DOCTORAL THESIS

Automatic Recognition of Human Behaviour in Sequential Data

A thesis submitted to Brunel University London
in accordance with the requirements
for award of the degree of Doctor of Philosophy

in

Department of Electronic and Computer Engineering

Rui Qin

March 26, 2019

Declaration of Authorship

I, Rui Qin, declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE: ....
(Signature of student)

Abstract

Automatic human behaviour recognition is a very important element for intelligent Human Computer Interaction (HCI) in which the machine or computer can recognize human behaviour and respond to humans accordingly. Among human behaviour recognition tasks, dynamics represents key information of the action, and it is one of the hardest tasks for automatic recognition. Significant progress has been made in the Artificial Intelligence (A.I.) area recently that provides new tools and technologies to recognize visual objects and make better decisions, logical deductions, mathematical optimization, etc. In this thesis, A.I. technologies have been applied to the analysis of human behaviour, especially using dynamic clues for touch gestures in which a human touches an animal or an object, etc.; micro-gestures that can be extensively utilized on wearable devices; body gestures that are not only used as a paralanguage, but also as an indicator of body behaviour; and 3D facial expression analysis that extracts the emotion information from high quality high-resolution 3D video recordings.

Firstly, an automatic touch gesture recognition system has been proposed, including pre-processing, multiple feature extraction, feature selection, pattern recognition and fusion. Both statistical and video features were extracted, including Motion Statistical Distribution (MSD), Spatial Multi-scale Motion History Histogram (SMMHH), Binary Motion History (BMH), Statistical Distribution (SD) and Local Binary Pattern on Three Orthogonal Planes (LBPTOP). Two powerful machine learning methods, Random Forest and multiboosting, have been utilized. A Sobel edge detection is utilized as pre-processing, and a Minimum Redundancy and Maximum Relevance (mRMR) feature selection is used to reduce the dimension of features after feature extraction. A decision-level fusion method Hierarchical Committee (HC) has been used as a post-processing tool to combine all the predictions. The main contribution of the system is the versatility of it, which can be applied in different dataset. This system also achieves a high performance with maintaining the versatility.

Secondly, another automatic 3D micro-gesture recognition system has been proposed and tested on a Holoscopic Micro 3D Gesture (HoMG) dataset for which a holoscopic 3D video was recorded and annotated. A new system including frame selection by score has been proposed on the video-based dataset. Video-based recognition used LBPTOP and the Local Phase Quantisation from Three Orthogonal Planes (LPQTOP) as feature selection and Support Vector Machine (SVM) as machine learning. Then an SVM prediction

has been utilized, and a score of each frame has been predicted. After using the SVM score to reduce the frames on the video-based dataset, the performance of video-based recognition is improved. This 3D micro-gesture recognition system achieves the best performance comparing with other current works by considering the non-linear relationship of features.

Thirdly, an automatic body gesture recognition system has been proposed to help older people with Chronic Lower Back Pain (CLBP). The proposed system can recognize the behaviours of CLBP patients, like abrupt actions and guarding. A new two-stage machine learning method has been proposed that combines k-nearest neighbour k-NN and HMM and achieves a better recognition performance of body gestures than traditional methods. The contribution of the system is it could detectt and analysis CLBP related body behaviour frame by frame and provide more detailed information about CLBP including the starting, ending and different level of CLBP according to the time series, which would help the future research of CLBP.

Fourthly, a 3D feature expression recognition (FER) system has been proposed to achieve better performance on the most popular posed face 3D FER dataset, BU4D. A latest Background Subtraction method was applied based on tensor for pre-processing to extract the dynamic information on the face. This is the first utilized Tensor Background Subtraction From Compressive Measurements (BSCM) on FER. A deep learning method (e.g. Dynamic Image Net) is used on the dynamic facial information. A comparison between with and without Background Subtraction is made to prove the effectiveness of this method. The contribution of this system is providing a new solution of reduce the calculation resources and reduce the computing time. It also helps to remove noise from the original data and optimise the classification performance.

Finally, a real-time FER system was built for an interactive movie system. In the proposed automatic emotion recognition system, CNN has been used as feature extraction, and SVM is used for classification. The real-time system has been prototyped and exhibited on several occasions. In the final system, several practical problems have been considered, including brightness control and the right face selection from all of the audience. That has improved the accuracy of the practical application. The contribution of the system is to achieve the aim of real-time FER while take care of the balance between system speed, classification accuracy and optimisation in real life using environment.

In summary, several automatic recognition systems for human behaviour have been proposed and applied in real recording data and practical applications. Some methods have versatility, and some are specialized for distinct tasks. All these studies contribute to the development of intelligent HCI.

Acknowledgements

First, I'd like to thank my supervision team, who gives me great help in my PhD research, especially my principal supervisor, Dr. Hongying Meng, who is very serious on academic and has lots of ideas in research. Every time when I don't know how to continue my work, he always finds a new idea to help me with my work. Dr. Hongying Meng has not only helped me on knowledge things like mathematics methods and programming, but also taught me how to do research and how to think. My second supervisor Dr. Alireza Mousavi and Research Development Advisor Prof. Maozhen Li also gives me lots of help on researches and life. It is an honor to have a great supervision team, and I learnt a lot from them.

Also, I want to thank Dr. Huibin Li and his team in Xi'an Jiaotong University. He gives us a great chance to exchange to Xi'an Jiaotong University, and I learn a lot about deep learning from them, especially background subtraction used on dynamic behaviour recognition.

Then, I'd like to thank my classmates, Fan Zhang, Jingxin Liu and Yi Liu. We have cooperated and finished lots of work and researches together. It is a pleasure to work with these excellent classmates and we have learnt a lot from each other.

In the end, I'd like to thank my family, who fully support me on all my decisions and researches. I appreciate their understanding and support on me.

List of publications

In preparation

Qin, R., Liu J., Meng, H. (2018) 'Real-time facial expression recognition system for interactive film application'. 8th International Conference on Affective Computing Intelligent Interaction (ACII 2019).

Submitted:

[1] **Qin, R.**, Meng H. (2018) 'Touch gesture recognition based on distinct features and decision fusion'. IEEE Access.

[2] **Qin, R.**, Liu Y., Swash MR., Li M., Meng H., Lei T. (2019)'A fast automatic holoscopic 3D micro-gesture recognition system for immersive applications'. The 2019 15th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2019).

Published:

[1] Liu, J., Meng, H., Li, M., Zhang, F., **Qin, R.** and Nandi, A. (2018) 'Emotion detection from EEG recordings based on supervised and unsupervised dimension reduction'. Concurrency and Computation: Practice and Experience. ISSN: 1532-0626, 10.1002/cpe.4446

[2] Liu, Y., Meng, H., Swash, M. R., Gaus, Y. F. A., **Qin, R.** (2018) 'Holoscopic 3D micro-gesture database for wearable device interaction'. Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China. pp. 802-807, 10.1109/FG.2018.00129.

[3] **Qin, R.**, Meng, H. and Li, M. (2016) 'Continuous Pain Related Behaviour Recognition from Muscle Activity and Body Movements'. Proceedings of 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Changsha, China. 15 August, 10.1109/FSKD.2016.7603435

[4] Gaus, YFA., Olugbade, T., Jan, A., **Qin, R.**, Liu, J., Zhang, F., et al. (2015) 'Social Touch Gesture Recognition Using Random Forest and Boosting on Distinct Feature Sets'. Proceedings of International Conference on Multimodal Interaction. Seattle, Washington, USA. 9 - 13 November ACM. pp. 399 - 406, 10.1145/2818346.2830599

Contents

List of Figures

List of Tables

List of Acronyms

3D-CNN	3 Dimension Convolution Neural Network
AAM	Active Appearance Model
ADs	Action Descriptors
A.I.	Artificial Intelligence
ASM	Active Shape Model
AUs	Action Units
BMH	Binary Motion History
BSCM	Background Subtraction from Compressive Measurements
BU-3DFE	Binghamton University 3D Facial Expression
BU-4DFE	Binghamton University 4D Facial Expression
CLBP	Chronic Lower Back Pain
CLM	Constrained Local Model
CNN	Convolution Neural Network
CoST	Corpus of Social Touch
DCT	Discrete Cosine Transform
DHT	Discrete Hartlet Transform
DFT	Discrete Fourier Transform
DIN	Dynamic ImageNet
DNN	Deep Neural Network
EEG	Electroencephalography
EMFACS	Emotional Facial Action Coding System
EMG	Electromyography
EOH	Edge Orientation Histogram
FACS	Facial Action Coding System
FACSAID	Facial Action Coding System Affect Interpretation Dictionary
FER	Facial Expression Recognition
FFT	Fast Fourier Transform
GM-LSTM	Long Short Term Memory with Geometric Moment Features
GRU	Gated Recurrent Unit
H3D	Holoscopic 3D
HAART	Human-Animal Affective Robot Touch
HC	Hierarchical Committee
HCI	Human Computer Interaction
HoMG	Holoscopic Micro 3D Gesture
HMM	Hidden Markov Model
k-NN	k-Nearest Neighbours

LBP	Local Binary Patterns
LBPTOP	Local Binary Patterns from Three Orthogonal Planes
LDA	Linear Discriminant Analysis
LPQ	Local Phase Quantisation
LPQTOP	Local Phase Quantisation from Three Orthogonal Planes
LRCN	Long-term Recurrent Convolutional Network
MCDNN	Multi-Column Deep Neural Network
MHH	Motion History Histogram
MHI	Motion History Image
mRMR	Minimum Redundancy and Maximum Relevance
MSD	Motion Statistical Distribution
NN	Neural Network
PCA	Principal Component Analysis
PDM	Point Distribution Model
RF	Random Forest
RPCA	Robust Principal Component Analysis
SD	Statistical Distribution
SMMHH	Spatial Multi-scale Motion History Histogram
SVM	Support Vector Machine
TSC	Two-Stage Classifier

Chapter 1

Introduction

1.1 Background

The development of Artificial Intelligence (A.I.) [10] has been one of the most important technology progresses in 21st century [11]. One common definition of A.I. is human-made machine demonstrated intelligence. Usually A.I. refers to human intelligence technology implemented by ordinary computer programs[12]. The evolution of A.I. has promoted the advantages in multidisciplinary and been widely employed in the areas including voice assistants in smart phones, car navigation, and bio-metric identification for security, etc[10].

Among all these research topics applying in A.I., Human Computer Interaction (HCI) is one of the most important element in its structure[13]. HCI technology can provide the platform that makes the machine understand human behaviours and human being can operate the machine in a natural ways such as voice, hand gesture, facial expression, etc. Among all these human behaviours in HCI area, like in facial expression recognition, they can be simply divided into two kinds, dynamic and statistic[14].

Lots of studies [15] [16] [17] [18] have already shown the effectiveness of dynamic information in HCI area. As there are too many kinds of dynamic human behaviours in the area of A.I. utilised HCI, it would be impossible to research all of them in one thesis. The human behaviour would be mentioned in this thesis are including body movements, facial expressions, gestures, etc.

One of the most important human behaviour is body language, it is not the only method we human beings used to communicate with each other, but also a very effective way to communicate with other animals [19]. As in the research of touch gesture recognition[20], the touch gesture is a very effective method for us to communicate with many different kinds of animals. Substantial research[21] has investigated the creation of robot animals that can be perfect alternatives for companion animals and work dogs. There are already some mature products in the market utilised touch gestures to achieve communication between human and robot animals like [22].

longsighted by [23], most of human daily communications are expression by different kinds of body language, most of which we do not even notice. For example, when we see a person walking haltingly, we will know she/he is an old person even if we are behind him/her. There are also some kinds of body languages that we do not notice ourselves, like some specific way of moving or using our arms and legs. Some actions may be habits from childhood, while others may be caused by disease. For example, because of joint wear over time, older people may have difficulty moving around, and they may not notice the changes themselves. The recognition of chronic pain will help cure such diseases.

Even in communication by languages, research has shown that facial expressions play an important part [24] [25]. When we say the same word with different facial expressions, the meaning may be totally opposite. This is also one of the most important parts of HCI. It is one of the hottest topics in the research area for robots to recognise and imitate human facial expressions. The application for Facial Expression Recognition (FER) has been explored for many years. Most applications are in the entertainment field. Interactive movies [26] are one of the most excellent ideas. There are also lots of applications for games and in many other areas.

1.2 Motivation

The core issues of A.I. include the ability to structure, know, plan, learn, communicate, perceive, move and manipulate objects that resemble or even transcend people. One classifies of A.I. [27] divides A.I. to weak A.I. and strong A.I. Weak A.I. is focused on one narrow task which are most currently A.I. systems. Strong A.I. refers to a machine that have flexible and skillful behaviours like human beings or have awareness of ideas and self-awareness. Strong A.I. is still the long-term goal of the field. At present, strong A.I. has already achieved initial results, and even in some aspects of video identification, language analysis, board games, etc., the ability to surpass humans has been achieved. The versatility of A.I. indicates that all these problems can be solved. The same A.I. program can directly use existing A.I. to complete the task without re-developing the algorithm, which is the same as human processing ability. However, it takes time to achieve integrated A.I. with thinking ability. The more popular methods include statistical methods to calculate intelligence and the traditional meaning of A.I. Many tools apply A.I., including search and mathematical optimisation and logical deduction. Bionics, cognitive psychology and algorithms based on probability and economics are also being explored.

Among all these effort on weak A.I. to achieve a glance of strong A.I., this thesis try to make some negligible breakthrough on human behaviours recognition based on sequential data. The human behaviours mentioned here include touch gestures, micro gesture, body movements,3D facial expressions and real-time facial expressions.

In touch gestures recognition, the exist state-of-the-art is more focused on one single dataset and doesn't have versatility. Like [28] has the best performance in CoST dataset but poor in HAART dataset; [29] has the best performance in HAART dataset but not suitable for CoST dataset. The research gap of this thesis is to propose a general method have good or even best performance in both dataset.

In micro-gesture recognition, the current method is not accuracy enough.This thesis proposed a better method based on the modification of base-line and achieved a best recognition accuracy.

In lower back chronic pain related body movement, the current works are already having a high recognition accuracy. But the recognition based on detecting the body movement for a long period of time and only recognise one result, have pain related behaviour or not. It could not recognise when the pain related behaviour starts and end. The detection of details including the starting and ending of pain related behaviour would be helpful for the research of pain related body movement.

In 3D facial expression recognition, the exists methods are focused on using powerful deep learning tools analysis whole 3D facial data to achieve high accuracy. It would be resource expense and low effective. The research gap in this thesis is to divide the raw facial data into dynamic part and statistic part to achieve the goal of saving calculation resources.

In real-time facial expression methods, the gap of research is to find a balance solution on recognition accuracy and the speed of the whole system to achieve the goal of real-time.

1.3 Research question

To recognise different human behaviours, many different kinds of A.I. technologies, pattern recognition and machine learning methods would be applied. Here A.I. technology refers to machines have high ability on solving single mission, some even better than human beings[11]. According to recognise these different kinds of human behaviours, these behaviours should be translated to some information that computers or machines can recognise, which is called pattern[30]. The mathematical models or neural networks utilised to achieve the pattern recognition are called machine learning[30].

Using machine learning to achieve human behaviour recognition, there are several research questions should be solved. There should be a balance among computing time, performance and calculation resources expense. Normally, the performance has a positive correlation with computing time and calculation resources. When aims to increase recognition accuracy, like in 3D facial expression recognition, complex deep learning networks would be applied; when aims to reduce the computing time, simply CNN network would be utilised like in real-time facial expression recognition.

1.4 Aim and objectives

The aim of the research is to develop new systems for automatic recognition of human behaviours, especially from human gestures and facial expressions. Depending on the different tasks, the recognition systems will be slightly different, but the basic construction is similar. The basic method for recognizing human behaviour is a combination of feature extraction and machine learning. In order to carry out the research and compare it with other researchers in the world, different human behaviour public datasets will be used. Especially, due to the complexity of the human behaviour, the research here is restricted on sequential recording data where the whole process of the behaviour is fully recorded. In these data, the dynamic information is the key for exploring.

In order to achieve the above aim, the objectives are set as the followings:

- Gesture recognition identifies human gestures through mathematical algorithms and machine learning methods to recognize gestures with computers. Gesture recognition can come from the movement of various parts of the body, but generally refers to the movement of the body and hands. Gesture recognition can be considered a means of letting the computer understand the language of the human body. Therefore, HCI is not only the text interface or the users image interface, but also is controlled by the mouse and keyboard. Among all the HCI methods, gesture recognition is more accurate and stable.

In touch gesture recognition, the objective is to develop a system to recognize touch gesture with better accuracy. The system includes multiple features extraction, high accuracy machine learning, feature selection and fusion. The comparison of systems with and without these pre-processing and post-processing features will explore the necessity of these methods. Also, the system will consider generation because there are two datasets in the Social Touch Database. A more generic system will be more practical in more situations of touch gesture recognition.

- In micro-gesture recognition, the objective is to recognize holoscopic 3D micro-gestures in a video-based dataset and an image-based dataset. Comparing both angles of recognition will develop a better method of holoscopic 3D micro-gestures recognition.

- In body movement recognition, the dataset of body movement and Electromyography (EMG) is collected from 22 participants with widely varying ages and of both genders. Each participant did different exercises several times, including one leg stand, sitting still, reaching forward, sitting to standing, standing to sitting, bending and walking. To recognize and detect Chronic Lower Back Pain (CLBP) continuously is the first step of pain management. The automatic detection of CLBP can provide long-term, continuous detection of chronic pain and supplement medical treatment. This is a research area strongly related to emotion detection and affective computing. Behaviours indicating chronic pain include guarding, hesitation, bracing, abrupt action, limping and rubbing. Any of these can indicate the existence of chronic pain.

- FER has been developed for many years. Comparing with traditional methods like SVM, k-NN, the recognition accuracy on FER using deep learning methods are much better, such as Convolution Neural Network (CNN) and Deep Neural Networks (DNN). The resource of calculation for deep learning is huge, so it is necessary to find a way to use deep learning methods more economically.

In 3D FER, dynamic extraction process will be discovered before deep learning. It will take lots of time for deep learning to act directly on 3D data, but using deep learning methods on texture maps and depth maps separately saves lots of time. To increase the accuracy of recognition on both maps, a latest Tensor Compressive Measurements (BSCM) method will be considered.

- In real-time FER on interactive movies, the objective is to build a practical real-time FER system with acceptable accuracy and speed. A simple CNN has been applied, and many practical modifications have been utilized. The objective is to make a usable interactive movie system and exhibit it in public. The practical entertainment system also can prove the effectiveness of using FER in real life.

1.5 Thesis contribution

The contribution of this thesis is mainly about building different automatic recognition systems to detect or recognize human behaviours. Those human behaviours have some common points in that they are all dynamic behaviours and they all also have some differences.

- An automatic touch gesture recognition system has been proposed, including pre-processing, multiple feature extraction, feature selection, pattern recognition and fusion. Both statistical and video features were extracted, including Motion Statistical Distribution (MSD), Spatial Multi-scale Motion History Histogram (SMMHH), Binary Motion History (BMH), Statistical Distribution (SD) and LBPTOP. Two powerful machine learning methods, Random Forest (RF) and multiboosting, have been utilized. A Sobel edge detection is utilized as pre-processing, and a Minimum Redundancy and Maximum Relevance (mRMR) feature selection is used to reduce the dimension of features after feature extraction. A decision-level fusion method Hierarchical Committee (HC) has been used as a post-processing tool to combine all the predictions.

- An automatic 3D micro-gesture recognition system has been proposed and tested on a Holoscopic Micro 3D Gesture (HoMG) dataset for which a holoscopic 3D video was recorded and annotated. A new system including frame selection by score has been proposed on the video-based dataset. Video-based recognition used Local Binary Patterns from Three Orthogonal Planes (LBPTOP) and the Local Phase Quantisation from Three Orthogonal Planes (LPQTOP) as feature selection and Support Vector Machine (SVM) as machine learning. Then, a non-linear SVM has been applied and the recognition accuracy is improved comparing with baseline and state-of-the-art.

- An automatic body gesture recognition system has been proposed to help older people with CLBP. The proposed system can recognize the behaviours of CLBP patients, like abrupt actions and guarding. A new two-stage machine learning method has been proposed that combines k-nearest neighbour (k-NN) and Hidden Markov

Model (HMM) and achieves a better recognition performance of body gestures than traditional methods.

- A 3D feature expression recognition (FER) system has been proposed to achieve better performance on the most popular posed face 3D FER dataset, Binghamton University 4D Facial Expression (BU-4DFE). A latest Background Subtraction method was applied based on tensor for pre-processing to extract the dynamic information on the face. This is the first utilized Tensor BSCM on FER. A deep learning method (e.g. Dynamic Image Net) is used on the dynamic facial information. A comparison between with and without Background Subtraction is made to prove the effectiveness of this method.

- A real-time FER system was built for an interactive film system. In the proposed automatic emotion recognition system, CNN has been used as feature extraction, and SVM is used for classification. The real-time system has been prototyped and exhibited on several occasions. In the final system, several practical problems have been considered, including brightness control and the right face selection from all of the audience. That has improved the accuracy of the practical application.

1.6 Thesis overview

This thesis has eight chapters. The first chapter, the introduction, which is this chapter, includes the background, research questions, motivations, aims and objectives of my research.

The second chapter is the literature review.The general human behaviour analysis methods were reviewed and the existing human behaviour recognition systems were addressed in term of feature extraction, feature selection, classification and fusion.

In chapter 3, an automatic social touch gesture recognition system is proposed that contains edge detection, multi-feature extraction, feature selection, classifier and decision-level fusion 5 sections. The Social Touch dataset has two parts, the Human-Animal Affective Robot Touch (HAART) dataset that contains 7 different touch gestures and the Corpus of Social Touch (CoST) dataset that contains 14 gestures. The proposed system achieves best results on both datasets. Also, it is not a deep learning method and saves calculation resources.

In chapter 4, an automatic micro-gesture recognition system is proposed. The basic construction of this system is similar to an automatic social touch gesture recognition system. The differences include using LPQTOP feature extraction and non-linear SVM classifier. The proposed system considers non-linear characteristic of features and achieves the best accuracy among all the publications.

In chapter 5, an automatic body gesture recognition system is proposed. This system can detect unnatural body movements caused by CLBP frame by frame. The advantage of detecting chronic pain frame by frame is that when utilized in real life, the system can be easily transformed into a real-time automatic CLBP detection application. Also, this system verified the performance of a Two-Stage Classifier (TSC) method by combining k-NN and HMM.

In chapter 6, an FER system on 3D dynamic dataset is proposed. The dataset utilized is the most popular posed 3D facial dataset, BU-4DFE. Already, some very accurate results have been achieved with this dataset using deep learning methods. The proposed system in this chapter is not the best in terms of performance, but it verified that BSCM, which has been used widely on video processing, can be used in facial recognition and can improve performance.

In chapter 7, a real-time FER application using deep learning is proposed. This application is installed in a real-life entertainment system. The application contains several pre-processing modules for use in real life, including face detection, face selection and brightness controlling. The proposed system is a good example of academic technology combined with industry application.

Finally, the last chapter is the conclusion and future works.

Chapter 2

Literature Review

2.1 Human behaviour analysis

Human behaviour is the response of human being opposite the stimuli from environment or internal bodies. Including all the human-related body movements and observable emotions. With the development of A.I. technology, human behaviours are important not only in human social communication, but also in intelligent HCI. HCI is defined as a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them by the Association for Computing Machinery. It involves study, planning, design and uses of the interaction between users and computers. The aim of HCI is to improve interactions between users and computers by making computers more usable and receptive to users' needs. A long-term goal of HCI is to design systems that minimise the barrier between humans' mental model of what they want to accomplish and the computer's support of the users' tasks. In HCI studies, the most commonly used human behaviours include voice, gesture, facial expression and some physiological information like heart rate, body temperature and Electroencephalography (EEG)[31]. Surpassing ordinary HCI, Intelligent HCI also seeks to explicate the mechanisms of human perception, cognition and action.

In this research, I not only target physical human behaviour like different type of gestures, but also advanced human behaviour like human emotion from facial expression. Gestures are the most commonly used human behaviours when humans communicate with each other. There are studies that show that most of our communication information is given by gestures, including body gestures and hand gestures. Facial expression is the most used modality to capture human emotion. These two are the most used ones for intelligent HCI.

2.1.1 Human gesture behaviour

In the area of computer science, gesture recognition is a topic using mathematics algorithms to recognise human gestures. It could to applies to recognise several different kinds of gestures including touch gestures and micro-gestures which would be introduced later[32].

Many methods has been proposed to detect gestures, like building 3-D models or detecting the track of hands or body movement. According to different kinds of data input, gesture recognition modelling could be divided into 3 types approximately: 3D model-based algorithms, skeletal-based algorithms and appearance-based models. Among these 3 types of modelling, 3D model-based algorithms and appearance-based models are most common used as it would be easy to achieve dynamic tracking of hand gestures and hand gestures has been most important part of traditional gesture recognition [33].

