (pp. 97–107). Springer.

Bhattacharyya, S., Maulik, U., & Dutta, P. (2009). High-speed target tracking by fuzzy hostility-induced segmentation of optical flow field. *Applied Soft Computing*, *9*(1), 126–134.

Binh, N. D., & Ejima, T. (2005). Hand gesture recognition using fuzzy neural network. In *Conference on graphics, vision and image proces* (pp. 1–6).

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Ieee international conference on computer vision* (Vol. 2, pp. 1395–1402).

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267.

Bobick, A. F., & Wilson, A. D. (1995). A state-based technique for the summarization and recognition of gesture. In *International conference on computer vision* (pp. 382–388).

Bosch, A., Zisserman, A., & Munoz, X. (2006). Scene classification via plsa. In *European conference on computer vision* (pp. 517–530). Springer.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern recognition*, *37*(9), 1757–1771.

Bouwmans, T., El Baf, F., et al. (2009). Modeling of dynamic backgrounds by type-2 fuzzy gaussians mixture models. *MASAUM Journal of of Basic and Applied Sciences*, *1*(2), 265–276.

Bregler, C., Malik, J., & Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, *56*(3), 179–194.

Bregonzio, M., Gong, S., & Xiang, T. (2009). Recognising action as clouds of space-time interest points. In *Ieee conference on computer vision and pattern recognition* (pp. 1948–1955).

Calvo-Gallego, E., Brox, P., & Sánchez-Solano, S. (2013). A fuzzy system for background modeling in video sequences. In *Fuzzy logic and applications* (pp. 184–192). Springer.

Campbell, L. W., & Bobick, A. F. (1995). Recognition of human body motion using phase space constraints. In *International conference on computer vision* (pp. 624–630).

Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B., & Kasturi, R. (2010). Understanding transit scenes: A survey on human behavior-recognition algorithms. *IEEE Transactions on Intelligent Transportation Systems*, *11*(1), 206–224.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, *3*(5), 698–713.

Cédras, C., & Shah, M. (1995). Motion-based recognition a survey. *Image and vision computing*, *13*(2), 129–155.

Chan, C. S., Coghill, G. M., & Liu, H. (2011). Recent advances in fuzzy qualitative reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *19*(03), 417–422.

Chan, C. S., & Liu, H. (2009). Fuzzy qualitative human motion analysis. *IEEE Transactions on Fuzzy Systems*, *17*(4), 851–862.

Chan, C. S., Liu, H., Brown, D., & Kubota, N. (2008). A fuzzy qualitative approach to human motion recognition. In *Ieee international conference on fuzzy systems* (pp. 1242–1249).

Chan, C. S., Liu, H., & Brown, D. J. (2007). Recognition of human motion from qualitative normalised templates. *Journal of Intelligent and Robotic Systems*, *48*(1), 79-95.

Chan, C. S., Liu, H., & Lai, W. K. (2010). Fuzzy qualitative complex actions recognition. In *Ieee international conference on fuzzy systems* (pp. 1–8).

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*.

Chen, G., Xie, Q., & Shieh, L. S. (1998). Fuzzy kalman filtering. *Information Sciences*, *109*(1), 197–209.

Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, *34*(15), 1995–2006.

Chen, X., He, Z., Anderson, D., Keller, J., & Skubic, M. (2006). Adaptive silouette extraction and human tracking in complex and dynamic environments. In *Ieee international conference on image processing* (pp. 561–564).

Chen, X., He, Z., Keller, J. M., Anderson, D., & Skubic, M. (2006). Adaptive silhouette extraction in dynamic environments using fuzzy logic. In *Ieee international conference on fuzzy systems* (pp. 236–243).

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Cheung, S.-C. S., & Kamath, C. (2004). Robust techniques for background subtraction in urban traffic video. In *Proceedings of spie* (Vol. 5308, pp. 881–892).

Chowdhury, A., & Tripathy, S. S. (2014). Detection of human presence in a surveillance video using fuzzy approach. In *International conference on signal processing and integrated networks* (pp. 216–219).

Cristani, M., Raghavendra, R., Del Bue, A., & Murino, V. (2013). Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, *100*, 86–97.

Cucchiara, R., Grana, C., Prati, A., & Vezzani, R. (2005). Probabilistic posture classification for human-behavior analysis. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *35*(1), 42–54.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Ieee international conference of computer vision and pattern recognition* (Vol. 1, pp. 886–893).

Delaitre, V., Laptev, I., & Sivic, J. (2010). Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *British machine vision conference* (Vol. 2, p. 7).

Desai, C., Ramanan, D., & Fowlkes, C. (2010). Discriminative models for static human-object interactions. In *Ieee computer society conference on computer vision and pattern recognition workshops* (pp. 9–16).

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Ieee international conference on computer vision* (pp. 726–733).

El Baf, F., Bouwmans, T., & Vachon, B. (2008a). A fuzzy approach for background subtraction. In *Ieee international conference on image processing* (pp. 2648–2651).

El Baf, F., Bouwmans, T., & Vachon, B. (2008b). Fuzzy integral for moving object detection. In *Ieee international conference on fuzzy systems* (pp. 1729–1736).

El Baf, F., Bouwmans, T., & Vachon, B. (2008c). Type-2 fuzzy mixture of gaussians model: application to background modeling. In *Advances in visual computing* (pp. 772–781). Springer.

El Baf, F., Bouwmans, T., & Vachon, B. (2009). Fuzzy statistical modeling of dynamic

backgrounds for moving object detection in infrared videos. In *Ieee computer society conference on computer vision and pattern recognition workshop* (pp. 60–65).

Elliott, R. J., Aggoun, L., & Moore, J. B. (1995). *Hidden markov models*. Springer.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Ieee computer society conference on computer vision and pattern recognition* (Vol. 2, pp. 524–531).

Forguson, L., & Gopnik, A. (1998). The ontogeny of common sense. *Developing Theories of Mind*, 226–243.

García, J., Molina, J. M., Besada, J. A., Portillo, J. I., & Casar, J. R. (2002). Robust object tracking with fuzzy shape estimation. In *International conference on information fusion* (Vol. 1, pp. 64–71).

Garcia, J., Patricio, M. A., Berlanga, A., & Molina, J. M. (2011). Fuzzy region assignment for visual tracking. *Soft Computing*, *15*(9), 1845–1864.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer vision and image understanding*, *73*(1), 82–98.

Gkalelis, N., Tefas, A., & Pitas, I. (2008). Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1511–1521.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(12), 2247–2253.

Gorelick, L., Galun, M., Sharon, E., Basri, R., & Brandt, A. (2006a). Shape representation and classification using the poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1991–2005.

Gorelick, L., Galun, M., Sharon, E., Basri, R., & Brandt, A. (2006b, dec.). Shape representation and classification using the poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1991 -2005.

Guo, Y., Xu, G., & Tsuji, S. (1994). Tracking human body motion based on a stick figure model. *Journal of Visual Communication and Image Representation*, *5*(1), 1–9.

Gupta, A., Kembhavi, A., & Davis, L. S. (2009). Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1775–1789.

Haering, N., Venetianer, P. L., & Lipton, A. (2008). The evolution of video surveillance: an overview. *Machine Vision and Applications*, *19*(5-6), 279–290.

Heisele, B., Kressel, U., & Ritter, W. (1997). Tracking non-rigid, moving objects based on color cluster flow. In *Ieee computer society conference on computer vision and pattern recognition* (pp. 257–260).

Hoai, M., & De la Torre, F. (2012). Max-margin early event detectors. In *Ieee conference on computer vision and pattern recognition* (pp. 2863–2870).

Holte, M. B., Tran, C., Trivedi, M. M., & Moeslund, T. B. (2011). Human action recognition using multiple views: a comparative perspective on recent developments. In *Proceedings of the joint acm workshop on human gesture and behavior understanding* (pp. 47–52).

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. In *Technical symposium east* (pp. 319–331).

Hosseini, M.-S., & Eftekhari-Moghadam, A.-M. (2013). Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. *Applied Soft Computing*, *13*(2), 846–866.

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34*(3), 334–352.

Hu, W., Xie, D., Tan, T., & Maybank, S. (2004). Learning activity patterns using fuzzy self-organizing neural network. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(3), 1618–1626.

Huntsberger, T. L., Rangarajan, C., & Jayaramamurthy, S. N. (1986). Representation of uncertainty in computer vision using fuzzy sets. *IEEE Transactions on Computers*, *100*(2), 145–156.

Hussain, B., & Kabuka, M. R. (1994). A novel feature recognition neural network and its application to character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(1), 98–106.