In this thesis, three types of human gestures will be investigated: 1). Touch gesture that records human finger movement; 2). Hand micro-gesture that is used to control small devices like watches in the Augmented reality (AR) and Virtual reality (VR) immersion experience; 3). Body gesture that is normally used in the analysis in sports or health care.

Touch gesture

Touch gesture refers to how human hands response and interactive on a directly contact surface[34]. My research of social touch gesture started with the Social Touch Challenge 2015 [35]. Touch gesture is a gesture touch on a surface, usually an entity screen, using more than one finger. Sometimes touch gesture would also include the touch of palm. Touch gesture on controls often have some predefined motions such as rotating [36].

Touch gesture has been widely used in controlling mobile interface as the development of smartphones. The earlier screens are only support one touch point [37], which can hardly support touch gesture technology in these system. The most common used touch gesture technology on smartphones is Multi-touch, which was first introduced by stumpe [38] and make touch gesture control possible.

Besides physical touch on entity screens, touch gesture can also been used on visual surface. Like in optical touch technology, the cameras or sensors would capture the scatter or reflection of light caused by touch gestures. There would not be any entity touch screens or walls exists [39].

As the development of touch gesture technology, touch gesture has been used in many HCI systems not limited on smartphones or tablets controlling system. In Kanevsky et al. [40], a touch gesture based interface for motor vehicle has been proposed to control motor vehicle. In Zhou et al. [41], a wearable device with visual touch gesture controller has been proposed. In Knight et al. [42], social touch gesture is used in sensate robot to achieve interactive with kids as a visual pet.

Hand micro-gesture

Micro-gesture refers to using some defined micro movement on hands to achieve some specified objective like controlling or recognition[43]. Because of convenience and effectiveness, hand micro-gestures have been used in many control systems. Many controllers utilizing hand micro-gesture have been developed. For example, virtual keyboard used micro-flick [44], Nod used micro-tap gesture and Soli used micro-slide gesture [45].

In [46], hand micro-gesture was utilized in wearable module, fusion with environmental module and in a control system in a vehicle. In [47], the control system utilized hand micro-gestures in the car combined with speech and eye detection. Many application have used micro-gestures in a car, like [48] and [49].

All the literature suggests that one big use of micro-gesture is in cars because it is inconvenient for large scale movements. The other usage of micro-gesture is in wearable devices, as in [4]. Lots of experiments have been based on this dataset, like [50] and [51].

Body gesture

Body gesture is one kind of para-language refers to a way of using body movements or movements to replace or assist in the communication of voice, verbal or other communication. Body gesture is an important communication of human beings, including the movement whole body or any part of it. The science used to understand and explain body gestures also been called kinesics [52] and first being introduced in [53]. Body gesture doesn't have a strict grammar or an absolute meaning like sigh language, the explanation of it should be interpreted broadly.

On the other hand, the usage of body gesture has been increased as the development of HCI technology. Some wearable equipment has been developed like [54] and [55] using body gesture as a controller. A smart TV system use body gesture controller has been proposed in [56].

Besides the using as a controller, other research like [57],[58] and [59] showed that automatic body gesture recognition can be used to treat CLBP. The detection of CLBP is achieved by recognizing unnatural body gestures, like guarding behaviour and hesitating. There are already some studies on detecting CLBP body gestures in [60].

2.1.2 Human emotion behaviour

According to some psychologists' studies, the multi-model judgement of human affect such as combinations of facial expression, sound events, body movement and gestures have advantages over using a single model like using visual or audio only. On the other hand, some studies have shown that facial expression is the most important among all the affective recognition methods [61] [62], although some linguists have different opinions about the importance of audio and vision.

Modelling of emotions seems to be a widely debated and interdisciplinary issue. Although many models exist, audio emotion recognition is mainly divided into discrete and continuous models [63]. In the discrete case, the emotions are divided by different verbal descriptions. The most common model is the basic emotion model and list of adjectives. Firstly, researchers introduced a set of primary emotions like happiness, sadness and anger, and then their combinations can change to other emotions [64]. The model has been challenged as compared to emotional preferences to learn physiological responses, such as being directly connected to neurological research in special areas of brain activity in which lie a particular set of basic emotions. Some researches show the connection of active edge or paralimbic belonging to a basic set of emotions [65].

The list of adjectives model group was set at a distinct emotional state, each corresponding to a set of synonyms for mood [66]. Many studies also propose to extend the list of adjectives model with some different emotional states [67]. On the other hand, the continuous emotion model is a result of a series of basic emotional feelings to consider, such as arousal and valence. These values are represented as an orthogonal coordinate system of continuous emotional states. Specific emotions in this coordinate system are represented as specific vectors. Although usually the number of axes is 2, which is a 2D representation or coordinate system, some researches have more dimensions and changed the coordinate system to a 3D coordinate [68]. Typical affective states used as axes in a 3D space are valence, arousal and dominance. However, the two-dimensional model can be considered one of the advantages of 3D space because it reduces complexity [69]. Through the cluster or qualitative dimension, representation is quantified by describing the relevant points arising from the use of verbal labels to describe the emotion value.

Words used to express different emotions or the same as the former are used as labels, which are the same as circle indicator points. The schematic of 2D Emotional Graph is shown in Figure 2.1

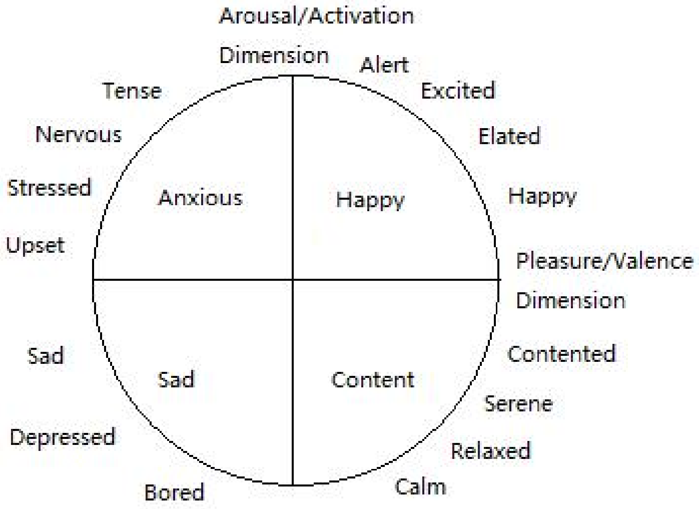


Figure 2.1: Schematic of 2D Emotional Graph [1]

The value of continuous space between the discrete space qualitatively describes the relationship between, but still have not found a quantitative relationship. In the continuous model, there is no emotional state to define the values that can be mapped to a discrete model of independent emotions. An exception is the case of 2D space into different quadrants. A continuous 2D model in a quadrant describes more than one discrete emotion, but no reports of an actual value can be used to quantify group arousal and emotional valence values, which are distinct and separate [1]. In addition, studies have shown that emotions had previously classified the 2D plane as uncertain, but the ambiguity of the situation can be resolved by continuous process, assuming that the value of each 2D graph is a distinct emotional state and not in a separate denomination of any text description [70].

Affective computing [71] is an interdisciplinary field concludes computer science, psychology, and cognitive science and aims to develop a system could recognise, interpret, process, and simulate the effects of human beings. Affective computing could be potentially applicable to multiple fields, including neuro-science, sociology, education, and psychology. One of the most important and common applications is HCI. Affective computing expands HCI by including emotional communication together with appropriate means of handling affective information. HCI could be improved by having computers naturally adapt to their users when communication about when, where, how, and how important it is to adapt involves emotional information, possibly including expressions of frustration, confusion, dislike, interest and more.

Recently, affective computing has become an important method in the research of psychology, cognitive science, computer science and many other areas. Also, affective computing is an important part of human computer communication. Computer design in the next generation should focus more on natural human expressions like facial expressions, sounds (including voice and sound events like coughs and laughs), gestures and body movement. There are two directions in affective computing area. One is detecting and recognising human emotions, and the other is designing computers innate or simulating emotions[72]. In this literature review, we focus on the former.

In the early works of affective computing, the most common ways for using it were training and testing on series of deliberately affective expressions rather than natural expressions. Thereafter, many works until recently focused on spontaneous expression of human emotions. In the early works, scientists recognised a series of basic expressions [73]: anger, disgust, fear, happiness, sadness and surprise. These emotions are present in Plutchik's Wheel of Emotions shown in Figure 2.2.

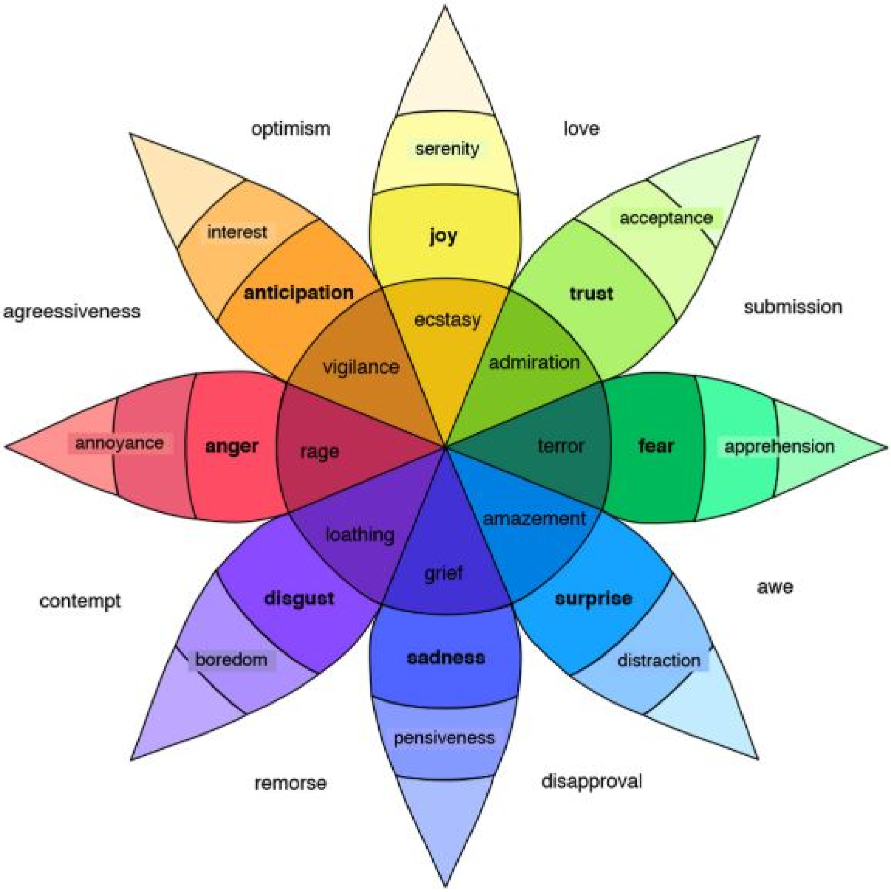


Figure 2.2: Plutchik's Wheel of Emotions [2]

Then, in the 1990s, the basic emotions were expanded to include another 11 emotions [64]: amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame. Some of these emotions are not expressed by facial muscles. But at the International Workshop on Emotional Representations and Modelling for HCI Systems [ERM4HCI] [61], Klaus R. Scherers presentation showed that strong emotions such as fear is rare in daily life and subtle emotions such as interest and regret are common. He also suggested promising ways to recognise subtle emotions in natural settings with facial cues and semantic rule structure. In the early works of affective computing, scientists used a single model for detecting emotion and recognition such as images and voice signals only. Recently, many works have also used multi-model recognition [74].

Ways to detect and recognise emotions include capturing speech facial expressions, body posture and gestures, heart rate and body temperature. Through analysing the characteristics of speech signals, such as speech patterns, vocal parameters and prosody features, one can recognise emotions. Analysing body gestures and physiological monitoring such as blood volume pulse, facial EEG and galvanic skin response is also a way to collect data to detect emotions.

For classifying different emotions in processing speech data quickly and accurately, a reliable data base or vector space model should be built, such as SVM, k-NN and HMM. Choosing a classifier is important because the appropriate classifier can significantly enhance the overall performance of the whole system.

Representing the sequence of speech feature vectors allows the deduction of states sequences through which the model progresses. The states can consist of various intermediate steps in the expression of an emotion, and each has a probability distribution over the possible output vectors. The states sequences allow us to predict the affective state that we want to classify. This is one of the most commonly used techniques within the area of speech affect detection.

Various methods such as HMM, neural network processing, optical flow or Active Appearance Model (AAM) are also used in combination or fused to detect and process facial expression. In 1972, Paul Ekman [73] identified six basic emotions: anger, disgust, fear, happiness, sadness and surprise. In the 1990s, Ekman [75] expanded his list of basic emotions, adding amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame. Speech and face recognition are also used in conjunction with gestures to detect a particular emotional state more efficiently.

Studies of vocal affective recognition [76] are also identified mainly by the impact of the basic theory of emotion. In turn, most of the efforts in this direction aimed at recognition of basic emotions from a subset of the voice signal. However, some recent studies have examined speech signal in terms of certain application-dependent affective states. There is study [76] has investigated detecting fraud, certainty and stress, confidence, confused and frustrated voice-based analysis, designed to detect trouble, annoyance and frustration. Some explored voice recognition-based detection performed in empathy.

Emotional identification from audio can be seen in the need to train or ground truth dataset to indicate emotional A.I. or pattern recognition tasks. The annotations include emotional patterns in close coordination with the selected emotion label or class. Signal technology featuring the same dataset is used to extract the appropriate choice and value and feature extraction of emotional comments, while feedback to the machine learning algorithms end up with a classification model. Technical characteristics of comments and the final set of tests are used to evaluate the development model to provide an assessment of the classification accuracy [77].

There are many different methods could achieve the goal of emotion recognition, including speech recognition, FER and gesture recognition. Especially combined these methods together, the performance of emotion recognition often could be improved. Gesture recognition concludes many different kinds, from simple kinds like conditioned reflex to complex kinds like sign Language. Many gestures could contain emotion information, like waving arms in exciting situation would be different from waving arms to alarm other people. So recognise these different emotions contain in different gestures and response them could be an important progress in HCI[78].

Emotion from 3D facial expression

Facial expressions are the result of one or more actions or states of facial muscles. These movements express the emotional state of the individual to the observer. Facial expressions are also a form of nonverbal communication. 3D facial expression is collected 3D data utilise to describe facial expressions[79].

In facial expression detection and recognition, one of the most important and commonly used methods is the FACS [75], which taxonomies expression of face based on a system developed by Carl-Herman Hjortsjo [80]. FACS can code nearly all the possible facial expressions in the human face. It can deconstruct a facial expression into several Action Units (AU) that are independent of any interpretation, such as recognizing the basic emotions according to Emotional FACS (EMFACS), various affective states according to FACS Affect Interpretation Database (FACSAID) and other psychological states like depression and pain [79].

FACS uses muscles to define AUs, as when they are relaxed or shrunk. It also defines some Action Descriptors, which are used to distinguish specific behaviours exactly same as they have for the AUs [81]. For example, FACS can distinguish the insincere smile and voluntary Pan-Am smile by tracking the movement of zygomatic. Also, the definition of involuntary Cheyenne smile and sincere are determined by the zygomatic is contracted on major part or just inferior part [81].

The advantage of 3D FER is more accuracy because the 3D facial data contain more details. The most popular datasets on 3D FER are BU-4DFE [5] and BP4D [82]. There are already lots of studies on these two datasets, some of which have achieved really good accuracy. In the development of A.I. technology, the most commonly used 3D FER systems are all deep learning methods.

The 3D FER dataset was introduced by Yin, who first proposed a facial expression label map (FELM) [83] that tracks a 3D facial model by a method called Three Points-based Pose Tracking [84]. After points tracking, the tracking models are fitted into the 3D face scan. Based on this method, the Binghamton University 3D Facial Expression (BU-3DFE) [85] dataset has been built. The same FER performance has been used on the BU-4DFE dataset. An AAM is also used in a 2D texture map to track feature points according to [86].

Real-time emotion analysis from facial expression

FER has been developed for years. Although 3D FER is more accurate than normal FER, the speed of 3D FER is very slow, and it cannot achieve real time currently. It would take several minutes to process one 3D model with deep learning methods. In that case, when there is the necessity of real time FER, 2D FER is the first to consider.

Posterity proposed encryption FER based domain uses the local fisher discriminant analysis achieved as good as ordinary image accuracy. Then, same subspace explained the same expression has been proposed. Through different emotions projected, new expressions can be generated from the same image[87].

Most of these methods use full face images. Other methods use features extracted from some specific face patches. The facial image is divided into a plurality of sub-regional and local features, and then Adaboost is used to improve the LBP histogram for classification [88].

In other studies, the face has been divided into 64 sub-regions to explore the most common facial expression active face patch and in particular the special facial expression active face patch. For multitasking sparse learning methods, they use some patches of facial features for FER [89]. Based on specific landmark, some studies used eight facial patches to show the change of skin. However, some patches are not included in the expression recognition of mouth texture, which is important. In addition, hair on the forehead occlusion will lead to an error [90]. Some researchers also extracted different scales of Gabor features from face images and used Adaboost to train and select salient features of each facial expression. But the location and size of different databases' trained prominent patches are different, so a unique standard cannot be used to define the location of the image expression [91].

The affective state can also be recognized by speech, both through explicit (linguistic) and implicit (paralinguistic). Linguistic messages (like the words we use) are rather unreliable means of analysing affective behaviours and are difficult to generalize from one language to another [92]. Paralinguistic message researchers have not identified an optimal set of voice cues that reliably discriminate among emotions. But listeners seem to be accurate in decoding some basic emotions from prosody [93] and some non-basic affective states like distress, anxiety, boredom and sexual interest from non-linguistic vocalizations like laughs, cries, sighs and yawns [61].

Many studies show the correlation between some affective displays (especially prototypical emotions) [94] and specific audio and visual signals [75]. The human judgement agreement is typically higher for facial expression modality than for vocal expression modality. However, the level of agreement drops considerably when the stimuli are spontaneously displayed expressions of affective behaviour rather than posed exaggerated displays. In addition, facial expression and the vocal expression of emotion are often studied separately [61].

Based on previous studies (temporal dynamics of facial behaviour represent a critical factor for distinction between spontaneous and posed facial behaviour and for categorization of complex behaviours like pain, shame and amusement [75]), we may expect that the temporal dynamics of each modality (facial and vocal [61]) and the temporal correlations between the two modalities play an important role in the interpretation of human naturalistic audio-visual affective behaviour. These issues are waiting to be researched. Another important area is human behavioural signals that are context-dependent (a smile can be a display of politeness, irony, joy or greeting; what is important is who the receiver is and who the expresser is [61]).

2.2 Sequential data

For both gestures and facial expressions, the most effective data are sequential data. Sequential data is a dataset remain an order of continuously changed sequence, usually is a video or a series of pictures that change continuously. Sequential data includes dynamic information of human behaviour, which is very important for automatic system building. A typical example of sequential data is video data. Video datasets are very common in FER areas, and some datasets of gestures also have a similar structure to videos.

A most popular published touch gesture is Social Touch dataset [35]. The data is collected by a conductive fur sensor, which has 8 × 8 sensors under the fur and can record the touch intensity in real-time. Each gesture would last for several seconds. The collected data is a dynamic change physical signal like some samples shown in Figure 2.3.

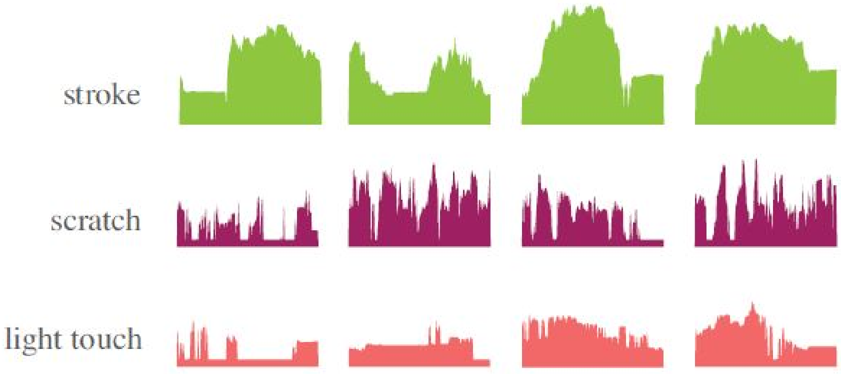


Figure 2.3: Samples of touch gestures in 2 seconds [3]

The other dataset using in this thesis is HoMG dataset [4]. The micro-gestures are collected by a Holoscopic 3D (H3D) camera. The whole system is shown in Figure 2.4. The HoMG dataset is a 3D video dataset. Each gesture is presented by a high solution video. The gestures are acted in the black frame of the system, and the camera will collect the videos in two different distances.

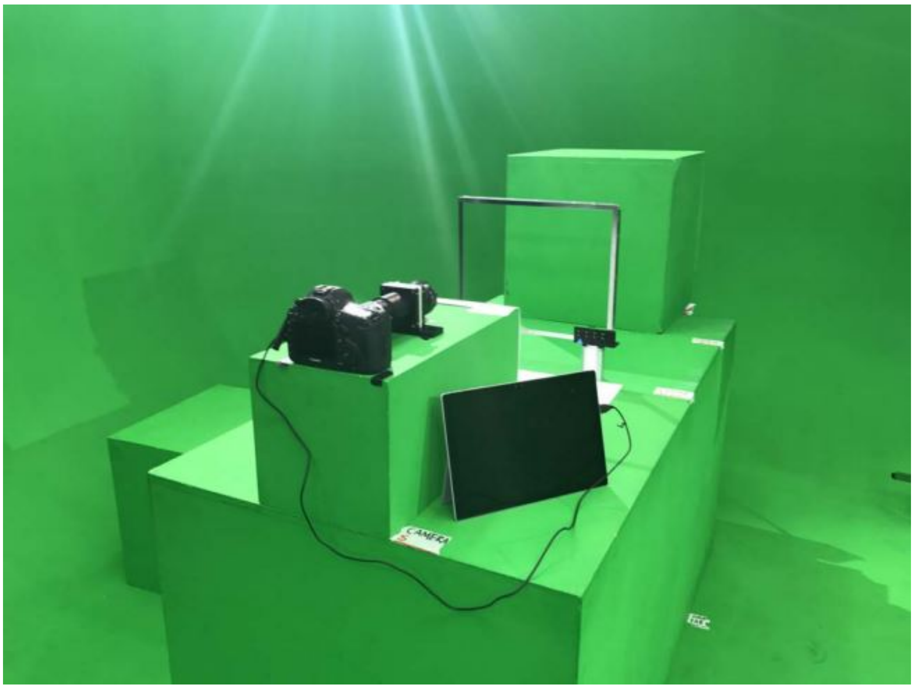


Figure 2.4: HoMG data collection system [4]

Computer-based facial recognition research has evolved since the first Automatic Face and Gesture Recognition in 1995 [95]. Although there have been a lot of successes in 2D databases and 3D image data, when dealing with low visibility appearance, large head rotation, subtle skin movements and changes in illumination in different postures, the performance will be noticeably decline.

Due to the limitation of facial surface deformation when evaluating three-dimensional features in two-dimensional space, two-dimensional images with few features may not accurately reflect real facial expressions (such as three-dimensional head posture, three-dimensional wrinkles and deep motions of skin extrusion, and on the cheeks, the area of the forehead, eyebrows, decree and crow's feet).

The face is a 3D object, and the expression of facial emotions needs to be realized by 3D information such as the transformation of the depth and the rotation of the head. Another major limitation of existing databases is that there are only deductive and posed expressions that differ in time, complexity and intensity of expression. There are currently no datasets available that contain dense, dynamic 3D facial representations of spontaneous facial expressions based on Facial Action Coding System (FACS) [75] annotations.

BU-4DFE [5] contains 6 different posed expressions: anger, disgust, fear, happiness, sadness and surprise. The database collects 101 dynamic 3D data of Asians, blacks, Latinos and whites aged 18 to 45. The BU-4DFE has a total of 606 videos, each of which is about 4 seconds long, with a video resolution of 1040×1392. The video shows the subject's head and neck. A sample of BU-4DFE is shown in Figure 2.5.