Ikizler-Cinbis, N., & Sclaroff, S. (2010). Object, scene and actions: Combining multiple features for human action recognition. In *European conference on computer vision* (pp. 494–507). Springer.

Iosifidis, A., Tefas, A., Nikolaidis, N., & Pitas, I. (2012). Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis. *Computer Vision and Image Understanding*, *116*(3), 347–360.

Iosifidis, A., Tefas, A., & Pitas, I. (2011). Person specific activity recognition using fuzzy learning and discriminant analysis. In *European signal processing conference* (pp. 1974–1978).

Iosifidis, A., Tefas, A., & Pitas, I. (2012a). Activity-based person identification using fuzzy representation and discriminant learning. *IEEE Transactions on Information Forensics and Security*, *7*(2), 530–542.

Iosifidis, A., Tefas, A., & Pitas, I. (2012b). Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis. *Signal Processing*.

Iosifidis, A., Tefas, A., & Pitas, I. (2013). Minimum class variance extreme learning machine for human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(11), 1968–1979.

Iwai, Y., Ogaki, K., & Yachida, M. (1999). Posture estimation using structure and motion models. In *Ieee international conference on computer vision* (Vol. 1, pp. 214–219).

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, *108*(1), 116–134.

Ji, X., & Liu, H. (2010). Advances in view-invariant human motion analysis: a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *40*(1), 13–24.

Ju, S. X., Black, M. J., & Yacoob, Y. (1996). Cardboard people: A parameterized model of articulated image motion. In *International conference on automatic face and gesture recognition* (pp. 38–44).

Juang, C.-F., & Chang, C.-M. (2007). Human body posture classification by a neural fuzzy network and home care system application. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *37*(6), 984–994.

Kakadiaris, I. A., & Metaxas, D. (1996). Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *Ieee conference on computer vision and pattern recognition* (pp. 81–87).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*(1), 35–45.

Kamel, H., & Badawy, W. (2005). Fuzzy logic based particle filter for tracking a maneuverable target. In *Symposium on circuits and systems* (pp. 1537–1540).

Karayiannis, N. B., & Pai, P.-I. (1995). Fuzzy vector quantization algorithms and their application in image compression. *IEEE Transactions on Image Processing*, *4*(9), 1193–1201.

Karnik, N. N., & Mendel, J. M. (1998). Type-2 fuzzy logic systems: type-reduction. In *Ieee international conference on systems, man, and cybernetics* (Vol. 2, pp. 2046–2051).

Karnik, N. N., Mendel, J. M., & Liang, Q. (1999). Type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, *7*(6), 643–658.

Kim, I. S., Choi, H. S., Yi, K. M., Choi, J. Y., & Kong, S. G. (2010). Intelligent visual surveillance—a survey. *International Journal of Control, Automation and Systems*, *8*(5), 926–939.

Kim, Y.-J., Won, C.-H., Pak, J.-M., & Lim, M.-T. (2007). Fuzzy adaptive particle filter for localization of a mobile robot. In *Knowledge-based intelligent information and engineering systems* (pp. 41–48).

Kimura, M., & Saito, H. (2001). 3d reconstruction based on epipolar geometry. *Transactions on Information and Systems*, *84*(12), 1690–1697.

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., & Hebert, M. (2012). Activity forecasting. In *European conference on computer vision* (pp. 201–214). Springer.

Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In *Eee applied imagery pattern recognition workshop* (pp. 1–8).

Kobayashi, K., Cheok, K. C., Watanabe, K., & Munekata, F. (1998). Accurate differential global positioning system via fuzzy logic kalman filter sensor fusion technique. *IEEE Transactions on Industrial Electronics*, *45*(3), 510–518.

Kohler, M. (1997). *Using the kalman filter to track human interactive motion: modelling and initialization of the kalman filter for translational motion*. Citeseer.

Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Ieee conference on computer vision and pattern recognition* (pp. 1446–1453).

Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, *1*(2), 98–110.

Kuipers, B. (1986). Qualitative simulation. *Artificial intelligence*, *29*(3), 289–338.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*, 107-123.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Ieee conference on computer vision and pattern recognition* (pp. 1–8).

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys*, *15*(3), 1192–1209.

Lee, C.-S., & Elgammal, A. (2007). Homeomorphic manifold analysis: Learning decomposable generative models for human motion analysis. In *Dynamical vision* (pp. 100–114). Springer.