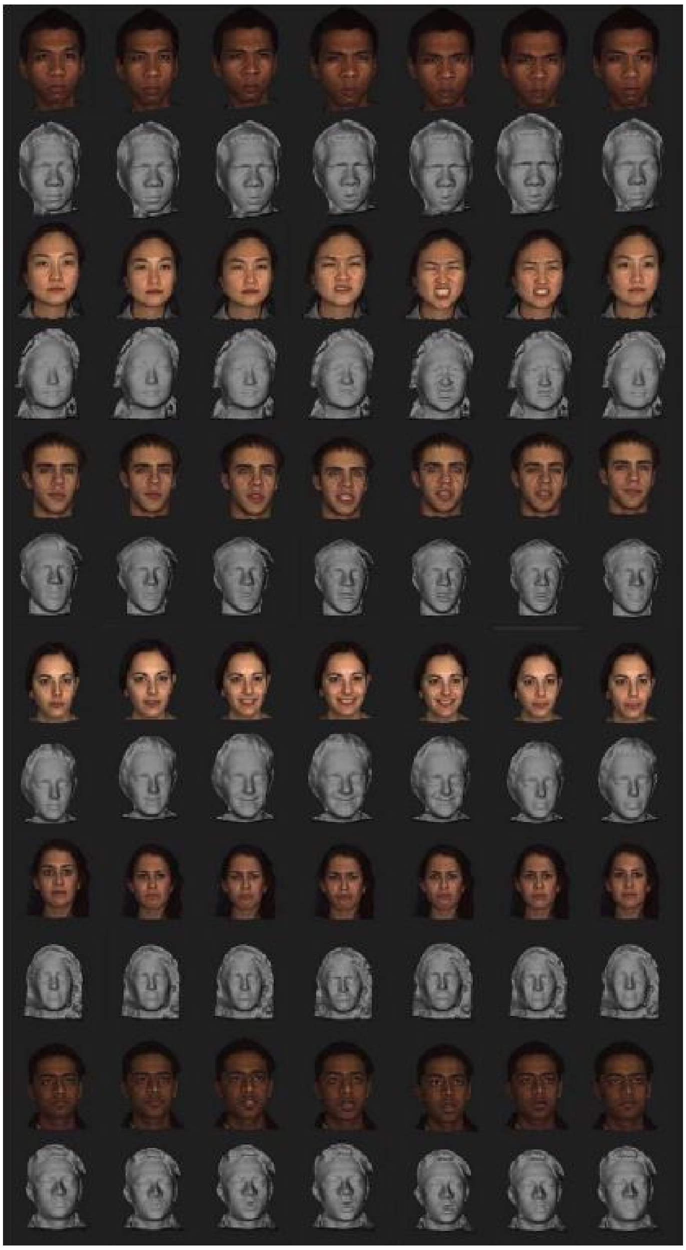


Figure 2.5: BU-4DFE video samples in texture map and depth map [5]

2.3 Automatic recognition system

A general automatic recognition system consists of 4 main components: signal pre-processing, feature extraction, pattern recognition and post-processing or fusion.

2.3.1 Pre-processing methods

Pre-processing refers to some methodology utilise before machine learning and feature extraction processing aims to extract more effective information or reduce noise from dataset. Like dimensional reduction, feature selection, sample optimisation, class balancing, image compression, etc.

Sobel edge detection

Sobel edge detection is an edge detection algorithm widely used in image processing and computer vision [96]. Technically, it is a discrete difference operator for the operation gradient approximation of image brightness function. At any point, this image operator will have its corresponding gradient vector or vector method. Sobel edge algorithm includes two kernel matrices convolved with the original image to calculate approximations of the derivatives horizontal and vertical, respectively. As the equation below shows, here the * represented the 2-dimensional signal processing convolution operation.

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A \qquad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \tag{2.1}$$

In equation 2.1, G_x and G_y represents horizontal and vertical edge detection separately. Horizontal and vertical gradient of each pixel of the image using the equation 2.2 approximation formulas combined to calculate the magnitude of the gradient G.

$$G = \sqrt{G_x^2 + G_y^2} \tag{2.2}$$

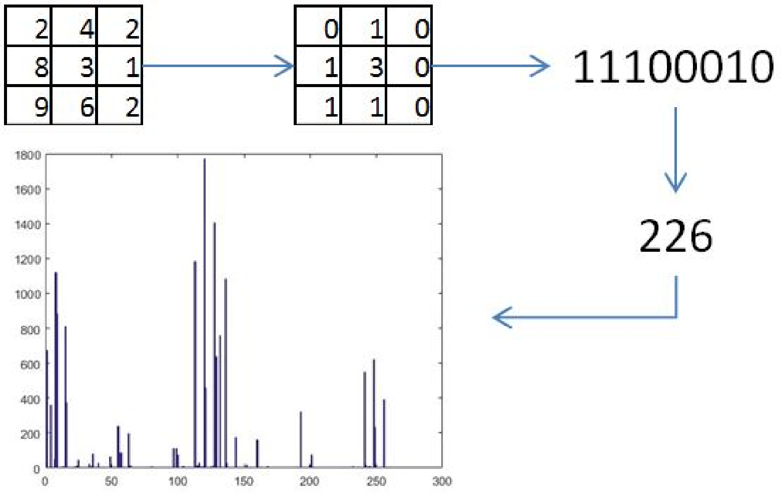
2.3.2 Feature extraction

Feature extraction refers to constructs information-rich and non-redundant derived values from an original dataset.It can help with the learning process and the steps of induction, and in some cases makes it easier for people to better interpret the data. The aim of feature extraction is to reduce the original dataset to a more manageable ethnic groups (features) for learning while maintaining the accuracy and completeness of the original dataset[97].

Local Binary Pattern (LBP)

In automatic FER, one essential step is face alignment, usually done by detecting the horizontal position of the eyes. The next step is feature extraction, including the very commonly used LBP and some special methods like MHH [98]. The features selected affect the accuracy of classification. Facial landmark detection is followed by feature extraction, like using an active Infra-Red illumination along with Kalman filtering to track facial components [99]. Using both geometric and appearance features can also improve performance. It is not convenient to determine the position of the eyes to figure out the initial positions of the facial landmark. It is also feasible for recognizing face AUs of the lower face by using the relative distance like eye, brow and lip corner and transient features like wrinkles and furrows [100].

In the simplest case, a simplified Local Binary Pattern feature vector can be calculated as follows. The detection window is cut into blocks (cells, for example, each block of 16x16 pixels). Each pixel in the block is compared with its eight neighbouring pixels (top left, middle left, bottom left, top right, etc.) in either clockwise or counter-clockwise order. The central pixel that is greater than a certain neighbourhood is set to 1; otherwise, it is set to 0. It won an 8-bit binary number (usually converted to decimals), a feature of that location. A histogram for each block is computed. At this point, one can choose a histogram normalized of all blocks of the histogram series that have been a feature of the current vector detection window. The schematic of LBP is shown in Figure 2.6.



2	4	2
8	3	1
9	6	2

0	1	0
1	3	0
1	1	0

Figure 2.6: Schematic of LBP

Because of the uncertainty of the instability threshold under different lights, it is difficult to determine the presence of the groove by Canny edge detector, which detects the edge by 5 steps. First, use a Gaussian filter to remove the noise and smooth the image. The larger the size of the kernel, the lower the detectors sensitivity to noise, so the localization error to detect the edge will slightly increase.

Local Phase Quantisation (LPQ)

The LQP [101] characterizes texture or appearance by using the differences in sign, magnitude and orientation. The proposed LPQ algorithm conclude two stages, learning satge and inference stage. The LPQ algorithm contains 4 steps: 1. extract local sign information by , magnitude information and orientation patterns by using orientation estimation and quantification; 2. using vector quantisation to learn an S, M and O separate codebooks; 3. utilizing lookup table (LUT) to map sign, magnitude and orientation patterns into the corresponding codebook; 4. combine 3 different histograms and generate the feature vector.

Motion History Histogram (MHH)

MHH is a feature extraction method for video based on well-known motion history image (MHI) [78]. MHI is a feature section that selects the pixels that have changed in the video. For a $M \times N$ video, there is a motion mask $D(m,n,:)$. If the movement does not occur in this frame, k, the MHI maps the $D(m,n,k)$ as 0. If the motion has happened, the motion mask is 1. Then we will get a $M \times N$ picture, using 0 and 1 to represent whether the pixel of the video has changed or not.

Unlike MHI, MHH uses a P_i to represent the pixels in the sequence of one pixel in the motion mask sequence $D(m,n)$. When the 1 in MHI is connected, the P_i shows that i of 1s are connected. Combining all the P_i together for each pattern, we can draw a grayscale image, and each image is a histogram. The MHH decomposes MHI into different parts. For the range of i, the larger the i is, the better the motion information that is extracted [102]. P_i is represented in equation 2.3.

$$\begin{aligned} P_1 &= 010 \\ P_2 &= 0110 \\ P_3 &= 01110 \\ &\dots \\ P_M &= 01\dots10, \quad which \quad has \quad M \quad of \quad 1 \end{aligned} \tag{2.3}$$

$$C_{I,k} = b_I, b_{I+1}, \ldots, b_k (1 \leq I < k \leq N) \tag{2.4}$$

$$MHH_{m,n,i} = \sum(I,k)\chi\{C_{I,k} \in A\{D(m,m,:)\}\}(1 \leq I < k \leq N, 1 \leq i \leq M) \tag{2.5}$$

In equation 2.4, $C_{I,k}$ represent a sequence of all sub-sequence in $D(m,n)$ where I and k represent the start frame and end frame. So the MHH can be represent as equation 2.5.

Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT)

FFT is an algorithm that calculates DFT sequence or an inverse transform algorithm [103]. Fourier analysis of the signal is converted from the original domain (usually time or space) to indicate the frequency domain or the reverse over the conversion. Through FFT, the DFT matrix can be calculated into sparse (mostly zero) quickly to the product of factors such as transformation. DFT is defined by the equation 2.6, where X_k is frequency component and x_n is time or space component.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1. \tag{2.6}$$

The value of DFT directly based on the equation 2.6 will need to be calculated $O(N^2)$ times; there are N outputs for X_k, and each output needs N terms sum. An usually simplified theory is equation 2.7. Suppose an $M * N$ matrix S can be resolved to a multiply of a column vector and a row vector.

$$S = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix} \tag{2.7}$$

If there are M_0 distinct nontrivial values for $[a_1 a_2 \ldots a_m]^T$ and $a_m \neq \pm 2^k$,$a_m \neq \pm 2^k a_n$ where $m \neq n$. Matrix $b_1 b_2 \ldots b_m]$ has N_0 distinct nontrivial value. So there are $M_0 * N0$ amounts of multiples in S.

$$\begin{bmatrix} Z[1] \\ Z[2] \\ \vdots \\ Z[N] \end{bmatrix} = S \begin{bmatrix} X[1] \\ X[2] \\ \vdots \\ X[N] \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \ldots & b_m \end{bmatrix} \begin{bmatrix} X[1] \\ X[2] \\ \vdots \\ X[N] \end{bmatrix} \tag{2.8}$$

From equation 2.8 we can figure out $Z_a = b_1 X[1] + b_2 X[2] + \cdots + b_n X[N]$, and for each

Z, there is $Z[1] = a_1 Z_a, Z[2] = a_2 Z_a, \dots, Z[N] = a_m Z_a$. The simplified model can be transferred as equation 2.9.

$$S = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \dots & b_m \end{bmatrix} + S_1 \tag{2.9}$$

S1 is also an $M * N$ matrix, and if there are P_1 non-zero values in S_1, the maximum amount of multiples of S is $M_0 + N_0 + P_1$. Assume $N = P_1 \times P_2 \times \cdots \times P_k$, and among them, $P_1, P_2, \dots, P_k$ are co-prime for each other. The amount of multiples of P_k equals to B_k, and the DFT and the multiples of N are calculated as: $\frac{N}{P_1} B_1 + \frac{N}{P_2} B_2 + \cdots + \frac{N}{P_k} B_k$. Assume $N = P^c$ and P are prime numbers. If the amount of multiples of $N_1 = P^a$ is B_1, and among the $n_1 \times n_2 (n_1 = 0, 1, \dots, N_1 - 1, n_2 = 0, 1, \dots, N_2 - 1)$, there are D_1 values that are not multiples of $N/12$ and $N/8$, and there are D_2 values that are multiples of $N/12$ and $N/8$ but not multiples of $N/4$. So the amount of multiples for DFT in N is $N_2 B_1 + N_1 B_2 + 3D_1 + 3D_2$.

One of the most common algorithms of FFT is the Cooley-Tukey algorithm [104]. This method uses a divide and conquer strategy recursively length $N = N_1 N_2$ Discrete Fourier Transform (DFT) of length N_1 decomposition of N_2 a shorter sequence of DFT, and with $O(N)$ a complex multiplication for the twiddle factors. In the equations below, the W_N is used to represent $e^{-j\frac{2\pi}{N}}$. The W_N has characteristics including periodicity $W_N^{k+N} = W_N^k$ and symmetry $W_N^{k+\frac{N}{2}} = -W_N^k$ and assumes m is a divisor of N, $W_N^{mkn} = -W_{\frac{N}{m}}^{kn}$.

According to these characteristics, when calculating $y_k = \sum_{n=0}^{N-1} W_N^{kn} x_n$, the sum part can be divided into two as equation 2.10.

$$y_k = \sum_{n=2t} W_N^{kn} x_n + \sum_{n=2t+1} W_N^{kn} x_n = \sum_t W_{N/2}^{kt} x_{2t} + \sum_t W_{N/2}^{kt} x_{2t+1} = F_{even}(k) + W_N^k F_{odd}(k) \tag{2.10}$$

In equation 2.10 $F_{even}(k)$ and $F_{odd}(k)$ are two transforms about sequence $\{x_n\}_0^{N-1}$ in even and odd separately at $N/2$. According to these, only $N/2$ values in front of y_k have been calculated. The $N/2$ values that are behind can be calculated using the equation 2.11 and the equation 2.12.

$$y_{k+\frac{N}{2}} = F_{even}(k) - W_N^k F_{odd}(k) \tag{2.11}$$

$$y_k = F_{even}(k) + W_N^k F_{odd}(k) \tag{2.12}$$

DCT is a Fourier-related transform that is similar to the DFT, but using only real numbers [105]. It corresponds to a DCT length that is about twice its DFT. The DFT (because a real and even function in the Fourier transform is still a real even function of the dual function of a real) in some deformations requires the input or output position to be shifted by a half unit. The basic DCT is shown as the equation 2.13 [106].

$$f_m = \frac{1}{2}(x_0 + -1^m x_{n-1}) + \sum_{k=1}^{n-2} x_k cos[\frac{\pi}{n-1} mk] \tag{2.13}$$

The boundary conditions of the equation above include that x_n is even around $n = 0$ and even around $n = N - 1$, similarly for x_k. The most commonly used DCT defined for all possible N is the equation 2.14.

$$f_m = \sum_{k=0}^{n-1} x_k cos[\frac{\pi}{n} m(k + \frac{1}{2})] \tag{2.14}$$

At one time, it was thought that it would be more efficient to use a Discrete Hartley Transform (DHT) to process a pure real DFT, but then it was found that for a pure real DFT of the same number of points, the designed FFT can save more operations than DHT. The Bruun algorithm is the first algorithm to try to reduce the amount of DFT operations input from real numbers, but this method has not become popular.

For real input, when the input is even symmetric or odd symmetric, time and memory can be further saved. In this case, DFT can be replaced by DCT or discrete sine transform (discrete cosine/sine transforms). Since DCT/DST can also design an FFT algorithm, in this case, this method replaces the FFT algorithm for DFT design.

Most people who try to reduce or prove the lower bound of the FFT complexity focus on the complex data input because it is the simplest case. However, the FFT algorithm for complex data input has a great correlation with the FFT algorithm of real data input, DCT, DHT and other algorithms. Therefore, if any algorithm has any improvement in complexity, other algorithm complexity will be improved immediately.

2.3.3 Feature selection

Feature selection refers to a processing in statistics or machine learning aims to build mathematical model through select subsets in relevant features. Feature selection could simplify models, reduce training time and optimised generalisation[107].

Principal Component Analysis (PCA)

Because there are more and more new data instances, deformable models are becoming more popular in facial landmark detecting. The point distribution model (PDM) [108] represents the shape's geometry characteristic and inferred geometric variation from some statistical modes. It has become a standard in computer vision for the statistical study of shape [109] and for segmentation of medical images [110] in which shape priors utilized to interpreted of pixels/voxels that are noisy or low-contrasted. The latter point leads to active shape models (ASM) and AAM representing the object combine models of shape and texture, and providing results to ASM [111]. AAM is widely used [112], [113], [114] for facial landmarks with non-rigid detection and tracking. But its performance is poor in subject independent scenarios. In training dataset, placement the landmarks manually for construction of the shape model is a process both tediously and time-consuming. Constrained Local Model (CLM) framework has been proved a better tool for subject independent facial landmark detection [115].

PCA is a method that can analyse and simplify datasets.PCA is often used to reduce the dimensional of the features or dataset while keeping the dataset variance that contributed the most features. Principal components with higher order would be ignored and those with lower order would be kept. Such low-level components are often able to keep the aspects with important information of live data. However, this is not certain. Because the main component analysis relies on the given data, it greatly influences the accuracy for data analysis [116]. PCA and Linear Discriminant Analysis (LDA) [117] are used as tools for reducing the dimension and classifying different expression. There are reports showing that PCA-LDA fusion method can produce a better performance [118].

Minimum Redundancy Maximum Relevance (mRMR)

PCA is a common used feature reduction method, but as the progress of technique, the performance of PCA are no longer satisfied the requirement of feature reduction and selection. So more and more feature reduction and selection methods has been proposed. Similar with PCA, mRMR is an algorithm used in feature selection frequently, which could narrow down the relevant of features and accurate the identify characteristics of them.

In machine learning, the subsets of dataset have relevant with the parameters in pattern recognition normally identified as Maximum Relevance. The information or features in these subsets of dataset or features often have relevant and also have redundant. mRMR is desigh to remove the redundant in the subsets of dataset or features. In this way, mRMR has variety use in many cases of recognition.

There are many feature selection methods, mRMR used a method called maximum-relevance selection. Features have strongest correlation with the classification objects would be selected. To achieve the selection, many algorithm like floating selections, sequential forward or backward could be utilized. However, the features mutually far away could have high correlation with each other or classification objects. mRMR has better performance on these cases and been proved more powerful [119].

2.3.4 Classification methods

Classification methods are one important part of machine learning. Its goal is to determine which known sample class a new sample belongs based on certain characteristics of known samples. Classification is an example of supervised learning. Based on the samples provided by the known training set, the feature parameters are selected by calculation, and a discriminant function is created to classify the samples. In contrast, unsupervised learning, such as cluster analysis[120].

RF

RF is a machine learning method that includes many different decision trees. This method combines bootstrap aggregating and a random subspace method to build decision trees [121]. It is widely used in classification and regression. The basic steps of building each decision tree in RF machine learning method are as follows: considering that the whole number of samples is N and the number of feature vectors is M, for each node of the decision tree, the m feature vectors are used to define a decision results of this node on the decision tree; with a sampling with replacement in the N samples composing a train set, predict the deviation with the samples not composed; every node on the decision tree is determined by these random m feature vectors; and every decision tree is grown by these steps and utilized in the final classifier.

There are many advantages to using RF as a classification method: it is a rather highly accurate classifier in many kind of dataset; the various inputs can be very large, and it can evaluate the importance of each input feature vector; when building a RF, it can evaluate the generalized error with no deviation; it can also estimate the missing data, and if a large part of data is missing, it still can retain a high level of accuracy; it provides an experimental method to detect variable interactions; and it balances the errors in unbalanced datasets. In summary, RF is a valuable method in data mining, detecting outliers and visualization of data [122].

SVM

SVM is a supervised learning model for analysing data in classification and regression analysis and related learning algorithms [123]. With a given set of training instances, each training instance is marked as belonging to one or the other of the two categories. The SVM training algorithm creates a model that assigns the new instance to one of the two categories, making it a non-probabilistic two linear classifier. The SVM model is an example of representing an instance as a point in space so that the mapping makes separate instances of separate classes as broadly as possible. Then the new instances are mapped to the same space, and the categories are predicted based on where they fall on the interval. In addition to linear classification, SVM can also use the so-called nuclear techniques to effectively non-linearly and implicitly map its input into high-dimensional feature spaces.

Multiboost

Multiboost is short for Multi-class Adaboost. The algorithm is called Stage-wise Additive Modelling. The first step is giving each feature an average weight. For an n feature sample, $w_i = 1/n$. Multiboost circles for M times, at m time, find a classifier $T_m(x)$ to fit the feature weight w_i. According to equation 2.15, we find the error in m time represented as $Error_m$, which can be used to calculate the reset weight of each features. From equation 2.16, the reset weight of each feature w_i can be found by equation 2.17. The output of the Multiboost system is presented by equation 2.18.

$$Error_m = \frac{\sum_{i=1}^{n} w_i \prod [c_i \neq T_m(x_i)]}{\sum_{i=1}^{n} w_i} \tag{2.15}$$

$$\alpha_m = log\frac{1 - Error_m}{Error_m} + log(K-1) \tag{2.16}$$

$$w_i^m = w_i^{m-1} exp(\alpha_m \prod [c_i \neq T_m(x_i)]), i = 1,2,3,\ldots,n. \tag{2.17}$$

$$R(x) = \arg\max_k \sum_{m=1}^{M} \alpha_m \prod [T_m(x) = k] \tag{2.18}$$

The algorithm of Multiboost is similar to that of Adaboost, but there are some differences. The method for calculating the error rate is the same, but for Adaboost, the α_m only needs $1 - Error_m > 1/K$ to be positive. The Multiboost algorithm gives error classified data more weight and combines weak classifiers in a different way to fit a stage-wise additive model better.

k-Nearest Neighbours (k-NN)

In machine learning area, k-NN algorithm has been proved to be a reliable and stable classification method. When a feature training with k-NN, an N-dimension space would be established where the N is the length of the feature vector. Corresponding to all the samples in the dataset, each one has a label located in this N-dimension space. After the model has been built, when the k-NN model predicts a new dataset, often called a test dataset, k-NN would choose those points that have been trained that are nearest to this test data point. The number of the points that have been chosen is k. Among all the trained points, which label has appeared most would be identified as the class of the test data. In a two class classification situation, to avoid two different classes having the same votes, k usually would be set to an odd number. To make the boards of each class clearer, the smaller k has would be considered.

Comparing with RF mentioned above, although k-NN would be more reliable and stable, the performance in classification of it usually is not as good as that of RF. But when considering the time cost on classification, k-NN has advantage than RF.

2.3.5 Fusion

Fusion in machine learning refers to utilise mathematical models to process multiple classification or clustering results in one sample based on different features or machine learning methods into one single output. The aim of fusion is to combine information of different mathematical models or features together to get the best result[124].

The data fusion methods of affective recognition commonly used include feature-level, decision-level and model-level fusion. For feature-level fusion, because features usually come from different modalities and the time scales and metric levels are also different, the performance of recognition based on feature-level fusion will be affected. One of the common methods to decrease the influence is normalization, but as the dimensions of feature vector are increasing, the recognition performance is still being affected.

Unlike feature-level fusion, decision-level fusion models independently input from different modalities and combine these recognitions from single-modal. Decision fusion works well in conditions in which inputs are independent from each other or do not have any influence on each other. But when the inputs complement each other, like audio and visual features of human expression that are seen in a complementary redundant manner to each other, the performance of decision-level fusion is not as good as that of feature-level fusion because it may lose some information about mutual correlation between several modalities.

The number of studies in this direction is increasing, and research affects most existing audio-visual surveys of recognition of basic emotions from intentional display. Some efforts have been reported from the deliberate display toward the detection of non-basic emotional states [125]. Related studies conducted on naturalistic data include designing a system for detecting hunger and pain, as well as sadness, anger and fear, from infant facial expressions and cries and have investigated separating speech from laughter episodes based on both facial and vocal expression [126]. The current methods now for audio-visual affect analysis mostly are based on display of deliberately posed affect.

The data fusion methods of affective recognition currently commonly used include feature-level, decision-level and model-level fusion. Because features usually come from different modalities and the time scales and metric levels are also different, the performance of recognition based on feature-level fusion will be affected. One of the common methods to decrease the influence is normalization, but as the dimensions of feature vector are increasing, the recognition performance is still being affected.

Unlike feature-level fusion, decision-level fusion has modelled independent input and combining these recognitions from single-modal together. Decision fusion works well in conditions in which has independent inputs or do not have any influence on each other. But when the inputs complement each other, like audio and visual features of human expression that are seen as a complementary redundant manner to each other, the performance of decision-level fusion is not as good as that of feature-level fusion because it may lose some information between several modalities. The differece of feature-level fusion and decision-level fusion is shown in Figure 2.7.

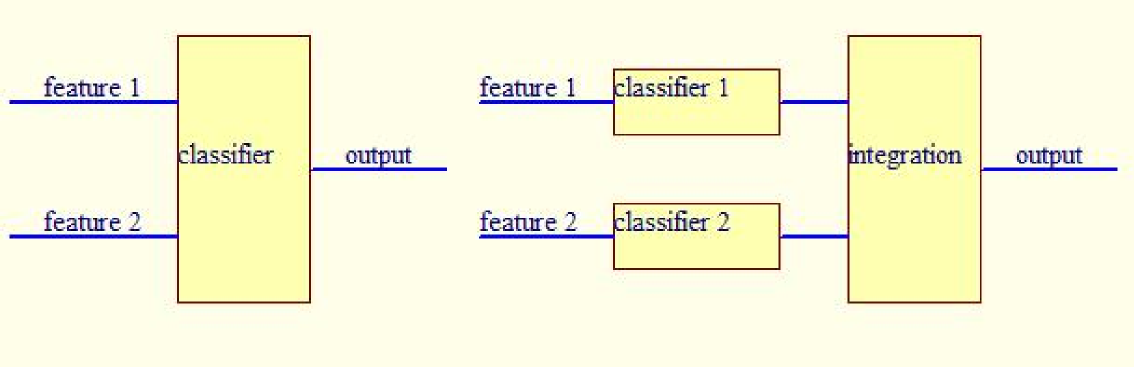


Figure 2.7: Feature-level and decision-level fusion

Aiming to deal with the problem of decision-level fusion, some researchers has worked out several model-level fusion methods. Model-level fusion not only uses correlation information between different modalities, but also handling the data streams is not a serious synchronization issue. There are many kinds of model-level fusion, like extending this fusion framework by introducing a middle-level training strategy, under which a variety of learning schemes can be used to combine multiple component HMMs [127]; presenting a tripled HMM to model the correlation properties of three component HMMs that are based individually on upper face, lower face, and prosodic dynamic behaviours [128]; or proposing an artificial neural network (NN) with a feedback loop called ANNA to integrate the information from face, prosody and lexical content [96].

Three different approaches all used SVM classifier with 2nd order polynomial kernel functions. In the first approach, a sequential backward feature selection technique was used to identify the features from both modalities that maximize the performance of the classifier. The number of features selected was 10. In the second approach, several criteria were used to combine the posterior probabilities of the mono-modal systems at the decision level: maximum, in which the emotion with greatest posterior probability in both modalities is selected; average, in which the posterior probabilities of each modalities are equally weighted and the maximum is selected; product, in which the posterior probabilities are multiplied and the maximum is selected; and weight, in which different weights are applied to the different unimodal systems.

Multisensory fusion, including audio-visual data fusion, linguistic and paralinguistic data fusion, and multi-visual cue data fusion would be highly beneficial for the machine analysis of human affect, but it is just beginning. Studies in neurology on the fusion of sensory neurons are supportive of early data fusion (i.e., feature-level data fusion) rather than of late data fusion (i.e., decision-level fusion). However, it is an open question how one can construct suitable joint feature vectors composed of features from different modalities with different time scales, different metric levels and different temporal structures. Simply concatenating audio and video features into a single feature vector, as is done in the current human affect analysers that use feature-level data fusion, is obviously not the solution to the problem.

Most researchers choose decision-level fusion, also called classifier fusion. The input coming from each modality is modelled independently, and these single-modal recognition results are combined in the end. Various classifier fusion methods (fixed rules and trained combiners) have been proposed in the literature, but optimal design methods for classifier fusion are still not available. In addition, since humans simultaneously employ the tightly coupled audio and visual modalities, the multimodal signals cannot be considered mutually independent and should not be combined only in the end, as in the case of decision-level fusion.

Model-level fusion or hybrid fusion aims at combining the benefits of both feature-level and decision-level fusion methods. However, based on existing knowledge and methods, how one can model multimodal fusion based on the multi-label multi-time-scale labelling scheme described above is largely unexplored. Several issues relevant to fusion require further investigation, such as the optimal level of integrating these different streams, the optimal function for the integration and the inclusion of suitable estimations of the reliability of each stream. In addition, how one can build context-dependent multimodal fusion is an open and highly relevant issue. Here, we want to stress those temporal structures of the modalities (facial and vocal) and that their temporal correlations play an extremely important role in the interpretation of human naturalistic audio-visual affective behaviour. Yet, these are virtually unexplored areas of research because the facial expression and vocal expression of emotion are usually studied separately.

Chapter 3

Touch Gesture Recognition Using Distinct Features and Decision Fusion

3.1 Introduction

In computer science, gesture recognition identifies human gestures through mathematical algorithms and machine learning methods with computers. Gesture recognition can come from the movement of various parts of the body, but generally refers to the movement of the body and hands. Gesture recognition can be considered a means of letting the computer understand the language of the human body. Therefore, HCI is not only the text interface or the user's image interface, but also is controlled by the mouse and keyboard [32]. Among all the HCI methods, gesture recognition is more accurate and stable.

Gesture and touch are among the most common communication methods by which humans interact with the environment and each other [125]. When we communicate with each other, body language encompasses a large amount of information we want to exchange. Studies show that gesture and touch will amplify the emotions and information during human communication [126].

Gesture recognition has been widely used in many devices and areas, including Radio Frequency Identification (RFID) [129], depth-aware cameras, gesture-based controllers and many other devices, because of its structural simplicity, universal usability and low cost [130]. That is another strong reason for developing gesture recognition methods in the HCI area.

Gestures are one of the most important communication methods human beings use to communicate with each other. There are lots of different types of gestures, like sign language and military signals. Here we focus on touch gesture, which is a communication method employing different angles and frequent touches on a surface.

Touch gesture is one of the most important ways we human beings use to communicate with each other and with other animals. Lots of research has been conducted on simulation touch communication with machines, and some products like robot dogs can respond to different touch gestures. Robot pets also can be a good solution to ethical arguments about people having real animal pets.

Touch gesture is also getting more important because of the number of touch pads we use in daily life. As more advanced screens have been utilized, more gestures can be recognised, and the control of the touchable devices is becoming more convenient.

To develop a better interaction between human beings and robot pets, it is necessary to research the model of touch gestures when human beings communicate with real pets. Studies have shown that most of the communication methods we use with pets are through touch. My research of touch gestures started the Social Touch Challenge. I used the dataset of touch gestures with my colleagues, and we got the second best results in the challenge. After the challenge, I kept working on the touch gesture dataset and achieved more and better results.

In this chapter, I will start with the social touch gesture challenge 2015 [35]. At the end, I will show some of my latest results from the dataset.

3.2 Related works

Social Touch Challenge 2015 was committed to improving the recognition of touch gesture. The advances of touch-based interaction are among the most important methods to improve human-robot interaction [33]. A touch gesture dataset has been collected in two parts, HAART and CoST. In this paper, we use the HAART dataset. The dataset uses an 8×8 accuracy sensor to collect the contact surface data in real time. The contact surface data are presented by an 88 matrix, and each pixel is a number presenting the strength of contact.

The sensor used in collecting touch gesture data of the HAART and CoST datasets is called a conductive fur sensor. The basic look and construction is shown in Figure 3.1.

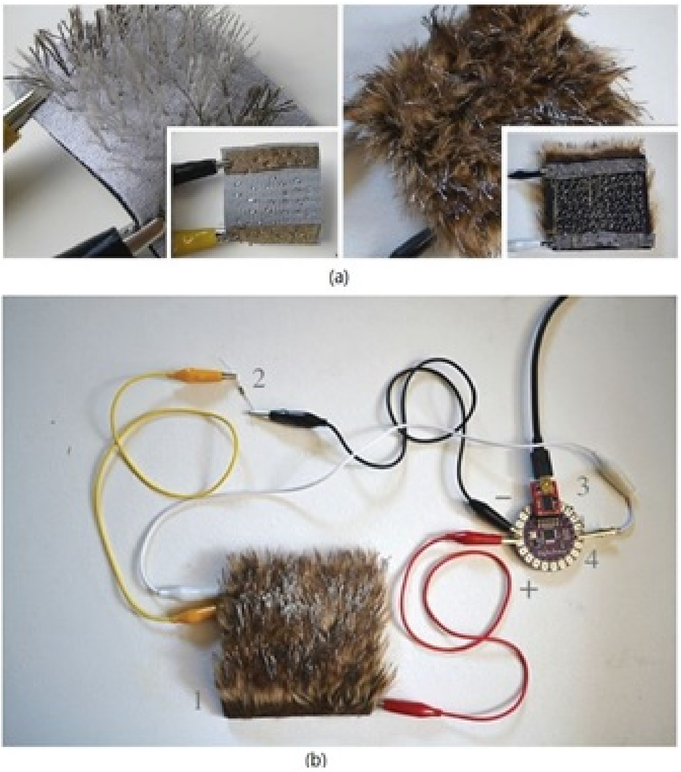


Figure 3.1: Pictures of the conductive fur sensor, (a) is basic look of conductive fur touch and gesture sensor and (b) is basic construction of fur sensor prototype [3]

Many studies have shown that gesture recognition is an intuitive and easy way to achieve human-robot communication [74]. Especially as A.I. is developing, the human-robot interaction with gesture recognition is becoming a more valuable subject [131].

Some studies also show that using video or image recognition methods on gesture recognition is effective and convenient to operate, for example in [132]. This paper uses two commonly used image processing methods on a gesture recognition dataset to prove that sometimes a video feature is better than typical gesture features in gesture recognition.

The best article was presented by the Grenoble System. The social touch gesture database has two datasets, the HAART dataset and the CoST dataset. The Grenoble System reaches the highest accuracy on both datasets. Currently, their CoST dataset result is still the highest. In the HAART dataset, a 3D-CNN method has been proposed, but because 3

Dimension Convolution Neural Network (3D-CNN) cannot process videos with different frames, this method cannot be used on the CoST dataset. We have also took part in Social Touch Challenge 2015 and our article was ranked second.

In the Social Touch Challenge 2015, the winning paper [9] provides many kinds of feature extraction methods and first introduced Sobel edge detection as a pre-processing method. Some traditional touch gesture recognition methods have been utilized such as number of frames, maximum and minimum value of all channels, average value of all channels and mean pressure over all frames of each column and row.

The Grenoble System is based on deep analysis of social touch dataset and uses a Sobel edge detection as a filter and a series features extract from temporal and spatial on the dataset. There are three sets of features: global features, which are the gestures statistics comprehensively; channel-based features extracted by different channels that absorbed in different channels' relationship in spatial and ignore their relationship in temporal; and sequence features that extract the features of gesture changes temporal.

Basically, the methods are divided into three parts, global features, channel-based features and sequence features. In group 1, global features, the main idea is to consider the overall statistical features of touch gestures, for example, the number of frames for each sample, which is fixed in the HAART dataset as 432; the average, maximum and minimum pressure for the overall dataset; and the number of frames are blanked in one sample. In group 2, the main idea is to ignore the relationship between each frame and only consider the features of the whole gesture sample, such as average pressure for all frames in one gesture sample and average pressure variation for all frames in one gesture sample. Group 3 includes the features that consider the relationship of sequences. The features include FFT and DCT.

Another paper [127] presented at the Social Touch Challenge 2015 has provided some similar features with [9]. But without Sobel edge detection, the final result is only 66.53% using a RF machine learning method. In addition, [127] provides some advanced points like considering the touch gesture as video data. Two video features have been utilized in [127], including LBP-TOP and SMMHH. The final feature was combined with typical touch features and video features. But without Sobel edge detection, the improvement of the video feature was not reflected.

Five sets of features have been extracted: SD of pressure surface, which is used to distinguish between different areas of different pressures; BMH, which is used to describe the image of each touch gesture on the 88 frames; Spatial SMMHH on touch dynamic, which usually used on video processing for dynamic changes of videos; and LBP-TOP, which is also a video processing method.

A feature selection method has been proposed, mRMR. It has the advantage than PCA and the basic theory has been introduced in Chapter 2. In my research after Social Touch Challenge 2015, I also used this method for feature selection and agree that the performance of it is better than that of PCA.

Also, the Grenoble System did a feature evaluation with auto-validation and cross-validation and calculated the contribution in each feature group. In all, 164 features have been finally chosen. The testing also proved that the 164 features achieved the best result. The results of Grenoble system are shown in Table 3.1.

Table 3.1: Classification Accuracy(%) of different features on the CoST dataset [9]

Features	Train set	Test set
Global features	58.22	52.17
Global features and Channel-based features	66.56	56.81
All features	66.89	58.96
Add Sobel Frames	68.61	60.10
164 features	69.77	61.34

Some samples in the dataset cause confusion, like squeeze and grab, rub and stroke. A manual sample selection has been utilized. The accuracy increased after the sample selection and has been shown in Table 3.2.

Table 3.2: Overall results on the CoST task [9]

Training data	manually selected subset (%)	all (%)
SVM	59.91	60.51
RF	61.34	60.81

For the HAART dataset, the highest result is proposed by a 3D CNN system. The article comparing three different deep learning methods and the 3D CNN achieved the best result. The three deep learning methods are (Long Short Term Memory with Geometric Moment Features) GM-LSTM, (Long-term Recurrent Convolutional Network) LRCN and 3D CNNs [29].

Table 3.3: Overall results on the HAART task

Author	Method	Accuracy (%)
Zhou et al. [29]	3D CNN	76.1
Zhou et al. [29]	GM-LSTM	65.3
Zhou et al. [29]	LRCN	60.6
Ta et al. [9]	Grenoble System on RF	70.9
Ta et al. [9]	Grenoble System on SVM	68.5
Gaus et al. [127]	Multi-features with RF	66.5
Gaus et al. [127]	Multi-features with Multiboost	64.5

According to the results in Table 3.3, among the three deep learning methods, only 3D CNN is much better than the others. One reason is that the sample of the HAART dataset is too small. Deep learning methods need large samples to reach high accuracy. 3D CNNs achieve a very good result, but because of the characteristic of 3D CNNs, they only work when the frames of videos are the same.

3.3 Touch gesture recognition system

3.3.1 System overview

In this work, there are five steps to processing the dataset: Sobel edge detection as pre-processing; multiple feature extraction including LBPTOP feature, SMMHH feature, MSD feature and SD feature; mRMR feature selection, which is used to reduce the dimension of features; classification with three different classifiers, SVM, multiboosting and RF; and HC decision-level fusion, which considers all the classified results of different features and calculates a final prediction.

Before Sobel edge detection and feature extraction, another important step is to analyse the HAART dataset. After looking deeply inside the dataset, it will be more reasonable to use video processing methods on this touch gesture dataset. The original HAART dataset is a 4D matrix that includes the length and height of each frame, which are 8 for both; number of frames, which is fixed at 432; and number of samples, which is 578 for train set and 251 for test set.

To process the dataset as a video, the most distinct feature of each touch gesture is their shape on this 8×8 image. Not only the shape on each frame, but also the shape on time line is an important feature. For example, the touch gestures ’scratch’ and ’tickle’ are very similar on the frame level, but as their contract times are different from each other, the shape on the timeline will be distinguished. The two methods used in this research are first Sobel edge detection to emphasize the shape of each gesture and then LBPTOP to extract features from three different planes.

LBP histogram has been utilized in computer vision and image identification areas for a long time. Based on LBP, a common method named LBPTOP has been introduced and is widely used in video recognition and facial expression detection. This paper has discussed the LBPTOP method used in gesture recognition area. The dataset being tested is the HAART gesture set in Social Touch Gesture Recognition Challenge 2015, which is not similar to a typical video. To improve the result, another video recognition preprocessing method, Sobel edge detection, has been utilized. Both LBPTOP and Sobel edge detection are typical video processing methods that have achieved advanced results in gesture recognition.

The machine learning system is built based on computer vision method. The basic system includes three parts: Sobel edge detection, feature extraction by LBPTOP and machine learning, which includes SVM and RF. The schematic of the whole system is shown below.

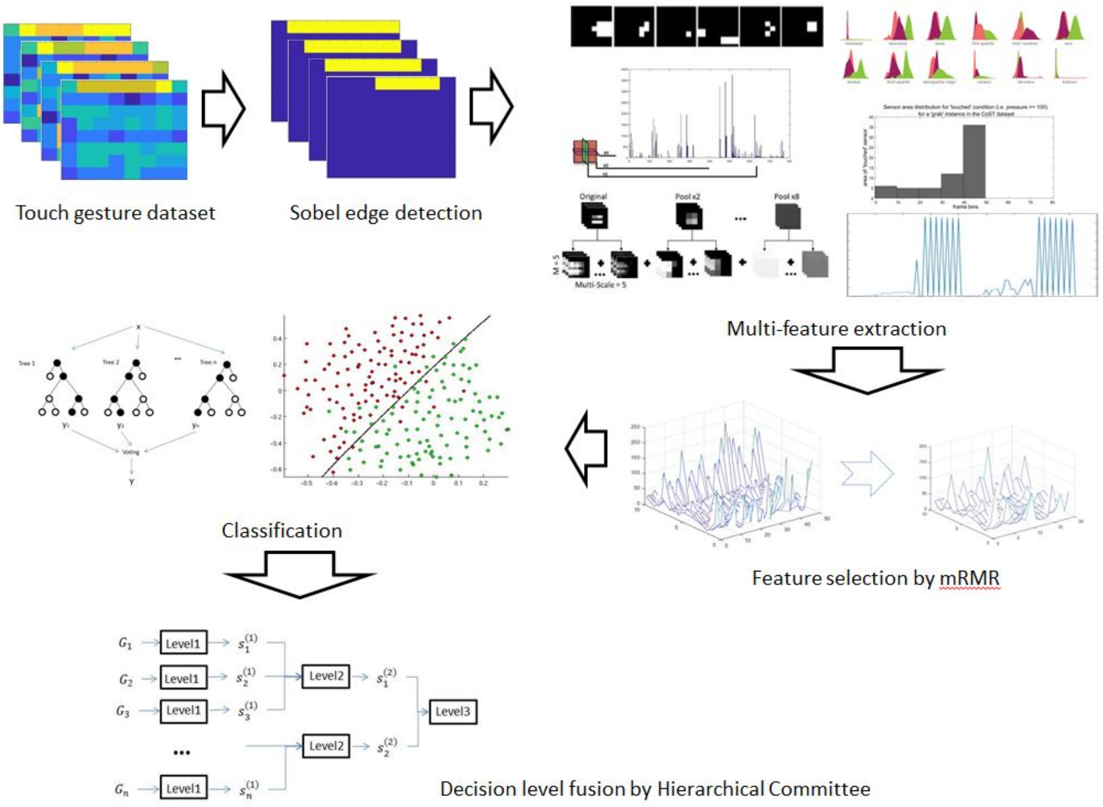


Figure 3.2: Overview of the whole touch gesture recognition system

Basically, the system is divided into five parts: Sobel edge detection, multiple feature extraction, dimension reduction, classification and decision-level fusion. The first is edge detection, and the method chosen is Sobel edge detection. The second is feature extraction, including many different feature extraction methods, like SMMHH and MSD. The third is feature selection, for which I chose mRMR as my feature selection method. The fourth is machine learning. I have used three machine learning methods: RF, Multiboosting and SVM. The last is decision-level fusion. I chose HC as my fusion method. The overview of system is shown in Figure 3.2.

3.3.2 Sobel edge detection

Sobel edge detection utilized here as a pre-processing method. The basic principle of Sobel edge detection is introduced in chapter 2. The basic idea of using edge detection method here is processing the data as image data. As the figure in system overview showed, the original images of gestures have complex waves and difficult to process by image-based methods. Edge detection method can draw the outline of images and will help to process the data as images.

The reason of chosen Sobel edge detection as pre-process method is it has been proved as a very effective pre-processing method in touch gesture recognition according to the research of Ta [9], which is the winning paper in Social Touch Challenge 2015. The thought behind chosen Sobel edge detection is to simplify the different areas in touch sensors by using 1 and 0. It would remove redundant information and increase the cleanliness of extracted features.

3.3.3 Multi-feature extraction

There are several different features have been chosen in the experiment. Some are traditional video-based features like LBPTOP feature and MSD features. MSD have been chosen because it is naturally present the basic characteristics of touch dataset, like average frames, maximum, minimum, mean, etc. LBPTOP has been chosen because it is a comprehensive feature to describe a video. An alternative feature for LBPTOP is LPQ-TOP, which would be applied in next chapter. Here the performance of LBPTOP is better.

LBPTOP feature extraction

Local binary patterns (LBP) were provided by [133] as a feature extracted method in image processing in 1994. It is a very powerful feature in texture classification, and combined with the directional gradient histogram, it can be very effective in raising the detection effect on image datasets. The basic method is comparing each pixel with its nearby pixels and saving the result as a binary number.

LBPTOP is a video feature modified on LBP that has been widely used on FER and human action recognition [76]. LBPTOP calculates LBP features in three orthogonal planes, including XY-LBP, XZ-LBP and YZ-LBP, where X, Y and Z are three directions in the space and XY, XZ and YZ are three planes in the space. LBPTOP is a feature conclude LBP feature from these three planes.

LBPTOP has been proved to be an advanced FER method, for example, in [134], [135] and [130]. Because the touch gesture dataset was collected frame by frame and each frame is an 88 matrix, it is possible to consider each touch gesture as a video and to extract the feature from three different planes. The HAART dataset has solid frames for each single touch gesture sample, so the three planes of LBPTOP are fixed.

MSD features

Among all these features, MSD feature is extracted. There are 12 features in version 2 features, including minimum, maximum, mean, first quartile, median, third quartile, total variation, area, interquartile range, variance, skewness and kurtosis.

For both the CoST and HAART datasets, each frame of the gestures is expressed as a 88 matrix. The number of frames for the HAART dataset is solid, and they are all 432 frames, but the number of frames for the CoST dataset is variable. Thus, we extract features that can ignore the impact of frames. Therefore, all the features in version 2 are based on the whole gesture movement of each pixel.

The first three features are minimum, maximum and mean, e.g. in the HAART dataset, minimum is the smallest number for each pixel in 432 frames. Then we extracted three quartiles and an interquartile range. Quartiles are numbers in the position of each quarter (25%, 50%, 75%) when sorting the sequence from smallest to largest. The interquartile range is the difference of the third quartiles (75%) and the first quartiles (25%). Area is the area of each pixel moving on the axis of time, specifically, the summation of all numbers of each pixel in all frames. Total variation is the summation of changes in the value for each pixel in all frames of a gesture. Variance, skewness and kurtosis are also calculated by pixels. Variance is based on equation 3.1, skewness is based on equation 3.2 and kurtosis is based on equation 3.3.

$$Var(X) = E[(X - \mu)^2] \tag{3.1}$$

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^(3/2)} \tag{3.2}$$

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^4} - 3 \tag{3.3}$$

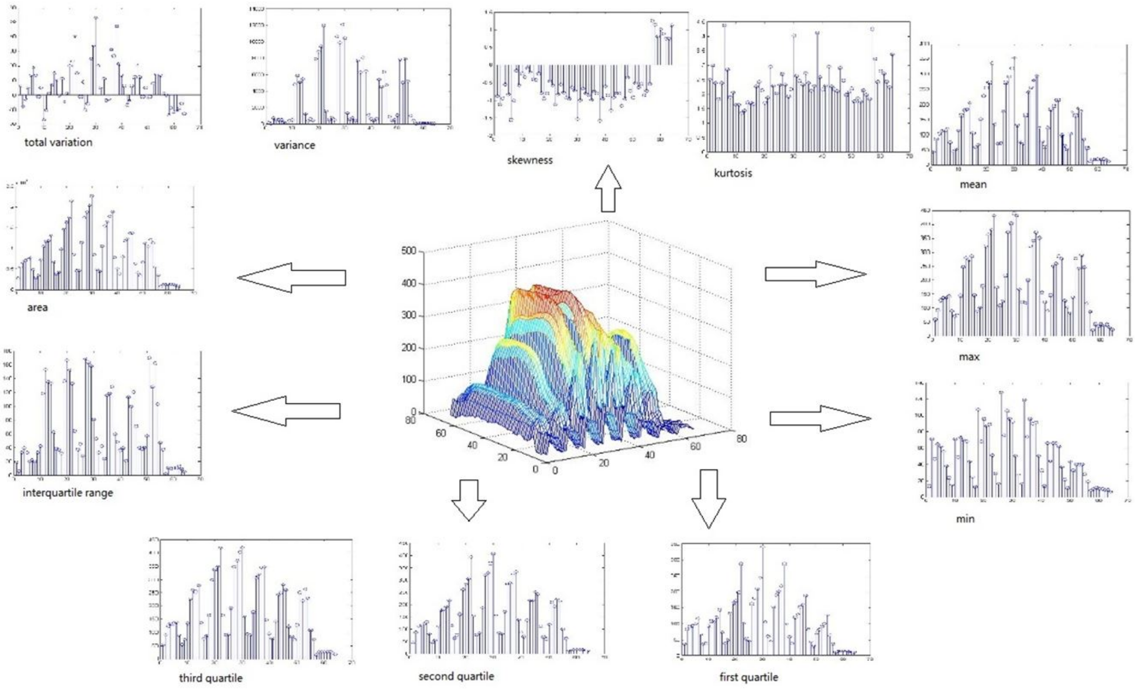


Figure 3.3: 12 features extracted by MSD

3.3.4 Feature selection with mRMR

The theory of mRMR has been introduced in literature review. mRMR is used to re-order the features by the correlation of features. By deleted features have low correlation with others and bad performance on classification, a feature selection process can be achieved and the performance can be increased.

3.3.5 Classification

There are many classification methods in touch gesture recognition in state-of-the-art. Among all these classifier, RF is the best, so it has been chosen as the main method. Also, SVM is chosen to be a comparison as a common classification method as in paper [127] and [128], it is usually the second best classifier.

RF

RF has been proved to be an effective method in [127]. Among all those traditional classifiers, RF has an unique feature selection mechanism and suitable for classify samples have long feature vectors. In Social Touch dataset, after multi-feature extracted, the feature vector can be very large and achieves a number over 1000. So RF is very suitable in this case and that is also why we choose RF in [127].

SVM

SVM is a very common used traditional classifier and very suitable for a control group of classifier to verify the performance of other classification methods. Also, SVM has been proved to be the best classifier in Grenoble System [9] during Social Touch Challenge 2015. The theory of SVM has been introduced in Literature Review.