Leung, M. K., & Yang, Y.-H. (1995). First sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(4), 359–377.

Lewandowski, M., Makris, D., & Nebel, J.-C. (2010). View and style-independent action manifolds for human activity recognition. In *European conference on computer vision* (p. 547-560). Springer Berlin Heidelberg.

Lewandowski, M., Martinez-del Rincon, J., Makris, D., & Nebel, J.-C. (2010). Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. In *International conference on pattern recognition* (pp. 161–164).

Le Yaouanc, J.-M., & Poli, J.-P. (2012). A fuzzy spatio-temporal-based approach for activity recognition. In *Advances in conceptual modeling* (pp. 314–323). Springer.

Li, X. (2003). Gesture recognition based on fuzzy c-means clustering algorithm. *Department Of Computer Science The University Of Tennessee Knoxville*.

Li, Z., Liu, W., & Zhang, Y. (2012). Adaptive fuzzy apporach to background modeling using pso and klms. In *World congress on intelligent control and automation* (pp. 4601–4607).

Liang, Q., & Mendel, J. M. (2000). Interval type-2 fuzzy logic systems: theory and design. *IEEE Transactions on Fuzzy Systems*, *8*(5), 535–550.

Lin, C., Chung, I., & Sheu, L. (2000). A neural fuzzy system for image motion estimation. *Fuzzy sets and systems*, *114*(2), 281–304.

Liu, H. (2008). A fuzzy qualitative framework for robot intelligent connection. In *Ieee international conference on fuzzy systems* (pp. 1556–1562).

Liu, H., Brown, D. J., & Coghill, G. M. (2008a). A fuzzy qualitative framework for connecting robot qualitative and quantitative representations. *IEEE Transactions on Fuzzy Systems*, *16*(3), 808–822.

Liu, H., Brown, D. J., & Coghill, G. M. (2008b). Fuzzy qualitative robot kinematics. *IEEE Transactions on Fuzzy Systems*, *16*(3), 808–822.

Liu, H., Coghill, G. M., & Barnes, D. P. (2009). Fuzzy qualitative trigonometry. *International Journal of Approximate Reasoning*, *51*(1), 71–88.

Loy, G., Eriksson, M., Sullivan, J., & Carlsson, S. (2004). Monocular 3d reconstruction of human motion in long action sequences. In *European conference on computer vision* (pp. 442–455). Springer.

Lyons, M. J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(12), 1357–1362.

Maddalena, L., & Petrosino, A. (2010). A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection. *Neural Computing and Applications*, *19*(2), 179–186.

Mahapatra, A., Mishra, T. K., Sa, P. K., & Majhi, B. (2013). Background subtraction and human detection in outdoor videos using fuzzy logic. In *Ieee international conference on fuzzy systems* (pp. 1–7).

Maji, S., Bourdev, L., & Malik, J. (2011). Action recognition from a distributed representation of pose and appearance. In *Ieee conference on computer vision and pattern recognition* (pp. 3177–3184).

Marichal, J.-L. (2000). On sugeno integral as an aggregation function. *Fuzzy Sets and Systems*, *114*(3), 347–365.

Marins, J. L., Yun, X., Bachmann, E. R., McGhee, R. B., & Zyda, M. J. (2001). An extended kalman filter for quaternion-based orientation estimation using marg sensors. In *International conference on intelligent robots and systems* (Vol. 4, pp. 2003–2011).

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Ieee conference on computer vision and pattern recognition* (pp. 2929–2936).

Mel, B. W. (1997). Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural computation*, *9*(4), 777–804.

Mendel, J. M., & John, R. B. (2002). Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, *10*(2), 117–127.

Mendel, J. M., John, R. I., & Liu, F. (2006). Interval type-2 fuzzy logic systems made simple. *IEEE Transactions on Fuzzy Systems*, *14*(6), 808–821.

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *37*(3), 311–324.

Miyamoto, S., & Umayahara, K. (1998). Fuzzy clustering by quadratic regularization. In *Ieee internation conference on fuzzy systems* (Vol. 2, pp. 1394–1399).

Moeslund, T. B., & Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, *81*(3), 231–268.

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2), 90–126.

Mozafari, K., Charkari, N. M., Boroujeni, H. S., & Behrouzifar, M. (2012). A novel fuzzy hmm approach for human action recognition in video. In *Knowledge technology* (pp. 184–193). Springer.