Multiboost

Like RF, Multiboost was firstly used in [127] and the reason utilized Multiboost is similar with RF. The feature vector of multi-feature extraction is very long and Multiboost calculate the weight of each feature and after each boost, Multiboost will renew the weight by calculate the error rate of each feature. With weighted the features, the length of feature vector would not impact the accuracy too much.

3.3.6 Decision level fusion

After dimension reduction, the features are sending to machine learning processing. We have tried to combine them directly together in Social Touch challenge 2015 [35]. But it will be more effective if there is some better way to combine them. One common used method is decision level fusion. Here I choose a method called HC [136] which has been proved effective on CNN in FER.

The basic idea of HC is to divided the features into different groups and put these groups into different levels, which is called Committee here. After calculate the weights of different features in different groups according to the performance of each feature, the processing will give them different scores. Then multiply by the scores, the final results will be calculated by all the features' predictions.

HC is decision-level fusion. It is a method aimed to improved the results of multi-column deep neural network (MCDNN) [137]. For decision fusion, there are three widely used methods: majority voting, median rule and simple average rule. The majority voting method directly uses prediction category labels and selects the most votes in the category. Instead of using labels, median rule and simple average rule, it uses continuous intimate class or fraction [136].

HC uses a validation-accuracy-based simple weighted average (VA-Simp-WA rule) to determine the importance of the decision as members of the weighted average calculation and verification of performance distribution rights for fractional weights. In determining the weights of each committee, HC uses an exponentially to expand the difference between them. The final ensemble of an $m = 1 \ldots M$ member model with a validation accuracy of Z_m and posteriori class probability vector S_m is shown as equation 3.4.

$$s_{final} = \frac{\sum_{m=1}^{M} (Z_m)^1 s_m}{\sum_{m=1}^{M} (Z_m)^q} = \sum_{m=1}^{M} d_m s_m \tag{3.4}$$

Where a decision weight d_m reflects the normalized significance of the model decision of m, where $(0 \leq d_m \leq 1)$ and an exponent q is a hyper-parameter to determine how much the qualified members are emphasized $(q > 1)$ or de-emphasized $(q \geq 1)$. Finally, a class with the highest value in exponentially-weighted class probabilities is chosen. The basic schematic of a 3 level HC is shown in Figure 3.4.

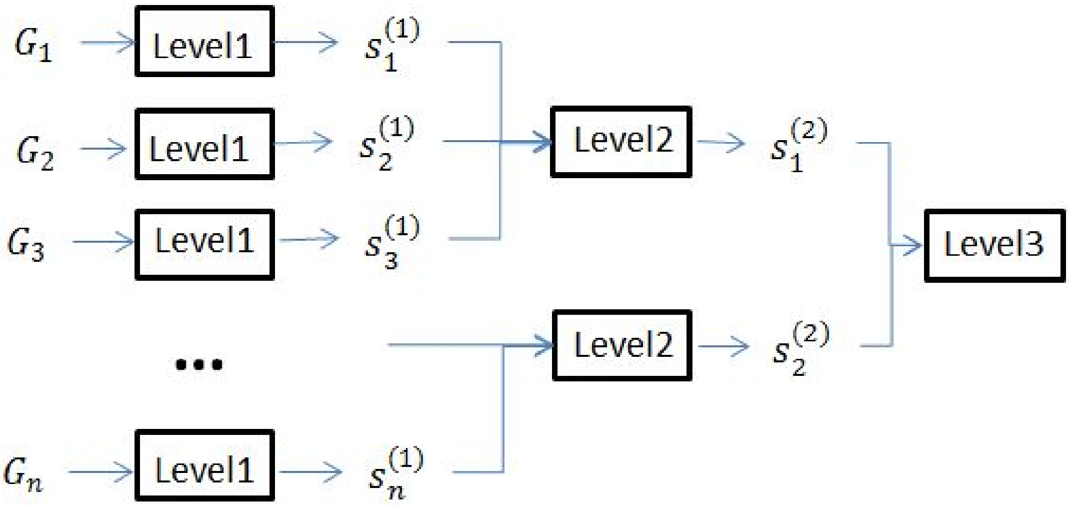


Figure 3.4: Schematic of 3 level HC

To bring various mistakes to the better committee, we first built multiple deep CNNs as individual committee members. Here, the depth model is trained by applying various network architectures, using several strategies for external data, and different input pre-processing and random initialization. Through these people, we formed a stratification committee that uses an exponential weighted average based on effective accuracy. This exponentially weighted decision fusion is superior to other commonly used collection methods because it increases generalization capabilities. In addition, the hierarchical structure does make more reliable decisions under the consensus of each subgroup.

3.4 Experimental results

3.4.1 HAART dataset

The machine learning methods used are RF and Multiboosting. As with [9], RF achieves the better result on the HAART dataset, 66.53%. The confusion matrix of RF result is shown in Table 3.4.

Table 3.4: confusion matrix of RF in HAART dataset for social touch gesture recognition using RF and boosting on distinct feature sets

Labels	constant	no-touch	pat	rub	scratch	stroke	tickle
constant	34	0	0	0	0	0	0
no-touch	1	36	0	0	1	0	0
pat	0	0	30	0	1	0	2
rub	0	0	1	16	2	13	0
scratch	0	0	3	17	26	3	24
stroke	0	0	1	2	0	18	3
tickle	0	0	1	1	6	2	7

There are also some good results on another Social Touch dataset, CoST, provided in [128]. In this paper, researchers divided touch gesture into different groups, such as male and female, gentle and rough. The results obviously increased in some emphasized touch gestures; for example, the group ’rough’ achieves better results than the group ’gentle’.

The basic idea of [128] is to group the train set into 3 different groups according to the pressure of CoST because the CoST dataset gives an additional 3 labels for each sample and describes the touch as ’normal’, ’gentle’ and ’rough’. After dataset analysis for each kind of touch gesture, 54 typical gesture features are extracted, including mean and maximum pressure over channels and time, different kind of peaks, and contract area and time for each gesture. This is different from the HAART dataset studies cited in this paper. The HAART dataset does not have this extra group of labels, so all the touch gestures in HAART are considered one kind of touch emphasis.

Then 5 different kinds of machine learning methods are used, including Bayesian, decision tree, SVM linear, SVM Radial Basis Function (RBF) and NN. Among all those methods, SVM-RBF achieves the best results, 62%.

Here, we seek a common solution of recognition of touch gestures as video data. Combined with the advantages of [9] and [127], a method of combined Sobel edge detection and LBP-TOP is chosen. In chapter 3, both methods will be introduced, and the dataset will be analysed. Chapter 4 describes the results of the experiments and some compression. Based on all the results in related works, we choose RF and SVM as our machine learning methods.

3.4.2 CoST dataset

Unlike the HAART dataset, the CoST dataset has 14 labels, and each touch gesture has a description of hard or normal that reflects the level of touch gesture on the sensor. The 14 labels are grab, hit, massage, pat, pinch, poke, press, rub, scratch, slap, stroke, squeeze, tap and tickle. The number of samples in the CoST dataset is much larger than in the HAART dataset. There are 3524 train samples and 1679 test samples in the CoST dataset. Another difference between the CoST and HAART datasets is the length of samples. As mentioned, all the samples in the HAART dataset are 432 frames, but the frames of the CoST dataset are not fixed. Consequently, some methods like 3DCNN that achieve very good results on HAART cannot be used on the CoST dataset. In Social Touch Challenge 2015, we used the same method as for the HAART dataset on CoST, although the accuracy was less than that of the HAART dataset. The results of CoST dataset is shown in Table 3.5.

Table 3.5: Recognition accuracy of each feature and their combination in CoST dataset

Dataset	Feature set	RF (%)	Boosting (%)
training	SD	41.24	41.31
	BMH	27.55	28.88
	MSD	44.82	44.93
	SMMHH	52.68	52.56
	LBPTOP	45.65	46.63
	Combine	64.52	64.44
testing	Combine	59.50	58.19

The experiment is based on the HAART dataset of Social Touch Challenge. The first step is pre-processing using Sobel edge detection on each frame of touch gesture. Then, we consider the gesture movement as a video and use LBP-TOP extracted feature.

Two parts of LBP-TOP feature have been extracted. One is after Sobel edge detection, and the other is extracted without Sobel edge detection. Both results and their combinations are shown in Table 3.6.

Table 3.6: Recognition accuracy of RF and SVM with and without Sobel edge detection

Classifier	Sobel + LBPTOP (%)	LBPTOP	Combination (%)
SVM	72.51	67.73	75.70
RF	71.71	67.33	73.31 - 76.10

The result obviously increases with both Sobel edge detection and LBP-TOP features. Furthermore, the combination of them achieves the highest result currently. One revision that should be pointed out is in [127], in which the LBP-TOP is actually only using two orthogonal planes and the results of LBP-TOP only is 10% lower than when using three orthogonal planes. In Table 3.7 is a confusion matrix of combined results in RF with a performance of 75.70%.

Table 3.7: confusion matrix of RF in HAART dataset for social touch gesture recognition of combine features using boosting on distinct feature sets

Labels	constant	no-touch	pat	rub	scratch	stroke	tickle
constant	33	0	0	0	0	0	0
no-touch	1	30	0	0	1	0	0
pat	0	1	31	0	0	2	1
rub	0	0	2	20	2	3	0
scratch	0	0	0	14	29	0	1
stroke	0	1	2	1	0	29	0
tickle	1	4	1	1	5	2	18
Recognition rate	94%	83%	86%	56%	81%	81%	50%

From the confusion matrix in Table 3.8, and compared with the results in [127] in Table I, as the edge has been emphasized, the chance of the gestures ’rub’ and ’tickle’ being recognized as ’scratch’ is dramatically decreased.

Table 3.8: confusion matrix of LBPTOP features without Sobel edge detection in RF 67.33% and the recognition rate of each label

Labels	constant	no-touch	pat	rub	scratch	stroke	tickle
constant	34	0	0	0	0	0	0
no-touch	1	32	0	0	0	0	0
pat	0	0	27	0	0	1	0
rub	0	0	5	16	1	9	0
scratch	0	0	1	16	31	1	28
stroke	0	0	1	1	0	22	1
tickle	0	0	2	3	4	3	7
Recognition rate	97%	89%	75%	44%	86%	61%	19%

Comparing Table 3.7 and Table 3.8 shows that it is obvious that Sobel edge detection makes a big contribution for distinguishing ’rub’ and ’tickle’ from ’scratch’. That is because the Sobel algorithm emphasizes the edge both on frame level and time level Then LBP-TOP features will make a big difference on both levels. For some typical touch gestures, the recognition rate is very high. The recognition for each gesture is shown below.

Except for ’rub’ and ’tickle’, all other distinguished touch gestures’ recognition rates are over 80%. Whether we consider ’rub’, ’tickle’ and ’scratch’ as touch gestures or a continuous video, it is hard to distinguish these three labels. But they do have some difference on the timeline. Even on each single frame, their images are similar, and their contract times are different. That is the improvement of ’rub’ and ’tickle’ recognition with video processing method.

The challenge includes two datasets, CoST and HAART. The CoST dataset includes 14 gestures: grab, hit, massage, pat, pinch, poke, press, rub, scratch, slap, squeeze, stroke, tap and tickle. The gestures are registered by a 88 pressure sensor grid that is wrapped around a mannequin’s arm. The corpus consists of the data of 31 subjects who performed 14 touch gestures in 3 variations (normal, gentle and rough). The HAART dataset contains 7 gestures: pat, contact without movement (press), rub, scratch, stroke, tickle and

’no touch’. Each touch action was performed on a 10x10 pressure sensor for 10 seconds. To assess feature robustness under realistic operating conditions when installed on a robotic animal model, each participant contributed gestures with the sensor mounted on all permutations of 3 substrate conditions (firm and flat; foam and flat; foam and curve) and 4 cover conditions (none; short minkee; long minkee; synthetic fur). The resulting dataset includes 840 gesture-captures (12 conditions × 7 gestures × 10 participants).

For the features extracted part, we chose several features, including LBPTOP, SMMHH and skewness. I was responsible for MSD features. A third set of features aimed to capture the main trends of the pressure behaviour. Twelve statistical features were extracted from each sensor sequence $F = (f_1, f_2, \ldots, f_L)$, and conclude minimum, maximum, mean, first quartile, median, third quartile, area, total variation, interquartile range, variance, skewness and kurtosis. The number of frames is presented as ’L’; the summation value for different sensor of all frames is presented as ’area’; and the summation values for change in sensors is presented as ’total variation’. The number of frames for the dataset is fixed at 432 frames, but the number of frames for the CoST dataset is different.

Based on these works, I have combined with the result of the winner of the challenge. They have two advantages. One is the Sobel edge detection, and the other is the DCT and FFT features. Table 3.9 is the results for features using Sobel edge detection.

Table 3.9: Results of different features with k-NN and RF classifier

Features	k-NN (%)	RF (%)
DCT and FFT without edge detection	37.1	53.8
DCT and FFT with edge detection	34.7	51.4
DCT and FFT with and without edge detection	34.3	53.0
LBPTOP without edge detection	49.0	69.3
LBPTOP with edge detection	47.0	70.1
LBPTOP with and without edge detection	49.8	72.9
MSD without edge detection	43.8	66.9
MSD with edge detection	44.6	64.9
MSD with and without edge detection	44.6	66.5
SMMHH without edge detection	46.6	56.6
SMMHH with edge detection	55.4	62.9
SMMHH with and without edge detection	50.6	64.5

The results in Table 3.9 show that for some of the features, Sobel edge detection is effective, but some need to combine Sobel edge detection and original features. Another improvement is mRMR, because there are too many features when all the features are combined. It is very important to choose the useful information.

The next step to verify the effectiveness of mRMR by comparing with the results of feature selection by PCA and results without feature selection. The results are shown in Table 3.10.

Table 3.10: Results of PCA and mRMR

Feature	PCA		mRMR	
	k-NN (%)	RF (%)	k-NN (%)	RF (%)
LBPTOP without edge detection	50.6	73.7	50.6	73.7
LBPTOP with edge detection	48.2	68.9	49.4	71.7
LBPTOP with and without edge detection	53.4	69.3	53.4	74.5
MSD without edge detection	44.6	60.6	45.4	66.9
MSD with edge detection	44.6	59.8	44.6	64.1
MSD with and without edge detection	44.6	57.4	44.6	64.1
SMMHH without edge detection	47.0	62.9	55.0	62.9
SMMHH with edge detection	55.8	67.3	54.9	62.6
SMMHH with and without edge detection	52.6	66.1	55.8	65.3

Comparing PCA and results without feature selection, the results of mRMR are obviously increased. In the touch challenge, the best result is 70.91%. Here with LBPTOP combined with Sobel edge, the result is much better. But when all the 12 features are combined, the result is not good enough. So HC has been utilized and the results are shown in Table 3.11

Table 3.11: Results of directly combination and HC

No HC	HC	
RF	k-NN	RF
70.92%	60.56%	76.10%

In Table 3.12 is comparison of results of paper [127] [9] and the latest experiment. In paper [9], the feature extracted method is named distinct feature sets which combined 5 different groups of features. In paper [127] the system is named Grenoble system, which combined 3 different groups of features.

Table 3.12: Recognition accuracy for all related work on HAART dataset

Author	Method	Performance (%)
Gaus et al. [127]	Distinct Feature Sets on Boosting	64.5
Gaus et al. [127]	Distinct Feature Sets on RF	66.5
Ta et al. [9]	Grenoble System on SVM	68.5
Ta et al. [9]	Grenoble System on RF	70.9
Zhou et al. [29]	3D-CNN	76.1
Ours	Proposed multi-feature system with decision fusion	**76.1**

According to the comparison of mRMR and PCA in HAART dataset, mRMR has been proved to be a better solution than mRMR, so in COST dataset, the feature selection method has been chosen as mRMR. The automatic touch gesture recognition system for CoST dataset is same with HAART. The classifier used in CoST system is also RF, but same as HAART, the score calculation in HC used k-NN, and then used RF as classifier. The result of each step of CoST dataset is shown in Table 3.13.

Table 3.13: Recognition accuracy for CoST dataset

Method	Performance (%)
Multi-feature without Sobel edge detection	60.27
Multi-feature with Sobel edge detection	61.17
mRMR feature selection	61.88
HC decision level fusion	62.66

Table 3.14 shows the results published on CoST comparing with our system.

Table 3.14: Recognition accuracy for all related work on CoST dataset

Author	Method	Performance (%)
Gaus et al. [127]	Distinct Feature Sets on Boosting	58.19
Gaus et al. [127]	Distinct Feature Sets on RF	59.50
Ta et al. [9]	Grenoble System on SVM	60.51
Ta et al. [9]	Grenoble System on RF	60.81
Hughes et al. [138]	CNN	42.34
Hughes et al. [138]	CNN-RNN	52.86
Hughes et al. [138]	Autoencoder-RNN	33.52
Jung et al. [128]	SVM linear	59
Jung et al. [128]	SVM-RBF	60
Jung et al. [128]	Nerual Network	59
Albawi et al. [28]	CNN	**63.7**
Ours	Multi-feature with decision fusion	62.66

Compared with other results on CoST, the proposed system is second, but the advantage is this system has generation and can be used on both HAART and the CoST datasets. Table 3.15 shows all the best results in both dataset.

Table 3.15: All the best results in both dataset

Author	Method	Performance (%)
Zhou et al. [29]	3D-CNN	76.1 in HAART
Ours	Proposed multi-feature system with decision fusion	**76.1** in HAART
Albawi et al. [28]	CNN	63.7 in CoST
Ours	Multi-feature with decision fusion	**62.66** in CoST

3.5 Summary

The most significant contribute of this pattern recognition system for touch gestures is the versatility. It can be used on both CoST dataset and HAART dataset and achieved the best performance on HAART dataset and second best on CoST dataset. On the other hand, the best methods on both CoST and HAART dataset in state-of-the-art could only be used on one dataset.

The total result is much better than the one we presented the first time in the competition [127] and also is better than the winning paper [9]. But both teams were facing the problems of distinguishing the touch gesture ’rub’ and ’tickle’ from ’scratch’. In this paper, the typical video processing method is proposed to solve this issue. This can be an initiated method for solving some similar problems in touch gesture recognition.

Furthermore, it not only works on solving gesture recognition, because there are more and more recognition problems in many areas. Some methods used in one area may also be good for issues in other areas. Comparing Sobel with non-Sobel results and results in Table 3.9, Table 3.10 and Table 3.11, adequate evidence shows the effectiveness of Sobel edge detection in distinguishing ’rub’, ’tickle’ and ’scratch’.

There are still lots of opportunities to improve the machine learning method. Some of the latest studies [98][99] have introduced deep learning methods in gesture recognition areas. With superior deep learning methods combined with features in this paper, the results should be improved a lot.

When we took part in the competition, as in [80], 5 parts of features had been extracted. But because of lack of good feature selection methods, different features were in conflict with each other, and the combination of all features did not achieve as good results as we expected. Here, only LBPTOP is used, and the result is relatively stable. But if there is a good method of fusion or feature selection, the combination of video feature and touch gesture feature will achieve a better result.

Chapter 4

Holoscopic 3D Micro-gesture Recognition based on LPQTOP and a Non-linear SVM classifier

4.1 Introduction

As discussed in chapter 3, touch gesture is widely used on many devices like pads and smartphones. All these operations are applied on a surface. Like the processing of HCI devices, the latest wearable devices need more accurate kinds of gestures. I will now discuss micro-gestures.

Micro-gestures are an important part of human behaviour automatic recognition. Unlike touch gestures, micro-gestures are more difficult to detect and recognize because they are more similar to each other. It is better to have more distinct micro-gestures, like 3D micro-gestures, to research than a static image of micro-gestures or normal videos.

Micro-gestures are defined by users' actions with small variations, and they have also been defined as micro-interactions. Because the variations of micro-gestures are very small, they allow fast and continuous interactions with productive tools. They offer some advantages for use on interactive equipment like tablets and wearable devices.

To recognize micro-gestures more clearly, the H3D camera is used. H3D sensors have unique abilities to capture RGB images as well as depth information. The data captured are full HD videos. The basic construction of an H3D camera is shown in Figure 4.1.

[redacted]

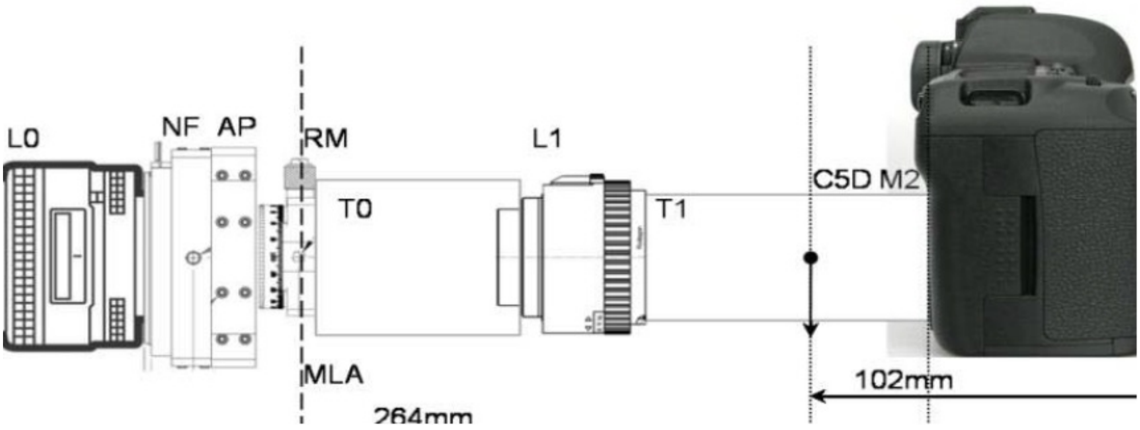


Figure 4.1: Construction of H3D camera [4]

Based on the research on touch gesture in chapter 3, these two kinds of gestures have many common characteristics. They both show some characteristics like videos, and they both need to extract dynamic information from videos. The most natural and common idea is using LBPTOP feature as in chapter 3. Also, a better method called LPQTOP has been described by [139], [140] and [141]. It achieves better performance than LBPTOP on image processing.

In touch gesture experiments, lots of classifiers have been utilized, including RF and SVM. RF has been applied because there are 7 different groups of features and RF has a selection process that can choose the most effective features. Linear SVM makes a simple linear separation between classes and ignores the non-linear relationship between classes. So in 3D micro-gesture experiments, I used both LBPTOP and LPQTOP as feature extraction and both linear SVM and non-linear SVM as classifiers for comparison.

4.2 Automatic 3D micro-gesture recognition system

After finishing the baseline of HoMG, I have tried to increase the performance of this automatic 3D micro-gesture recognition. In my previous work in HoMG, I have considered the problem by viewing whole videos. When considering the problem from images, we will find in some stages micro-gesture videos. The three gestures are distinctly different, but there are also lots of frames in all three gestures that look very similar.

The new system considers lots of classification methods and different feature extraction methods. The first step is to reduce the resolution of videos, because high-resolution videos will require huge computing resources. The reduced resolution is 38 × 66, which is enough for image and video processing. The second step is feature extraction. In my system, feature extraction is divided into two parts. One is deep feature, and the other is LPQTOP and LBPTOP feature, which have both been proven very effective in video-based recognition. The third step is classification. I have compared lots of different classifiers, including different SVM and trees. Through comparing different prediction accuracy, I found the non-Linear SVM to be the best one. The overview of the system is shown in Figure 4.2.

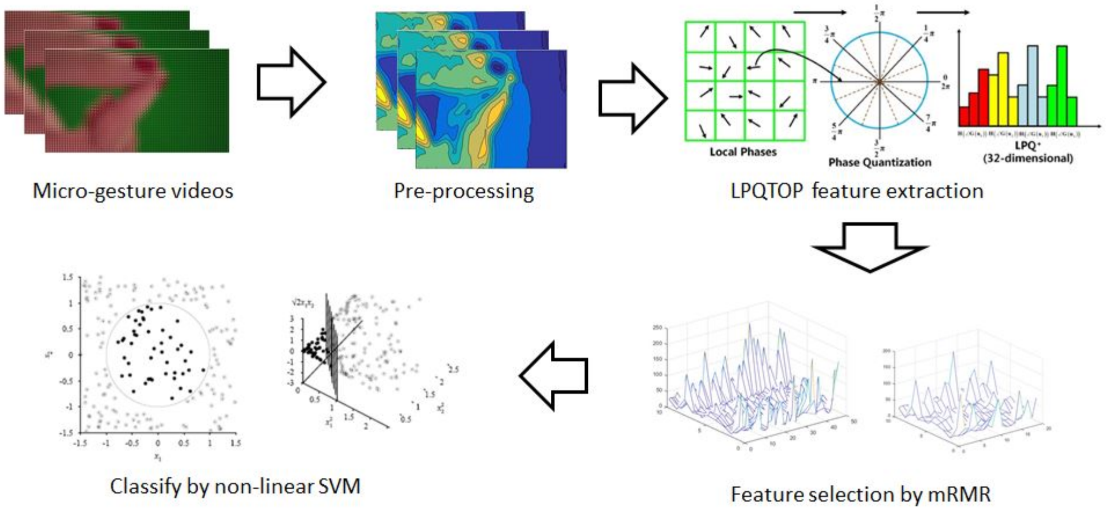


Figure 4.2: Overview of automatic 3D micro gesture recognition

4.2.1 Pre-processing

As the resolution of the videos is very high, 1080 × 1920, to directly use the original video will spend huge amounts of computing resources. It is natural to consider reducing the resolution and it would be easy for computer to process.