Murofushi, T., & Sugeno, M. (1989). An interpretation of fuzzy measures and the choquet integral as an integral with respect to a fuzzy measure. *Fuzzy sets and Systems*, *29*(2), 201–227.

Narukawa, Y., & Murofushi, T. (2004). Decision modelling using the choquet integral. In *Modeling decisions for artificial intelligence* (pp. 183–193). Springer.

Ning, H., Tan, T., Wang, L., & Hu, W. (2004). Kinematics-based tracking of human walking in monocular video sequences. *Image and Vision Computing*, *22*(5), 429–441.

Niyogi, S. A., & Adelson, E. H. (1994). Analyzing and recognizing walking figures in xyt. In *Ieee conference on computer vision and pattern recognition* (pp. 469–474).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 831–843.

Pal, N. R., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. *IEEE Transaction on Fuzzy Systems*, *13*(4), 517–530.

Parikh, D., & Grauman, K. (2011). Relative attributes. In *Ieee international conference on computer vision* (pp. 503–510).

Pece, A. E. (2002). From cluster tracking to people counting. In *Ieee workshop on performance evaluation of tracking and surveillance* (pp. 9–17).

Pentland, A. (2000). Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 107–119.

Piccardi, M. (2004). Background subtraction techniques: a review. In *Ieee international conference on systems, man and cybernetics* (Vol. 4, pp. 3099–3104).

Platt, J., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin dags for multiclass classification. *Advances in neural information processing systems*, *12*(3), 547–553.

Popoola, O. P., & Wang, K. (2012). Video-based abnormal human behavior recognition - a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C Applications and Reviews*, *42*(6), 865–878.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer vision and image understanding*, *108*(1), 4–18.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*(6), 976–990.

Prest, A., Schmid, C., & Ferrari, V. (2012). Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 601–614.

Rehg, J. M., & Kanade, T. (1995). Model-based tracking of self-occluding articulated objects. In *International conference on computer vision* (pp. 612–617).

Rogez, G., Orrite, C., Guerrero, J., & Torr, P. H. (2014). Exploiting projective geometry for view-invariant monocular human motion analysis in man-made environments. *Computer Vision and Image Understanding*(0), -.

Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *Image understanding*, *59*(1), 94–115.

Rudoy, D., & Zelnik-Manor, L. (2012). Viewpoint selection for human actions. *International journal of computer vision*, *97*(3), 243–254.

Ryoo, M. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Ieee international conference on computer vision* (pp. 1036–1043).

Sasiadek, J., & Khe, J. (2001). Sensor fusion based on fuzzy kalman filter. In *International workshop on robot motion and control* (pp. 275–283).

Sasiadek, J., & Wang, Q. (1999). Sensor fusion based on fuzzy kalman filtering for autonomous robot vehicle. In *Ieee international conference on robotics and automation* (Vol. 4, pp. 2970–2975).

Sasiadek, J., Wang, Q., & Zeremba, M. (2000). Fuzzy adaptive kalman filtering for ins/gps data fusion. In *Ieee international symposium on intelligent control* (pp. 181–186).

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *International conference on pattern recognition* (Vol. 3, pp. 32–36).

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *International conference on multimedia* (pp. 357–360).

See, J., Lee, S., & Hanmandlu, M. (2005). Human motion detection using fuzzy rule-base classification of moving blob regions. In *International conference on robotics, vision, information and signal processing* (pp. 398–402).

Senthil, R., Janarthanan, K., & Prakash, J. (2006). Nonlinear state estimation using fuzzy kalman filter. *Industrial & engineering chemistry research*, *45*(25), 8678–8688.

Shakeri, M., Deldari, H., Foroughi, H., Saberi, A., & Naseri, A. (2008). A novel fuzzy background subtraction method based on cellular automata for urban traffic applications. In *International conference on signal processing* (pp. 899–902).

Shen, Q., & Leitch, R. (1993). Fuzzy qualitative simulation. *IEEE Transactions on Systems, Man and Cybernetics*, *23*(4), 1038–1061.

Shen, Q., Leitch, R., & Coghill, G. (1993). Fuzzy qualitative modelling. In *Ieee colloquium on two decades of fuzzy control-part 2* (pp. 3–1).

Shimazaki, H., & Shinomoto, S. (2007). A method for selecting the bin size of a time histogram. *Neural Computation*, *19*(6), 1503–1527.

Sigari, M. H., Mozayani, N., & Pourreza, H. R. (2008). Fuzzy running average and fuzzy background subtraction: concepts and application. *International Journal of Computer Science and Network Security*, *8*(2), 138–143.

Silaghi, M.-C., Plänkers, R., Boulic, R., Fua, P., & Thalmann, D. (1998). Local and global skeleton fitting techniques for optical motion capture. In *Modelling and motion capture techniques for virtual environments* (pp. 26–40). Springer.

Sugeno, M., & Kwon, S.-H. (1995). A new approach to time series modeling with fuzzy measures and the choquet integral. In *Ieee international conference on fuzzy systems* (Vol. 2, pp. 799–804).

Sullivan, M., & Shah, M. (2008). Action mach: Maximum average correlation height filter for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*.

Tahani, H., & Keller, J. M. (1990). Information fusion in computer vision using the fuzzy integral. *IEEE Transactions on Systems, Man and Cybernetics*, *20*(3), 733–741.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Tsujinishi, D., & Abe, S. (2003). Fuzzy least squares support vector machines for multiclass problems. *Neural Networks*, *16*(5), 785–792.

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473–1488.

Várkonyi-Kóczy, A. R., & Tusor, B. (2011). Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models. *IEEE Transactions on Instrumentation and Measurement*, *60*(5), 1505–1514.

Verma, R., & Dev, A. (2009). Vision based hand gesture recognition using finite state machines and fuzzy logic. In *International conference on ultra modern telecommunications & workshops* (pp. 1–6).

Vogel, J., & Schiele, B. (2007). Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, *72*(2), 133–157.

von Luxburg, U., & Schölkopf, B. (2008). Statistical learning theory: models, concepts, and results. *arXiv preprint arXiv:0810.4752*.

Wachs, J., Kartoun, U., Stern, H., & Edan, Y. (2002). Real-time hand gesture telerobotic system using fuzzy c-means clustering. In *Biannual world automation congress* (Vol. 13, pp. 403–409).

Wachs, J. P., Stern, H., & Edan, Y. (2005). Cluster labeling and parameter estimation for the automated setup of a hand-gesture recognition system. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *35*(6), 932–944.

Wachter, S., & Nagel, H.-H. (1997). Tracking of persons in monocular image sequences. In *Ieee nonrigid and articulated motion workshop* (pp. 2–9).

Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, *51*(1), 59–64.

Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognition*, *36*(3), 585–601.

Wang, Z., & Zhang, J. (2008). Detecting pedestrian abnormal behavior based on fuzzy associative memory. In *Conference on natural computation* (Vol. 6, pp. 143–147).

Weinland, D., Boyer, E., & Ronfard, R. (2007, oct.). Action recognition from arbitrary views using 3d exemplars. In *International conference on computer vision* (p. 1 -7).

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, *104*(2), 249–257.

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, *115*(2), 224–241.

Welch, G., & Bishop, G. (1995). *An introduction to the kalman filter.*

Welch, G. F. (2009). History: The use of the kalman filter for human motion tracking in virtual reality. *Presence: Teleoperators and Virtual Environments*, *18*(1), 72–91.

Wilson, A. D., & Bobick, A. F. (1999). Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(9), 884–900.

Wongkhuenkaew, R., Auephanwiriyakul, S., & Theera-Umpon, N. (2013). Multi-prototype fuzzy clustering with fuzzy k-nearest neighbor for off-line human action recognition. In *Ieee international conference on fuzzy systems* (pp. 1–7).

Wu, H., Sun, F., & Liu, H. (2008). Fuzzy particle filtering for uncertain systems. *IEEE Transactions on Fuzzy Systems*, *16*(5), 1114–1129.

Wu, S., Moore, B. E., & Shah, M. (2010). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Ieee conference on computer vision and pattern recognition* (pp. 2054–2060).

Wu, Y., & Huang, T. S. (1999). Vision-based gesture recognition: A review. In *Gesture-based communication in human-computer interaction* (pp. 103–115). Springer.

Xie, D., Hu, W., Tan, T., & Peng, J. (2004). A multi-object tracking system for surveillance video analysis. In *International conference on pattern recognition* (Vol. 4, pp. 767–770).