4.2.2 LBPTOP and LPQTOP feature extraction

The LQP characterizes texture or appearance by using sign-based, magnitude-based and orientation-based differences. The procedure of the proposed algorithm consists of two components: the learning and inference stages. It includes four stages: (1) three kinds of information (local sign, magnitude and orientation patterns) are extracted from the image, in which a local orientation pattern is realized by using orientation estimation and quantification; (2) three separate codebooks (O, S and M, respectively) are learned by using vector quantisation; (3) the sign, magnitude and orientation patterns are mapped into their corresponding codebook by using lookup table (LUT); and (4) three histograms are concatenated into one vector. The inference stage consists of all stages except the second. The schematic of LPQ is shown in Figure 4.3. Similar to LBP and LBPTOP, LPQTOP is LPQ on three different directions of LPQ.

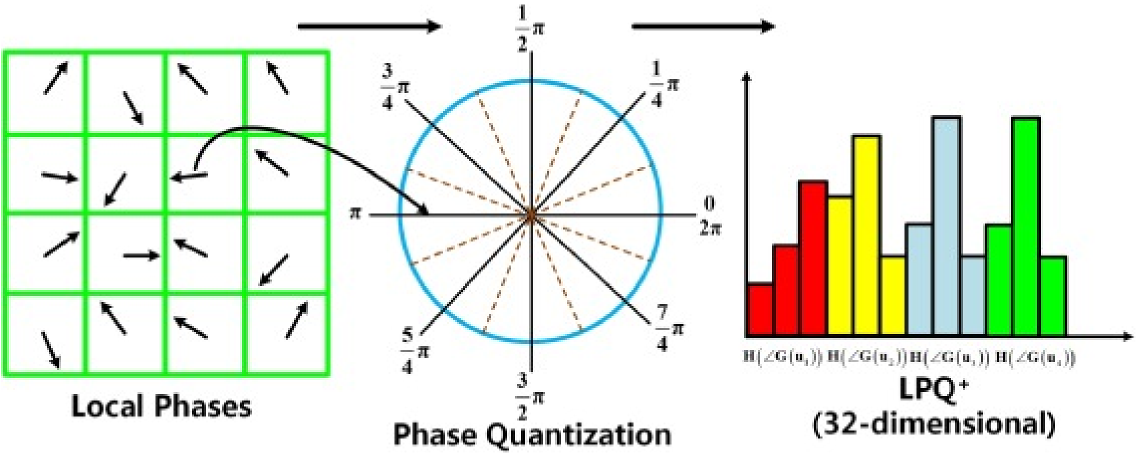


Figure 4.3: Schematic of LPQ [6]

4.2.3 Non-Linear SVM classifier

The basic idea for non-linear SVM is to project data into a higher dimension. Applied optimal hyperplane algorithm in new space for some data is not linearly separable. The basic step is to define a ϕ, calculate $\phi(x)$ for each training sample and then find a linear SVM in feature space. The basic idea of Non-linear SVM is shown in Figure 4.4.

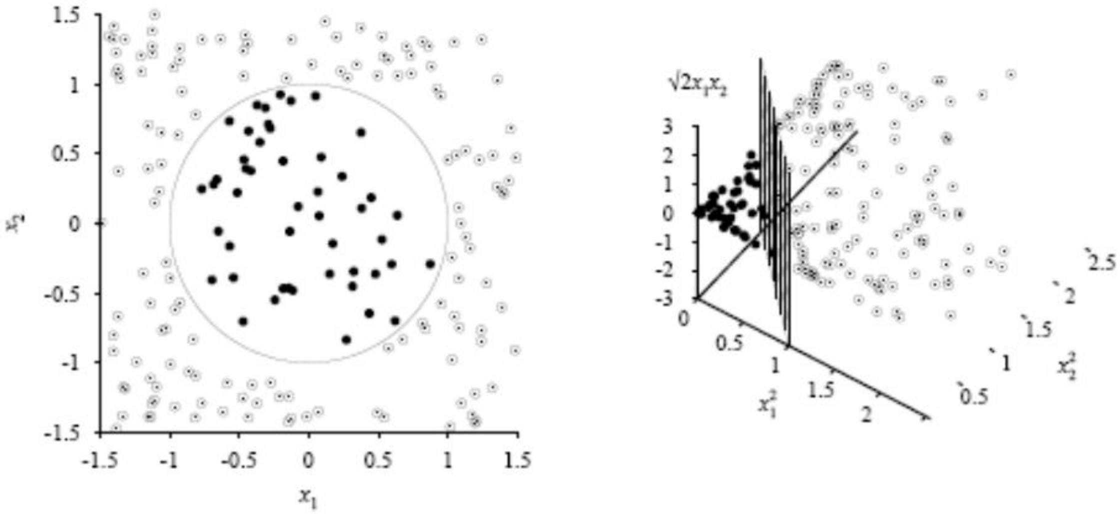


Figure 4.4: Schematic of non-linear SVM [7]

4.3 Experiments

4.3.1 HoMG database

This dataset is collected using a H3D camera by Liu [4] at Brunel University. Three micro-gestures were captured, Button, Dial and Slider. In all, 50 subjects' micro-gestures have been collected, 33 males and 17 females. In the final HoMG database, 40 subjects have been collected. The 3D micro-gesture has been collected from both left hand and right hand. The camera has been set at two different distances, 45 cm for close and 95 cm for faraway. Two different colours of background have been used, white and green. The length of a video is between 2 and 20 seconds, and the resolution of videos is 1902×1086.

For each subject, there are 12 videos. They are collected by separate right and left hands, far and close for distance, green and white background and 3 different micro-gestures, Button, Dial and Slider. The whole dataset includes 600 3D videos. The experiment has been done by independent subject. The subject has been divided into 3 groups, train set, development set and test set. In the baseline experiment, only train set and development set have been used. An example of video-based micro-gesture in HoMG is shown in Figure 4.5.

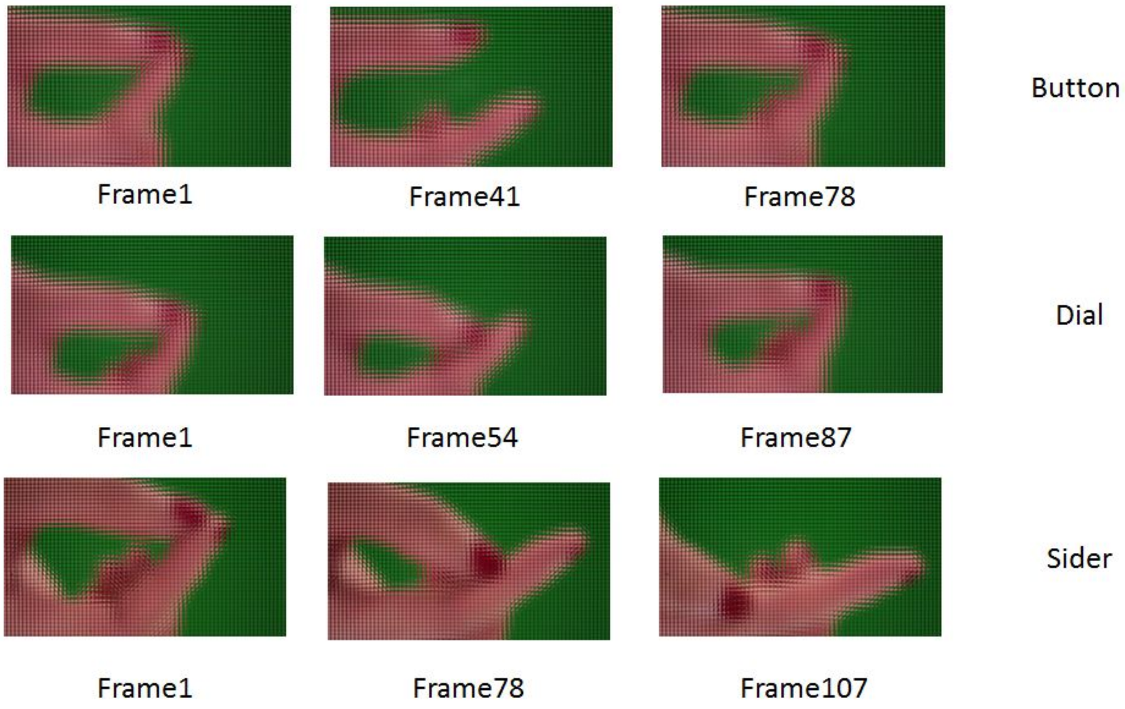


Figure 4.5: Samples of 3D micro gestures

4.3.2 Experimental results

In the recognition of micro-gesture, my experiments focus on dynamic features. I have used MATLAB classification learner for machine learning and LBPTOP and LPQTOP for feature extraction. The subjects has been divided into 3 groups, training set, development set and test set.

In the first version of the experiment, there are 30 subjects. Each subject is asked to record three different micro-gestures against two different backgrounds, with two different distances and using the right hand and the left hand. With the 30 subjects, two experiments were done. One subject left one out, so there were 29 subjects for training, and the other one is for testing. Another is 10 cross-validations. The 30 subjects have been divided into 10 groups, each with 3 subjects. Each time, 9 groups were used for training, and one group was used for testing. Both experiments are subject-independent. The result for 30 subjects, leaving one out, is shown in Table 4.1.

Table 4.1: Leave one out result for LBPTOP and LPQTOP

Feature	LBPTOP (%)	LPQTOP (%)
SVM	68.9	78.9
k-NN	50.6	51.9
Subspace Discriminant	72.5	81.1

There also two results comparing the influence of distance on accuracy. These results are based on one subject leaving out one experiment. The result for 10 cross-validation are shown in Table 4.2.

Table 4.2: 10 cross-validation result for LBPTOP and LPQTOP

Feature	LBPTOP (%)	LPQTOP (%)
SVM	71.4	78.6
k-NN	52.4	51.2
Subspace Discriminant	72.5	79.7

The HoMG dataset contains two distance of micro-gestures, to comparing the performance of recognition in different distances, the result of close gestures and far gesture is shown in Table 4.3.

Table 4.3: Classification accuracy (%) on far and close gesture subsets based on LBPTOP and LPQTOP features

Feature extraction	Close Gesture		Far Gesture	
	LBPTOP	LPQTOP	LBPTOP	LPQTOP
SVM	59.7	74.4	60.8	66.7
k-NN	36.1	37.5	35.0	42.5
Subspace Discriminant	55.0	68.3	56.1	70.0

The results show that the distance does have an effect for micro-gesture recognition. The closer the distance, the more accurate the recognition. In the second stage of the experiment, I gathered another 10 subjects' micro-gesture data. The dataset has been divided into 3 groups, train set, develop set and test set. According to the results, I choose LPQTOP as feature extraction and got the highest result from cubic SVM in classification learner. With the training on train set and testing on development set, I got the final results for baseline. According to these results, LPQTOP is much better than LBPTOP, so as the test set.

Table 4.4: Comparison of LBPTOP and LPQTOP in development set and test set

Data set	LBPTOP (%)	LPQTOP (%)
develop set	72.5	81.1
test set	60.4	84.2

Sharma et al. [142] proved the effectiveness of deep learning. Comparing my system and deep learning is necessary. Because deep learning did not consider the non-linear characteristics of the dataset, its performance should not be as good as my system's. The full system added mRMR for feature reduction, and the accuracy reaches 84.6%. Here are the comparisons for all results on the video-based HoMG dataset.

Table 4.5: Current results on HoMG dataset

	Method	Accuracy (%)
Zhang et al. [50]	ResNet	82.0
Zhang et al. [50]	Dense	82.0
Zhang et al. [50]	SE-ResNet	82.0
Zhang et al. [50]	Hybird NN	69.2
Sharma et al. [142]	LSTM	65.41
Sharma et al. [142]	Gated Recurrent Unit (GRU)	69.17
Ours	Proposed system	**84.6**

4.4 Summary

This chapter proposes an automatic recognition system for H3D micro-gestures. This system utilised traditional feature extracted and classification methods rather than other deep learning methods. But the final result is better than those of deep learning methods. Also the calculation resource cost is less than those of deep learning methods, and calculation speed is faster.

The key point of the proposed method is applied non-linear SVM based on the thought of classify the non-linear features of 3D micro-gesture. The value of this method is using low computing source expense method to achieve high performance. It achieves the best results until now, better than those deep learning methods in state-of-the-art.

From the results on micro-gestures, it is obvious that the results in a close distance is better than those in a far distance. That means that in the future equipment for automatic micro-gesture recognition, it is better to increase the resolution and try to decrease the distance between the recognition system camera and human hands.

The reason a deep learning method is not as good as the traditional method in this case may be that the sample video dataset is too small. There are only 30 subjects, and each subject has 8 videos in total. In the future, it is necessary to test the performance of deep learning again with a bigger dataset.

Chapter 5

Body Gesture Recognition based on Two-stage Classification

5.1 Introduction

Besides micro-gestures and touch gestures, another important communication tool commonly used in HCI human dynamic recognition is body gesture recognition. Unlike micro-gestures and touch gestures, body gestures are more widely used. One good example is in medicine and psychology. Body gesture has been used to recognize patients' emotions and assists in treating them, especially for depression.

Chronic pain is a disease that causes patients to suffer a lot in their daily lives, and difficult to relieve completely. The difficult of managing CLBP is coming from the characteristic of relapse and unpredictable. However, CLBP still could represent by pain-related behaviours like hesitation and guarding.

Here, a machine learning-based automatic recognition system that can detect pain-related behaviours continuously from the EMG signals of patients and body movements has been proposed. This automatic recognition system constitutes with data collection process, feature extraction methods, two different modelling and classification. A Biosensor sensor for EMG and motion capture for body movements has been used to collect the dataset. Some specific features based on body movement and EMG has been extracted. Then RF and a TSC scheme (k-NN coupled with HMM) were used for detecting pain-related behaviour. This proposed system is tested on the Emo-Pain corpus dataset [59].

In this chapter, body gestures are used to recognize the presence of CLBP. Lower back pain is common, and almost everyone will experience it at some time in their lives. When lower back pain is chronic, it is called CLBP. For some older people, CLBP may plague them for a long period of time, and they may change some daily movements inadvertently. With these changed movements like support and hesitation, the CLBP will be eased to some extent, which is not helpful for treating CLBP. So it is important to recognize CLBP automatically, because medicals and doctors cannot observe patients on any day and at any time.

5.2 Related works

Chronic Pain is a kind of neuropathic nociceptive that lasts for a long time. The definition of long time is last for more than half a year even if there is some arbitrary interval after first feeling[77]. Some epidemiological studies show that 10% to 55% people in the world have been plagued by chronic pain [63]. Originate of chronic pain could be inner brain, neurons or inner body and is very hard to treat. The participant of handling chronic pain by professional pain management teams, often including clinical psychologists and medical practitioners [65]. Some psychological treatments and non-paid medicines could relieve some kind of chronic pain, but they are only effective for some patients [66][67]. 10-year mortality has been found increased if the patients have chronic pain, especially from the respiratory system or heart. Anxiety, depression, neuroticism and sleep disturbances has also been proved could be caused by chronic pain [68].

CLBP is a lower back pain usually lasts for more than 3 month and disorder involving the nerves, bones and muscles [143] [1]. CLBP is one of the most common kinds of chronic pain and is more common in elderly people. Researches also shows in both Europe and North American there is a large number of disabilities caused by CLBP [70]. Although exercise could prevent CLBP according to some researches like [144], it is still no certain evidence that CLBP can be prevention [145].

To recognize and detect CLBP continuously is the first step of pain management. Automatic recognition of CLBP can supplement medical treatment. This is a research area strongly related to emotional detection and affective computing. The behaviour of indications of chronic pain include guarding, hesitation, bracing, abrupt action, limping and rubbing. Any of these can illustrate the existence of chronic pain [146].

There are several methods for detecting CLBP, including FER [147], voice detection [148], EMG signal analysis and motion capture. Through all these collected data, unnatural facial expressions and body movement could be captured and analysis, such as guarding behaviour when touching a sore spot or supporting the waist with the arms when standing up.

The Aim of this automatic CLBP recognition system is to detect CLBP-related behaviour continuously and automatically in daily life. Despite all the data that have been collected, because of the inconvenient of facial expression, voice signal collection and analysis, and also the poor operability of CLBP. Using musculoskeletal pain to detect CLBP is more practicable [149]. With a continuously collect of EMG signal and body movement, the aim of detecting CLBP could be achieved. In the market, there are already many wireless EMG sensors and wearable motion capture devices exist and available.

To recognize CLBP-related behaviours continuously and automatically through motion capture and EMG signals, machine learning can be used to build the model and make the predictions. This research aims at long-term continuous self-management for chronic pain patients [150].

In recent years, automatic pain related behaviour recognition using machine learning has been studied extensively. Such as the study of automatic shoulder pain detection utilizing FACS by Lucey et. al.[151]. The dataset is constituted by 200 videos, and the pain information is associated frame by frame. The feature extracted by AAM and it is a two class classification problem with pain and no pain labels.

The dataset using in this chapter for automatic CLBP recognition is Emo-Pain corpus database [59]. The dataset concludes facial expression data which used to manually make the ground truth labels, EMG signals and body movement which used as training set. There are some researches by Aung et al. [58] with an RF regression to automatically detection of guarding behaviour for patients. Through the existence of guarding behaviour, chronic pain can be ascertained. Data used in [58] include joint angles, joint energy and EMG.

In [58],some different body movements have been studied, like one leg standing and reaching forward. Detection is based on behaviours that include guarding, supporting and hesitating. In Emo-Pain corpus dataset, labels are rated by 4 different raters, and in Aung et al. [58], the labels are added together and normalised in a range of [0, 1] to achieve the aim of regression in RF. Among the behaviours related to CLBP, Aung tested several behaviours with guarding, hesitation and limping.

Recently, Olugbadel [60] used just part of the Emo-Pain corpus dataset with a focus on only one body motion, reaching forward. Instead of using all the features of body motion, Olugbadel used a very limit part of body motion, like neck combined with left arm. This is not a real-time detection system; it is based on a whole instance and uses a twin slide window to analyse all frames of the instance and determine the level of chronic pain of the instance.

Here, we will do further work to build the automatic recognition for CLBP-related behaviour like guarding, one leg standing and moving forward. We will select different subsets from the Emo-Pain Corpus dataset and build the system for continuous recognition of CLBP-related behaviours. Rather than detect the whole video and just one output, our automatic system will recognition behaviours frame by frame.

The machine learning system is based on the continuous classification method of affective emotion level prediction proposed in [152]. In [152], a continuous multi-stage level classification method was proposed based on k-NN in the first stage and an HMM in the later stages. It has been used for facial expression and vocal expression prediction on the frame level with good results. Here, we use k-NN in the first stage and then extend the HMM in the later stage and make it a Two Stage Classification (TSC) scheme. The results were compared with the ones from RF on the frame level.

5.3 Automatic recognition system

5.3.1 System overview

In Emo-Pain dataset, there are 22 subjects. In the test of our system, we divided them into 3 groups and the results would be test by 3-fold cross-validation, which means 2 groups would be used for training and the other one for testing. The system overview is shown in Figure 5.1. Both training set and testing set data would be classify frame by frame according to the ground truth labels rated by 4 raters.

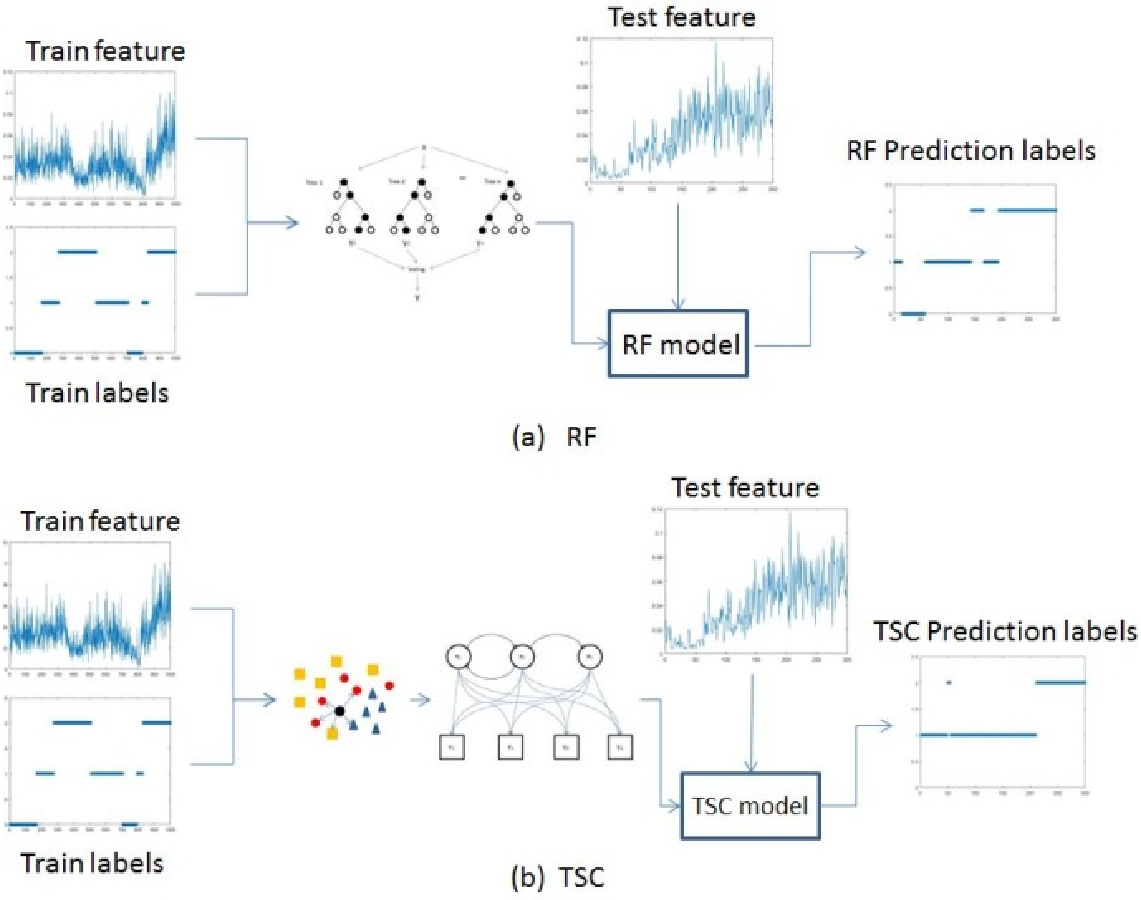


Figure 5.1: Overview of the system (a) RF prediction system; (b) TSC prediction system

5.3.2 Emo-Pain corpus dataset

The Emo-Pain dataset concludes two parts. One is facial expressions, which collected by 8 cameras set around the subjects in experimental site. These videos are watched by several psychologists and given a mark as the level of CLBP. This part of dataset are using as ground truth labels. The second part is EMG signals and motion capture data. The subjects are asked to equip wearable devices which can capture the spatial coordinate and EMG signals. Then, subjects are asked to perform different actions in a 5-meter-long walking and motion capture area. This part of dataset is training set and testing set.

The spatial coordinate of body motion of subjects are captured by a customized Animzaoo IGS-190 suit. There are 18 inertial sensors that are small and solid-state in the wearable device which can measure the rotations of the wearer accurately in real-time. Also, the wearable suit is designed both comfortable to wear and minimally obstructs movement of wearer when measuring data.

A BTS FREEEMG 300 with 4 wireless sensors is used to collect EMG signals. The distance of probes from the receivers could be up to 50 meters, and it concludes a BTS EMG-Analyser. During the data collection, there are 2 probes placed on subjects' paraspinal muscles respectively and 2 probes placed on subjects' trapezius muscles respectively.