Xu, R., Wunsch, D., et al. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678.

Yager, R. R., & Zadeh, L. A. (1992). *An introduction to fuzzy logic applications in intelligent systems*. Springer.

Yamato, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *Ieee conference on computer vision and pattern recognition* (pp. 379–385).

Yang, W., Wang, Y., & Mori, G. (2010). Recognizing human actions from still images with latent poses. In *Ieee conference on computer vision and pattern recognition* (pp. 2030–2037).

Yang, Y., Hao, A., & Zhao, Q. (2008). View-invariant action recognition using interest points. In *International conference on multimedia information retrieval* (pp. 305–312).

Yao, B., & Fei-Fei, L. (2010). Grouplet: A structured image representation for recognizing human and object interactions. In *Ieee conference on computer vision and pattern recognition* (pp. 9–16).

Yao, B., & Fei-Fei, L. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1691–1703.

Yao, B., Hagras, H., Al Ghazzawi, D., & Alhaddad, M. J. (2012). An interval type-2 fuzzy logic system for human silhouette extraction in dynamic environments. In *Autonomous and intelligent systems* (pp. 126–134). Springer.

Yao, B., Hagras, H., Alhaddad, M. J., & Alghazzawi, D. (2014). A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments. *Soft Computing*, 1–8.

Yilma, A., & Shah, M. (2005, oct.). Recognizing human actions in videos acquired by uncalibrated moving cameras. In *International conference on computer vision* (Vol. 1, p. 150 -157 Vol. 1).

Yoon, C., Cheon, M., & Park, M. (2013). Object tracking from image sequences using adaptive models in fuzzy particle filter. *Information Sciences*, *253*, 74–99.

Yu, M., Naqvi, S. M., Rhuma, A., & Chambers, J. (2011). Fall detection in a smart room by using a fuzzy one class support vector machine and imperfect training data. In *Ieee international conference on acoustics, speech and signal processing* (pp. 1833–1836).

Yun, X., Aparicio, C., Bachmann, E. R., & McGhee, R. B. (2005). Implementation and experimental results of a quaternion-based kalman filter for human body motion tracking. In *Ieee international conference on robotics and automation* (pp. 317–322).

Yun, X., & Bachmann, E. R. (2006). Design, implementation, and experimental results of a quaternion-based kalman filter for human body motion tracking. *IEEE Transactions on Robotics*, *22*(6), 1216–1227.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zadeh, L. A. (1988). Fuzzy logic. *Computer*, *21*(4), 83–93.

Zelnik-Manor, L., & Irani, M. (2001). Event-based analysis of video. In *Ieee conference on computer vision and pattern recognition* (Vol. 2, pp. II–123).

Zeng, J., Xie, L., & Liu, Z.-Q. (2008). Type-2 fuzzy gaussian mixture models. *Pattern Recognition*, *41*(12), 3636–3643.

Zhang, H., & Xu, D. (2006). Fusing color and texture features for background model. In *International conference on fuzzy systems and knowledge discovery* (pp. 887–893).

Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038 - 2048.

Zhao, Z., Bouwmans, T., Zhang, X., & Fang, Y. (2012). A fuzzy background modeling approach for motion detection in dynamic backgrounds. In *Multimedia and signal processing* (pp. 177–185). Springer.

Appendices

PUBLICATIONS

Journals

Lim, C. H., Ekta Vats, & Chan, C. S. (2015). Fuzzy Human Motion Analysis, *Pattern Recognition*, 48(5), 1773-1796.

Lim, C. H., & Chan, C. S. (2015). Fuzzy qualitative human model for viewpoint identification. *Neural Computing and Applications*, 1-12.

Lim, C. H., Risnumawan, A., & Chan, C. S., (2014). Scene Image is Non-Mutually Exclusive - A Fuzzy Qualitative Scene Understanding. *IEEE Transactions on Fuzzy Systems*, 22(6), 1541-1556.

Conferences

Lim, C. H., & Chan, C. S. (2013). Fuzzy action recognition for multiple views within single camera. *In Fuzzy Systems (FUZZ), 2013 IEEE International Conference on (pp. 1-8)*.

Lim, C. H., & Chan, C. S. (2012). A fuzzy qualitative approach for scene classification. In *Fuzzy Systems (FUZZ), 2012 IEEE International Conference on (pp. 1-8)*.

Lim, C. H., & Chan, C. S. (2011). A Framework on Fuzzy Intrusion Detection. In *Advanced Computational Intelligence and Intelligent Informatics (IWACIII)*.

INTUITION OF USING FUZZY MEMBERSHIP

In this section, The intuitive idea of using 4-tuples fuzzy membership function in the proposed framework is discussed. If loss function is defined as,

$$\ell(f_i(\mathbf{x}), y) = \begin{cases} 0 & \text{if } y = \max\limits_{k \in \{1,\ldots,K\}} \mathsf{r}^i(\mathbf{x}) \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathsf{r}^i(\mathbf{x}) = \{r_1^i, \ldots, r_K^i | r_k^i \in [0,1]\}$ is the output of the inference of function i, the scalar output r_k^i is defined in (5.6) and $\sum_{k=1}^{K} r_k^i = 1$. Suppose to have finitely g functions, then, the objective is to find a function $f^*(\mathbf{x})$ that minimize the loss function,

$$f^*(\mathbf{x}) = \underset{\mathbf{y} \in \{1,\ldots,K\}}{\arg\min} \sum_{i=1}^{g} \ell(f_i(\mathbf{x}), \mathbf{y}) \tag{2}$$

In order to get the interpretation of (2) we will use the concept of maximum entropy. In information theory, the principle of maximum entropy is to minimize the amount of prior information built into the distribution. More specifically, the structure of maximum entropy problem is to find a probability assignment (or membership function $\mu_{jk} \in [0,1]$) which avoid bias agreeing with any given information. In this case, while looking at (2), the membership function μ_{jk} captures such prior information. Inspired by Miyamoto and Umayahara Miyamoto & Umayahara (1998), the maximum entropy is utilized to get the interpretation of 4-tuples fuzzy number. For simplicity we omit i, and the objective of maximum entropy,

$$\max - \sum_j \sum_k \mu_{jk} \log \mu_{jk} \tag{3}$$

Subject to the constraint $\sum_k \frac{\prod_j \mu_{jk}}{Z} = 1$ and $f^*(\mathbf{x}) = c$, where c is a constant. Then

using Lagrange multipliers,

$$\begin{aligned}\mathscr{J} = &-\sum_j \sum_k \mu_{jk} \log \mu_{jk} + \lambda_1 \left(1 - \sum_k \frac{\prod_j \mu_{jk}}{Z}\right) \\ &+ \lambda_2 (c - f^*(\mathbf{x}))\end{aligned} \tag{4}$$

For simplicity, μ_{jk} is treated as a fixed length vector since $\mathbf{x}$ is assumed to be discrete, then yield,

$$\frac{\partial \mathscr{J}}{\partial \mu_{jk}} = -1 - \log \mu_{jk} - \frac{\lambda_1}{Z} - \lambda_2 \frac{\partial f^*(\mathbf{x})}{\partial \mu_{jk}} \tag{5}$$

By setting $\frac{\partial \mathscr{J}}{\partial \mu_{jk}} = 0$ and get μ_{jk} yields,

$$\mu_{jk} = \exp\left(-\left(1 + \frac{\lambda_1}{Z} + \lambda_2 \frac{\partial f^*(\mathbf{x})}{\partial \mu_{jk}}\right)\right) \tag{6}$$

Actually this result is similar when minimize or maximize the objective function of,

$$\text{min/max} - \sum_j \sum_k \mu_{jk} \log \mu_{jk} - \lambda_2 f^*(\mathbf{x}) \tag{7}$$

with subject to the constraint $\sum_k \frac{\prod_j \mu_{jk}}{Z} = 1$. After taking min-max sign change and make the constant $\lambda = 1/\lambda_2$ for brevity, the following objective is obtained,

$$\begin{aligned}&\min f^*(\mathbf{x}) - \lambda \sum_j \sum_k \mu_{jk} \log \mu_{jk} \\ &\text{subject to} \sum_k \frac{\prod_j \mu_{jk}}{Z} = 1\end{aligned} \tag{8}$$

If compare (8) with the formula of a classifier with regularization, $f + \lambda \mathscr{R}$, the 4-tuples membership function implicitly models the regularization. In details, the 4-tuples membership function with $\mu_{jk} = 1$ (mutually exclusive part) models the classifier while the transition of membership function $[0, 1]$ (non-mutually exclusive part) implicitly models the regularization.