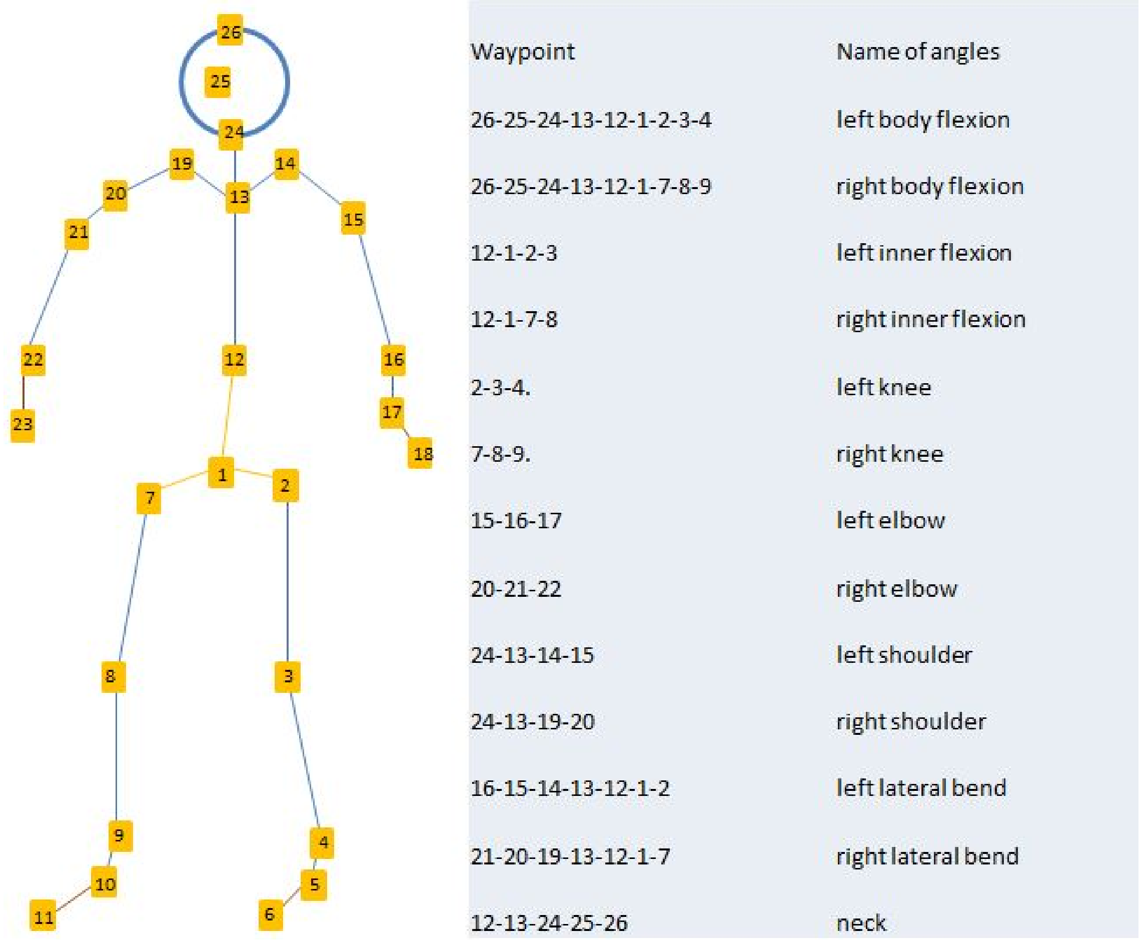
In Emo-Pain corpus dataset, the collected pain-related movement concludes rubbing, guarding, abrupt action, hesitation, limping and supporting. According to the situation of dataset, here abrupt action and guarding are chosen for experiment. According to the definition in [58], abrupt action is a body movement in suddenly and not in control because of extraneous reasons, guarding is a movement rigid or interrupted like preventing the pain point of the body from contacting a chair when sitting down.
There are 4 psychologists watching the facial expressions frame by frame and labelling the level of pain for subjects. Each psychologist holds a special designed remote controller to rate the subjects continuously with a score in the range of [0, 1] while watches the videos. Each score represents a level of CLBP according to the behaviour of CLBP. In the Emo-Pain corpus dataset, the continuous scores have already been translated into labels 0, 1 and 2, in which 0 means no behaviours of CLBP, 1 means low level of behaviours of CLBP and 2 means high level of behaviours of CLBP, respectively.

In each frame of the video, there are four labels which present four scores rated by four different psychologists. A majority voting has been used to combine these four scores into one. The combined one is the label used in experiment. The labels still present as 0, 1 and 2, in which 0 means no behaviours of CLBP, 1 means low level of behaviours of CLBP and 2 means high level of behaviours of CLBP, respectively.

5.3.3 Behaviour annotation

The dataset of body movement and EMG is collected from 22 participants with widely ranging ages and of both genders. Each participant does different exercises several times, including one leg standing, sitting still, reaching forward, sitting to stand, standing to sit, bending and walking.

In the body motion data, 26 points have been tracked and 78 XYZ spatial coordinates have been calculated. According to the spatial coordinates, 13 angles of different body movements have been work out. The names of angles and the position of each point of the angles are shown in Figure 5.2. These values are the feature vector in the machine learning process. At the same time, 4 channels of EMG data are collected.



Waypoint	Name of angles
26-25-24-13-12-1-2-3-4	left body flexion
26-25-24-13-12-1-7-8-9	right body flexion
12-1-2-3	left inner flexion
12-1-7-8	right inner flexion
2-3-4.	left knee
7-8-9.	right knee
15-16-17	left elbow
20-21-22	right elbow
24-13-14-15	left shoulder
24-13-19-20	right shoulder
16-15-14-13-12-1-2	left lateral bend
21-20-19-13-12-1-7	right lateral bend
12-13-24-25-26	neck

Figure 5.2: Features on the body

There are totally 226 features been calculated according to the 78 spatial coordinates and 13 angles. The detailed names and numbers of features are shown in Table 5.1. All of these features and original 78 spatial coordinates and 13 angles are contained in Emo-Pain dataset.

Table 5.1: Name and number of features

Features	Number of Features
Joint Angle	13
Joint Energy	13
Speed (Distance between samples)	49
Trajectory Smoothness (Spectral Arc Length)	49
Trajectory Directness	49
Movement Size (Alpha Volume)	49
EEG	4

5.3.4 RF classifier

In RF, a subset of features in the training dataset is used to determine a node of a decision tree. Then the dataset is sampled several times to constitute a training set, and leftover samples are used as a test set to assess the errors. For every node of the decision tree, randomly chosen features and all the nodes are determined by these features. According to these features, we calculate the best way to produce the decision tree and build a complete tree classification.

The advantages of RF include the high accuracy of classification it provides and the large number of input variables it can process. It can also assess the importance of parameters in determining categories. It can be generalized within the error after not producing an estimate of bias in the construction of the forest. It also is a good way to estimate missing data and can still maintain accuracy when part of the information is lost. RF provides an experimental method to detect variable interactions and can balance error for an unbalanced classified dataset.

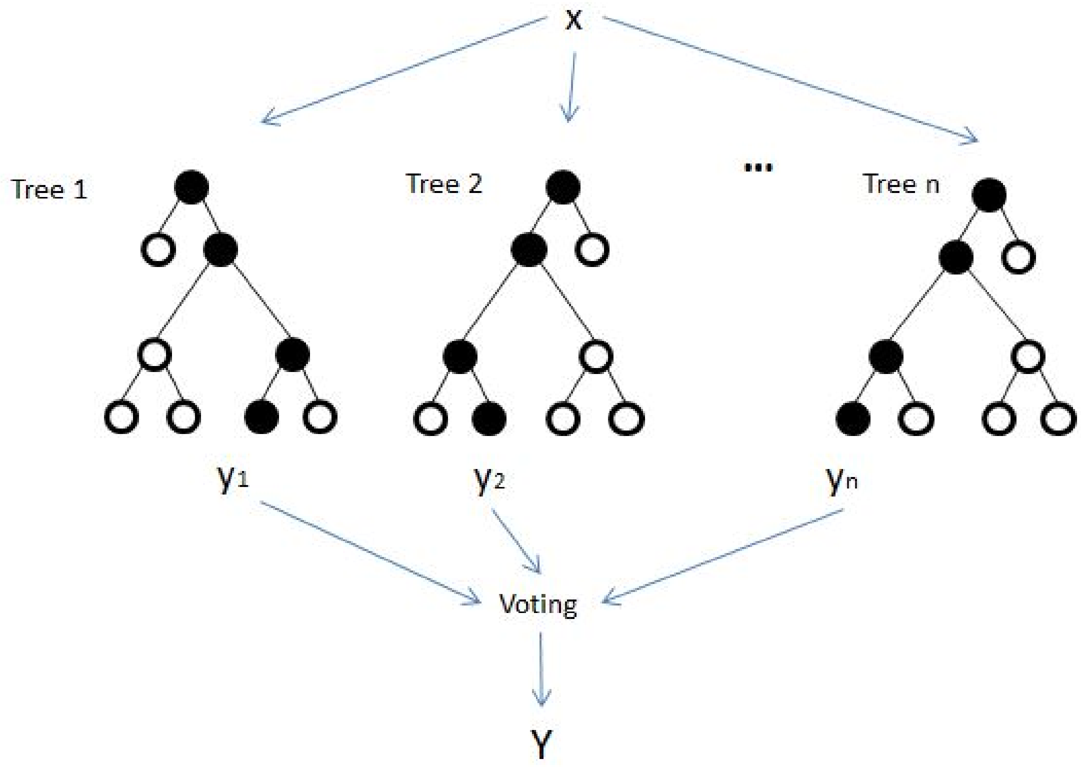


Figure 5.3: Scheme of RF

5.3.5 Two-stage classification (TSC)

As mentioned, this TSC method contains two parts of classification, k-NN and HMM. Firstly, k-NN is used for a prediction, and then HMM is used to determine the relationship between these predictions and to refine the prediction labels of k-NN based on the possibility of state changes.

TSC

There are two stages in our TSC system. In the first stage, a k-NN algorithm is used for all the features in the frame level, and the prediction labels will be generated as decision values. These decision values can be treated as the time series in which the relationship between consecutive frames can be modelled as a Markov process because the behaviour happens slowly.

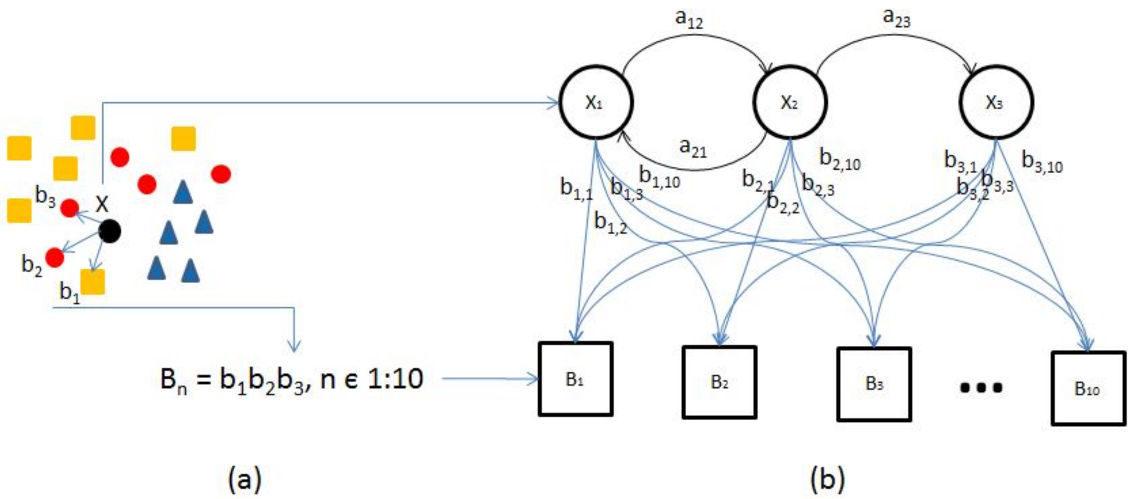


Figure 5.4: Scheme of TSC (a) k-NN prediction; (b) HMM prediction

k-NN

The theory of k-NN has been introduced in literature review. k-NN is a traditional classifier, but here it is not used to classify. Here k-NN is utilized as the first step of TSC. k-NN will calculate the distance of samples and find the closest one in test set in the multiple space learned by train set. Here we calculate the 3 closest points rather than one. Then send these 3 points to HMM.

HMM

HMM is a static model to describe a Markov process that contains implied unknown parameters. The hidden states of HMM are not directly visible, but some variables influenced by states are visible. It will be used in the decision level in the following.

In detail, firstly, we use k=3 k-NN classifier to predict the labels and get the labels of 3 training samples nearest to the testing sample. As we mentioned, there are three states, 0, 1 and 2, in labels. In HMM, these three states are X_1, X_2 and X_3 in the transition matrix. For the 3 nearest labels of the training set, there are 10 situations in total, so the HMM emission matrix is a 3 times 10 matrix, meaning under each X_n (n = 1, 2 or 3), there is the possibility of each situation of 3 nearest samples label. Figure 5.4 shows the whole TSC system.

The matrices below are examples of transition matrix and emission matrix. Equation 5.1 is a transition matrix. Each element in it means a possibility from the previous state changes to the next state. In Figure 5.4 and Equation 5.1, a_{mn} means the previous state is m, the next state is n and the possibility of m changing to n is a_{mn}. Equation 5.1 is the emission matrix. Each element in it means a possibility of combination of the 3 nearest samples labels. For $k = 3$ situation, the 10 situations are 000, 001, 002, 011, 012, 022, 111, 112, 122 and 222. These situations are combined without order. In Figure 5.4, $b_{m,n}$ means the current state is m, the hidden situation is n and $b_{m,n}$ is the possibility of n in state m.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{5.1}$$

5.4 Experimental results

In our experiments, k-NN, TSC and RF methods are compared. There are lots of data in the Emo-Pain corpus dataset, and the behaviour of labels include guarding, supporting, abrupt action and hesitation. But for most of these behaviours, there are too many 0 labels, and classifying them seems pointless. Hence, we choose the part of the data with the most kinds of labels.

As mentioned, we use a 3-fold cross-validation method on the testing. In the Emo-Pain corpus dataset, there are 22 participants. We divided them into three groups, and each time, we used 2 groups as training and 1 group as testing.

To calculate the accuracy of classification, we denote the values as Y_{mn} for numbers of predictions in row m and column n in a confusion matrix. Accuracy can be calculated by the Equation 5.2.

$$Accuracy = \frac{\sum_{n=1}^{3} Y_{nn}}{\sum_{m=1}^{3} \sum_{n=1}^{m} Y_{mn}} \tag{5.2}$$

The classification results are shown in Table II. The classification was done in 7 different situations with the combination of behaviour and exercise. The results of k-NN, RF and TSC are compared.

There are four scores from four raters for each frame. We use majority voting tactics to transfer four scores into one label. The final three classification labels are also 0, 1 and 2, in which 0 means no behaviours of CLBP, 1 means possible behaviours of CLBP and 2

Table 5.2: Prediction accuracy of behaviours of CLBP

Label	Exercise	Classification Method			Length of samples
		k-NN (%)	TSC (%)	RF (%)	
Guarding	One leg stand	48.80	53.93	49.21	14225
Abrupt action	One leg stand	37.58	56.50	53.27	14225
Guarding	Reach forward	37.25	49.79	42.20	25214
Guarding	Sit to stand	42.92	31.52	52.28	4948
Guarding	Stand to sit	45.56	44.70	58.49	5236
Guarding	Sit to stand not instruct	45.14	33.86	45.94	6004
Guarding	Bend to pick up	40.31	42.43	36.41	6037

means subject has behaviours of CLBP, respectively. Specifically, when adding the labels of 4 raters together, the sum is 0 or 1, the final label is 0, the sum is 2, 3 and 4, the final label is 1, the sum is 5 or larger and the final label is 2. The divided classification label is based on the voting principle. When half the raters think it is possible, the final label is considered possible, and when all raters think it is possible and one of them think it is confirmed, or more than half of the raters think it is confirmed, the final label is determined as the subject having this behaviour of CLBP. Thus, the problem becomes a 3-class classification problem.

Table 5.2 shows that the performance of the classification is related to the length of the features (number of the samples in a time series). For short sequences, RF achieved the best results, while for long sequences, TSC achieved the best performance. The HMM model needs to be trained using long sequences because the model can be fully learned only with efficient samples. For the short sequence, RF is a better classifier than k-NN.

HMM gives each state a probability of changing to the next state. So the continuousness of the prediction series is better than that of the k-NN prediction series. That is the reason for the better performance HMM has than k-NN and RF.

Because HMM is a series-based method, it works well only when the series are long enough. When the series are short, the result of HMM is very unstable, and the result of RF is more stable when samples are limited.

It is obvious that when the samples are long enough, more than ten thousand, the correct rate of TSC is higher than that of RF. But when the length of samples decreases to a few thousands, the performance of TSC decreases significantly. In the four groups of data of less than 10 thousand, only one groups result shows TSC is better than that of RF.

To give more detailed information on the classification, Table 5.3 gives the confusion matrix obtained in the classification for guarding action recognized in the exercise reaching forward under the TSC classification method. From this table, it can be seen that the majority of the labels are correctly classified.

Table 5.3: Confusion matrix of prediction

		Predicted Labels (Y_{mn})		
		No Guarding	Possible	Guarding
True Labels	No Guarding	2778	1541	2294
	Possible	2537	5109	3045
	Guarding	1628	1774	4508

Although in [59] and [153] there are results of recognition of pain related behaviour using Emo-pain dataset, but the calculation and recognition methods are totally different. The recognition of [59] and [153] are based on whole subject movement rather than frame by frame recognition introduced in this chapter.

Table 5.4: Comparison with published results on Emo-pain dataset

Author	Method	Movement	Result
Olugbade et al.[153]	SVM	Full trunk flexion	94%
Olugbade et al.[153]	RF	Full trunk flexion	88%
Olugbade et al.[153]	SVM	Sit-to-stand	76%
Olugbade et al.[153]	RF	Sit-to-stand	69%
Aung et al. [59]	RF	Bend	0.019 (mean-squared error)
Aung et al. [59]	RF	Sit-to-stand	0.016 (mean-squared error)
ours	RF	Stand-to-sit	58.49%
ours	TSC	One leg stand	56.50%

5.5 Summary

In this chapter, an automatic CLBP-related behaviour detection system has been built from EEG and motion capture data from the Emo-Pain Corpus database. The automatic recognition system detects chronic pain through some unnatural body gestures like guarding and abrupt actions. This new system can detect CLBP in a very short period of time-one frame is enough. This proves human behaviour can be used in the medical field.

The key point of this method is providing details information of CLBP in one period of time. This knn-hmm method could distinct when the CLBP start and when it is end. Rather than just one output in state-of-the-art, it would help to research on CLBP.

Different machine learning methods have been used in this automatic recognition system. Compared with RF, which has been proven to be a very effective traditional method for classification, the performance of TSC combined with k-NN and HMM has proven to be a better method.

Chapter 6

Facial Expression Recognition on 3D Videos based on Background Substraction and Dynamic ImageNet

6.1 Introduction

Unlike gesture recognition, FER is a more advantageous human behaviour recognition system. Gesture recognition directly recognises movements of hands or the body. FER recognises emotions through movements on the face.

In machine learning, 3D dynamic FER is one of the latest methods. There are many methods of 3D FER, and the basic way involves pre-processing, feature extraction, machine learning and post-processing. The idea is to reduce the dimensions of features and use machine learning to achieve a better result. The key is to extract the most distinct information from the dataset. In my method, I am using a BSCM with Tensor Robust Principal Component Analysis (RPCA) [154] to divide the dynamic part (Sparse) and static part (Low Rank) of the 3D videos. Then, a series of machine learning methods has been used. The most effective is a deep learning method called Dynamic Image Net with a pre-trained VGG19 CNN.

In 3D dynamic FER, two pre-processing methods have been used, Total Variation Regularised Tensor RPCA for BSCM and Robust On-line Matrix Factorisation for Dynamic Background Subtraction, to extract the dynamic information from facial expression videos, including both texture maps and depth maps [154].

6.2 Related works

FER has been researched for years, and 3D FER is one of the latest and most accurate approaches. Most of these results are predicted by deep learning methods. One of the most popular and integrated dataset is the BU-4DFE dataset from [5], which I used in this chapter. In 3D FER, one of the most different problems is that the scale of the dataset is very big and resource calculation is limited. So how to extract the most effective information from large 3D data is the most important pre-processing problem in a 3D FER system.

There are lots of methods to extract dynamic information from video data. [8] has been proven to be an effective method. This method has never been used in the 3D FER area before, so it is worth trying it and testing its effectiveness.

Lots of experiments and researches show deep learning methods are very effective for FER. [155] proved it a very effective method in the BU-4DFE dataset in [156]. Although this method of calculation is expensive, I have used it on the extracted data and proven its effectiveness.

The research gap of 3D FER is the cost of calculation resources and the large run time. Even the performance with ResNet is quite good in state-of-the-art. The equipment is not able to afford to anyone. So, it would be a realistic way to divide dynamic features out and save the computing resources.

6.3 3D dynamic FER system

For 3D dynamic facial recognition, the most distinct features of FER are movements on human faces. The basic idea is to extract the dynamic information on human faces and extract features from the dynamic parts of faces. The common method is to use a video-based feature extraction method like LBPTOP, LPQTOP and MHH or a deep learning method like CNN to process whole videos. But the BU-4DFE dataset [5] is very big, in fact, the largest published dataset for posed FER.

6.3.1 System overview

The traditional 3D feature extraction methods like LBPTOP and LPQTOP are limited because the BU-4DFE dataset is very high solution. High solution has many benefits, the most important one being that it can describe the details on the face more clearly. So the performance of recognition can be better if the machine learning system can extract all the right information and ignore all the disturbances. But the traditional LBPTOP and LPQTOP will extract all the features from videos without any selection. Thus, the performance will decrease when the solution is very high and useful information is very concentrated.

On the other hand, the high solution will make the deep learning process very slow. Many deep learning methods have proved very effective on FER. To increase the speed of the deep learning process, it is necessary to extract the useful information before sending it into the deep learning network. BSCM is used to extract dynamic information and then send into the Dynamic ImageNet (DIN).

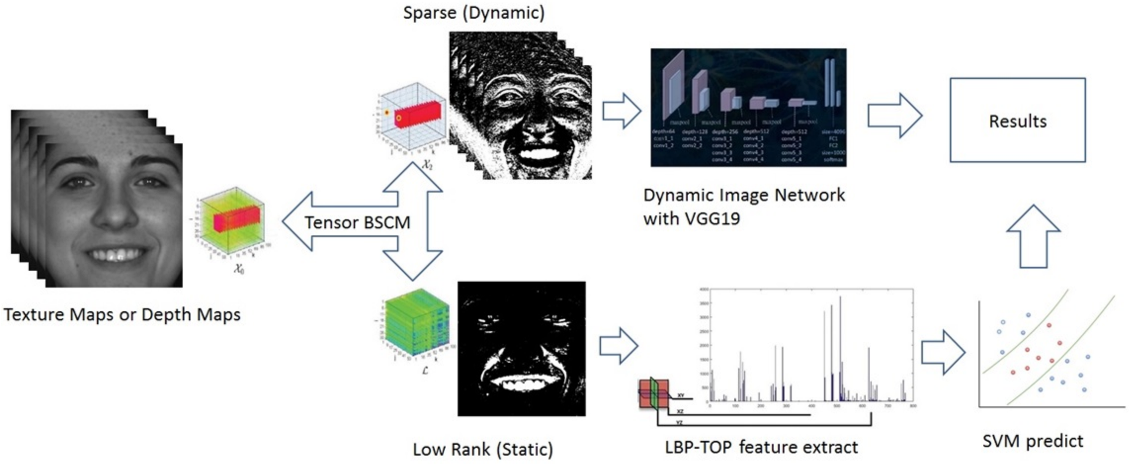


Figure 6.1: Structure of the whole system

6.3.2 Tensor RPCA for BSCM

Background Subtraction is a common method used in image processing. It has been widely used in noise decrease, feature selection and dimension reduction. Tensor RPCA for BSCM is one of the latest methods used on Background Subtraction. It considers the video a tensor, and the video can be divided into two parts, background and foreground. The foreground of the video has spatial-temporal continuity, and the background has spatial-temporal correlation. Thus, we can use a tensor resolved method to divide the video into two parts, low rank and sparse. The first is the dynamic part, and the second is the static part of the video.

Tensor RPCA for BSCM assumes the video tensor has some common characteristics: the video background has a self-similarity; the video foreground has continuously on spatial-temporal, and the video background has correlation on spatial-temporal. With these assumptions, the video volume is considered a 3-order tensor that can be represented as a dynamic component (sparse/foreground) and a static component (low rank/background).

First, we considering a video χ_0 can be divided into two parts, a static background χ_1 and a dynamic foreground χ_2 as shown in Equation 6.1.

$$\chi_0 := \chi_1 + \chi_2 \tag{6.1}$$

Then the original video χ_0 can be represent as Equation 6.2, and so does the χ_1 and χ_2 in Equation 6.3 and Equation 6.4, where i is the frames of videos.

$$\chi_0 := \{X_0^1, X_0^2, \ldots, X_0^D\}, where X_0^i \in \Re^{W\times H} \tag{6.2}$$

$$\chi_1 := \{X_1^1, X_1^2, \ldots, X_1^D\} \tag{6.3}$$

$$\chi_2 := \{X_2^1, X_2^2, \ldots, X_2^D\} \tag{6.4}$$

The vectorization of a video χ_0 is $x_0 := [x_0^1, x_0^2, \ldots, x_0^D]$. The same with χ_1 and χ_2. Then the compressive measurement y can be present as Equation 6.5

$$y = \mathscr{A}(x_0), \mathscr{A} = D \cdot H \cdot P \tag{6.5}$$

In Equation 6.5, D is random down sample operate; H is noise translate and P is random permutation matrix. So the reconstruction of the Video volume can be represent as Equation 6.6.

$$\min_{x_0,x_1,x_2} = \lambda\Omega_0(x_0) + \Omega_1(x_1), x_0 = x_1 + x_2, y = \mathscr{A}(x_0) \tag{6.6}$$

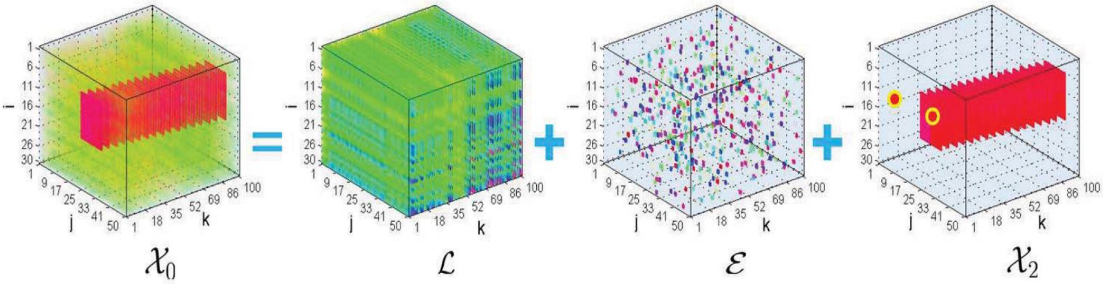


Figure 6.2: Tensor RPCA for BSCM [8]

The video volume χ_0 can be represent 3 parts as the Figure 6.2 and the equation can be re-write as Equation 6.7, where ε is disturbance and $\mathscr{L}$ is low rank component.

$$\chi_0 = \chi_2 + \varepsilon + \mathscr{L} \tag{6.7}$$

6.3.3 DIN

With the separation of sparse and low rank of each facial emotion video volume, the dynamic information is more obvious to recognise for the machine learning component in the system. Also, the static part is an important reference for recognition system. Because these two different parts have different characteristics in feature domain, for the dynamic part, it is wise to use some dynamic feature extraction methods like LBPTOP and Dynamic Image Net in this experiment; for the static part, it is good to use some static methods like LBP and LPQ.

For sparse features, because the facial emotions in the BU-4DFE dataset are posed, all the emotions start with a natural face that then becomes a posed emotional face. Then the faces of subjects revert to natural faces. The whole process lasts for about 4 seconds, with a frame rate of 25. The total frames of the video are about 100. To reduce the influence of natural faces in the videos, the features only extracted the 60% in the middle of the videos, about 60 frames.

For low rank features, the pictures in the video are barely moving and almost the same. It will save lot of time to just extract one frame in the low rank part after BSCM. In the machine learning process of the system, two different methods have been used for sparse features. One is traditional SVM, and the other is Dynamic Image Net, which is a deep learning method using a CNN network to extract the dynamic information from several frames of a video and use the dynamic information through CNN to predict a result.

The first step is to construct a dynamic image. We assume a video can be represented by its frames like $I_1,\ldots,I_T$. So the feature vector extract from videos can be represented as $\psi(I_T) \in \mathbb{R}_d$. In the time sequence t, the average of features can be represented as $V_t = \frac{1}{t}\sum_{\tau=1}^{t}\psi(I_\tau)$. Then a ranking function can be worked out, $S(t \mid d) = \langle d, V_T \rangle$. According to the ranking function, the learning rate d can be presented by Rank SVM:

$$d^* = \rho(I_1, I2, \ldots, I_T; \psi) = \arg\min_d E(d), \tag{6.8}$$

$$E(d) = \frac{\lambda}{2}\|d^2\| + \frac{2}{T(T-1)} \times \sum_{q>t} \max\{0, 1 - S(q \mid d) + S(t \mid d)\} \tag{6.9}$$

6.3.4 Frames selection and post-processing

As a dynamic machine learning system based on video sequences, 3D dynamic FER has the same weaknesses as the 3D micro-gesture recognition described in chapter 4. In the BU-4DFE dataset, at the start and end of each video, there are natural faces in the sequences. So the frame selection method is also very useful for 3D dynamic FER.

The steps are almost the same. The experiment is processed on texture maps, which is more obvious on images. The videos have been extracted frame by frame. Each sample becomes a 100-frame images. All those images are put together as train set. The machine learning method is SVM, and each image learns a score from SVM prediction. Using this score, those images that have almost equal scores on more than 2 categories will be deleted. The images having a distinct score in 6 categories will be retained.

After the frame selection, most frames of natural faces and characteristics are not obvious and have all been filtered. The remaining images are treated with two methods. One is using image processing learning method to process the images. The other method is using video-based machine learning method. The dispersed images are still treated as a video, although some frames will act very strangely. Most of the frames in the middle of a video are still retained.

6.4 Experiments

6.4.1 BU-4DFE dataset

The BU-4DFE dataset was collected by Yins team and published for several years [5]. Currently, it is still the biggest 3D dynamic facial dataset with different races and genders of people. The dataset is collected with one texture camera and two stereo cameras. Therefore, there are two maps in the dataset, depth maps and texture maps. There are 101 subjects in the dataset. Each subject was asked to act out 6 different emotions, anger, disgust, fear, happy, sad, and surprise. Each sample is around 4 to 5 seconds. The collection rate is 25 frames per second. In all, there are 606 subjects, each one having one texture model and one depth model. Through these models, a 3D high-resolution face model has been built. In my experiment, I only use the original texture model and depth model. I did not use the rebuilt 3D model because of the limitation of calculation. Also, all the good results have been achieved using texture maps and depth maps.

The data are collected with a dimensional imaging dynamic face capture system called Di3D. The system can capture both 3D model sequences and 2D texture videos. The system captures the texture videos and depth videos in parallel with two computers. Each subject was asked to sit one and a half meters away from the 3D cameras and was asked to act out 6 different emotions. Each expression contains natural expressions at the start and at the end of a video. The total length of a video is about 100 frames, lasting for about 4 seconds.

6.4.2 Experimental results

The first experiment is made by traditional LBPTOP and LPQTOP on original texture maps and depth maps as a comparison with my other methods. For all the 101 subjects and 6 emotions, the LBPTOP and LPQTOP extract features on the whole 100-frame videos. After features are extracted, a classification learner on MATLAB is used to predict the results and calculate the accuracy. Over 20 machine learning methods have been used to calculate the accuracy. The two most accurate are SVM and subspace discriminant.

With extracting the 60% frames in the middle of the videos, the accuracy is obviously increased because the process deletes the natural faces at the start and end of the video. After using BSCM, two parts of videos, low rank and sparse, are calculated with different machine learning methods. After BSCM, both results have obviously increased. Each step of results are shown in Table 6.1.

Table 6.1: Prediction accuracy for each processing (%)

Data	Texture Maps	Depth Map
LBPTOP in original video	42.2	49.8
LPQTOP in original video	40.3	55.4
LBPTOP in middle 60% video	54.0	55.1
LPQTOP in middle 60% video	63.5	65.5
BSCM of Spares	71.6	72.4
BSCM of Low Rank	63.5	67.5
Dynamic Image Net	**73.3**	**73.8**

Compared with other results in this dataset, my result is not the best and has lots of differences from published results.

Table 6.2: Comparison with state of the art methods

	Method	Accuracy (%)
Cao et al. [8]	AIM ResNet	86.67
Cao et al. [8]	AIM all maps	92.22
Sun et al. [157]	Tracking Vertex Flow and Model Adaptation	94.37
Ours	BSCM and DIN	75.41

From the results in Table 6.2, it is obvious that after twice optimising the system, the results have increased. The idea of deleting the natural faces at the start and end of the video works. And the idea of dividing the dynamic part and static part of faces and calculating them separately also yields good results. In particular, the sparse part, which means the dynamic part of the face, reaches a very high level of accuracy.

From the results, it is obvious that the results of deep learning are much better than those of LBPTOP and LPQTOP. Although the deep learning methods are slower than traditional methods, with more subjects and big data and more time, it is almost certain that deep learning methods can achieve higher results.

Some studies show that by using more of the latest nets like ResNet, the result will be much better. Also, it will need much more calculation resources than VGG19. Also, the pre-processing filtered most of the natural faces and increased the results.

6.5 Summary

Compared with other published results, there is a large gap between my result and the best result. The possible reason is the deep neural nets are different. I am using VGG19 ImageNet, which is almost the limit of my PC with a GTX 980 graphic card. I have tried using ResNet, but the calculation time is too long and I had to give it up halfway through. Also, the VGG19 ImageNet still has room for optimisation. The parameter can be optimised, and training time can be extension.

In any case, this is a test of BSCM used in FER area, and the result reflects the effectiveness of BSCM on FER. The extraction of dynamic information obviously increased the performance of machine learning. Also, extracting the middle of each video to find the typical features of each facial expression makes a great contribution to the results.

In comparing deep learning method and traditional SVM classifier, deep learning method is obviously better. Results in [8] and [157] also prove the advantage of deep learning. In future work, with enough calculation resources, a deep learning method would be an important consideration for FER.

The key point and contribution are to refer a latest dynamic separation method in 3D FER. As 3D FER is already computing source expense and cost lots of time; the value of this BSCM method is great. It could reduce the computing source, run time and increase the classification accuracy at same time.

Chapter 7

Development and Application of a Real-time FER System

7.1 Introduction

Unlike 3D FER in the last chapter, 2D FER in this chapter is more focused on real-time speed and anti-noise performance. This research comes from a project called RIOT, which is an interactive film simulating a riot situation and testing how people will act in a riot to increase the survival rate for normal people in a real riot.

The RIOT project[158] is a HCI system to test the emotional response of people in a riot situation. The system includes an interactive film with a set of stereo sound systems and a real-time FER system. According to the FER results in different clips of the movie, the movie will include different stories and will lead to different endings.

The interactive film simulates a situation of the audience facing an riot. A police officer will have some interaction with the audience in front of the screen. If the audience shows a high level of anger or fear emotion, the police will try to arrest the audience, and the movie will have a bad ending.

There are 4 clips connected with the real-time FER system. Each clips last from 15 seconds to more than 1 minute. When the audience watches these 4 clips, the real-time FER system will detect their faces and recognize their emotions. The real-time recognition system will detect emotions 9 times in one second, and then a majority vote would be applied after the clip. Only if the audience maintains a high level of fear or anger during most of the clip would the simulated police officer try to arrest the test subjects. Here is a link website of RIOT project [158].

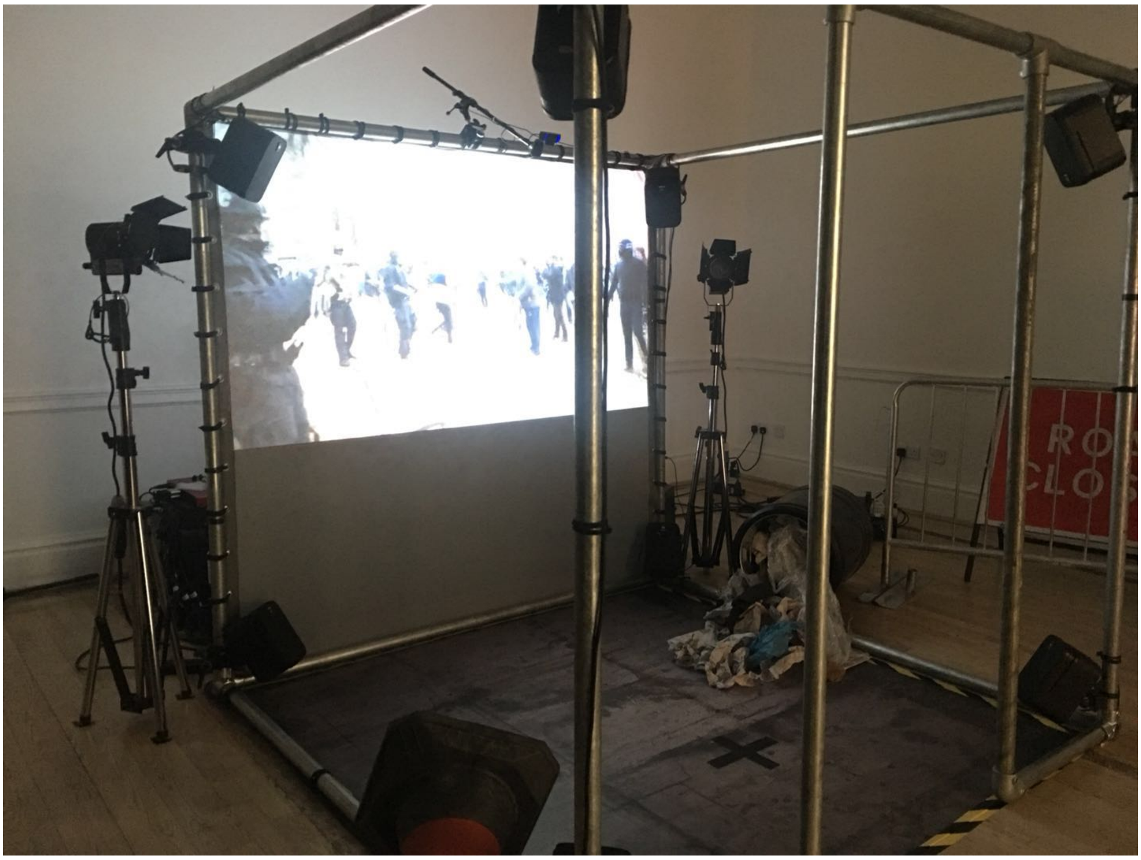


Figure 7.1: Interactive film testing system setting.

7.2 Related works

FER has been developed for years and has been used on lots of applications. Many are used in real time. For instance, [159] use FER as a control system for a music player. There is a system to achieve FER through Kinect proposed by [160]. Smartphone technology also has seen great development recently. [161] and [162] proposed real-time FER systems on smartphones. Those applications in industry areas are good examples of academic knowledge used in real life.

In our situation, a stereo-phony interactive movie is a very advantageous application, and the combination of this integrated entertainment system and the hottest A.I. technology would be a great business gimmick. The success of this system in many exhibitions in the world proves it.

7.3 Real-time FER system

This is a HCI system to test the performance of people in a riot situation. The system includes an interactive movie with a set of stereo sound systems and an emotion detection system. When the movie plays, it will move into different stories based on the emotional feedback of the audience.

We collected images of 20 peoples faces as training data. They were asked to watch the riot movie, and a camera was set up to collect their facial expressions in real time. After they watched the movie, we asked how they felt when they were watching it. Their emotions are used as training labels.

Three labels of subjects has been considered in this interactive movie system: calm, fear and angry. It is a riot 3D sound movie, and the subjects are assumed to be in a riot situation. When the subjects are acting out angry emotions, they will be taken by the police; when the fear emotions are detected from the subjects, they may not arrested by the police on the spot, but they will be queried after the riot ends; and when the subjects stay calm, the police will release them and everyone will go back home safely.

The three labels are real numbers, and the machine learning system is a regression problem. The machine learning recognition system will calculate the emotions of the subjects while they are within the sight of camera. Then by considering all the emotions they show during the specified period of the movie, the machine learning system will give a classification label of angry, fear or calm.

7.3.1 System overview

In the early version of the system, Edge Orientation Histogram (EOH) features are extracted and SVM is used as regression method. In the later version, facial detected and extracted systems are added, which improves the performance of the whole system. Then a brightness control system is added to make sure the whole system can work in darker situations like cinemas. In the final version of the system, CNNs are used as the feature extracted method, which highly increased the performance of the whole system.

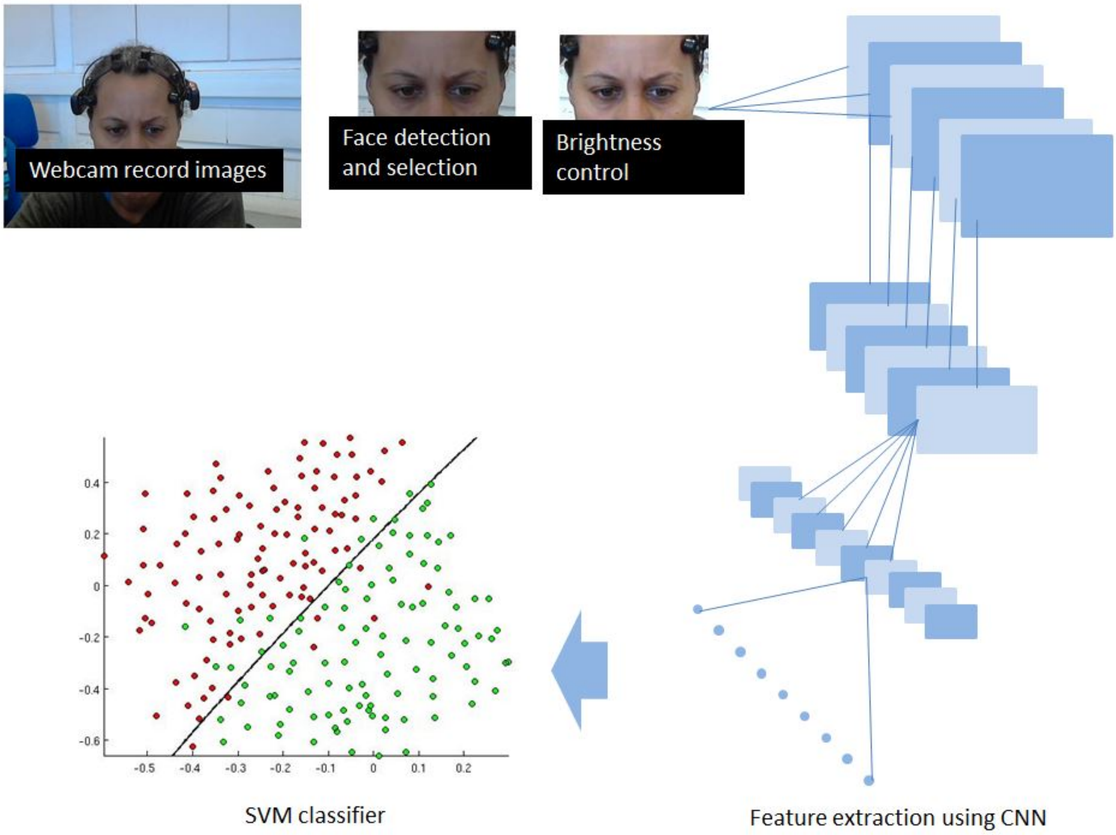


Figure 7.2: Overview of the system

The interactive movie is built on a Java system. The Java system is responsible for the movie fragments playing in the right sequence. Behind the Java system, my FER system is working all the time the movie is playing to record faces and predict the emotions on the faces in real time. The FER system only works when the camera is turned on. The camera will turn on when the movie play the essential 4 fragments.

The first important step in the system is facial recognition. When the camera captures the face, the facial recognition processing can recognize the human face, select the face with a frame and extract the face as an image. This image only contains the human faces and deletes all the other parts of the subjects and the background.

Using these images, a series of machine learning systems has been built. The first step of a machine learning system is feature extraction. In the early version of the interactive movie system, EOH feature has been extracted, and in the last version of the system, deep learning method CNN has been used as a feature extracted method.

After the features have been extracted, a SVM has been used for regression. The first step of the regression system is using the 20-subject database to build a training set. Then, some random persons (excluding the 20 subjects in the train set) have been invited to watch the interactive movie and become the test set. The test set's faces are used as predictions in SVM regression.

In actual exhibition of the interactive movie, the audience members are like a test set. Their faces are extracted by facial recognition processing, and a SVM regression has been used to predict the emotions on their faces. Then the results are sent to the interactive Java system to choose the next movie fragments to play.

7.3.2 Data collection and training

The database aimed to collect three emotions, calm, fear and anger, at different levels so a regression dataset of these three emotions can be built. The labels are levels of these three emotions, numbered from 0 to 1 representing from no fear/anger/calm to extreme fear/anger/calm.

The whole system is an interactive movie. A camera on the top of the screen will collect the facial data in real time. Unlike 3D FER in the last chapter, the facial emotions here are spontaneous. So the database built here is collecting spontaneous emotions on the faces. The subjects are asked to watch the movie fragments from the interactive movie. The fragments last from a few seconds to a few minutes. Then the subjects are asked to record their level of three emotions, ranking them from 0 to 1.

We have collected 20 subjects, including all races, cultures and ages from youths to old people, both male and female. Each subject is asked to watch 4 movie fragments from the interactive movie. In actual testing circumstances, the emotions of testing subjects are detected using the same 4 movie fragments. These 4 movie fragments are selected as the most radical and essential in the movie. If the subjects keep calm in all 4 movie fragments, the movie will keep playing and will have a good ending. If the subjects display a high level of fear or anger, the movie will take a different turn and have a bad ending.

7.3.3 CNN feature extraction and SVM classifier

In the final system, we use CNN as a feature extraction method and use SVM as classifier. comparing with EOH feature extraction, CNN is much better on performance, this will be proved in the experiment in this chapter.

CNN feature extraction

CNN has been developed for many years on human face detection and FER. Many of them have achieved very good accuracy on FER. There are lots of well developed Net in FER, but many of them are too large on size and running too slow. So here we choose a basic ImageNet, as the speed of it is very fast, about 9 frames per second. Also, the accuracy is acceptable.

SVM classifier

SVM is also a very classic classifier and has been used in gesture recognition system in previews chapters. Among all the traditional classifier, the speed of SVM is one of fastest and the accuracy is also very good. Here we choose a basic linear SVM as classifier. Comparing with CNN, the cost of time for SVM can be ignored.

7.3.4 Brightness control and post-processing

Unlike normal system building in the laboratory in the previous 4 chapters, the realistic application system faces lots of issues in actual use. Our interactive movie system has been sent for exhibition all over the world. The first problem is light problems. Some exhibition rooms are illuminated adequately, and some are dark. Sometimes the exhibition is in the morning, sometimes in the afternoon. And the sunlight is different at different times and different places.

So a brightness control system was built to adjust the illumination. The basic idea of a brightness control system is to average the light in the faces of subjects. Based on the average light, some faces that are too dark will be brightened, and some faces that are too bright will be dimmed. Without brightness control system, the facial recognition system may lose the face when it is too dark or too bright.

Another problem in the actual use is there may be too many faces in front of the camera. As the interactive movie system has been exhibited, the audience may stand in a circle, and the camera usually captures lots of faces and the facial recognition system will extract all the faces in the image.

Therefore, a facial selection processing has been built. Because the main audience usually stands closest to the screen, the images extracted by the facial recognition system will be the largest. Thus, a judgement program has been added after facial recognition. Only the largest images will be sent to the machine learning system.

Another problem is speed. It is fast enough for EOH feature extraction and SVM regression system. But with upgrade to a deep learning CNN feature extraction, the speed decreases. Therefore, instead of using more accurate CNN like VGG19 and ResNet, a simplest CNN has been chosen as feature extraction. The speed is about 11 frames per second, which is fast enough for the interactive movie.

To increase the accuracy of the system, a majority voting process has been used after all the regression has been made by SVM. Playing the movie fragments takes anywhere from dozens of seconds to a few minutes. In all this time, all the regression results will be added together. At the end of the movie fragments, the added results are the final results. The highest of three emotions will be considered the emotion of the subjects watching this fragment of the movie.

7.4 Evaluation

Although the accuracy of the RIOT project is not measured by a percentage, we still calculate an accuracy in the lab using the 20 subjects' data. The experiment is subject-independent. As in the exhibition of the RIOT system, the result is calculated by majority voting over a long period of time. The performance is much better than just one time test.

All the results are made by 10 cross-validations. As mentioned, in the actual use of this system, a majority voting system has been added, so the accuracy is much better than from single testing in the laboratory experiment. The real-time FER system will detect the faces and predict emotions 9 times in each second. Each interactive video clip last from ten seconds to more than 1 minute.

Table 7.1: Classification accuracy and frames per second

Method using in the system	Performance (recognition rate of the emotions %)	frames per second
EOH feature extraction and SVM regression	58.01	42
EOH feature extraction, brightness control, and SVM regression	61.44	40
CNN feature extraction, brightness control and SVM regression	**80.30**	9

The Table 7.1 shows the frames processed per second in different systems. it could be easy see even the CNN decrease the speed to 9 frames per second, it still satisfies the requirement of real-time.

7.5 Summary

The interactive movie system is an actually engineering application, which is much different from any other system I have built in the laboratory. There are many actual problems when building this system that I have never faced in the laboratory. It is a very good experience working on a real industry project after conducting lots of experiments in the laboratory.

Chapter 8

Conclusion and Future Works

8.1 Conclusion

The research aim of automatic human behaviour recognition is to explore the application of HCI. As HCI and A.I. technology develop, more and more applications are being manufactured. One of the most important steps of A.I. technology is HCI, which aims to automatically recognize human behaviour. Exploring automatic human dynamic behaviour recognition gives us some effective approaches to HCI. If the computer can understand emotions, health and other human states more quickly and in more detail, it will helps humans more. It is also a good tool for entertainment to use automatic human behaviour recognition.

Human behaviour is a very massive object to research. Here I choose gestures and facial expressions as my research object because these features are most commonly used detections and objectives in HCI systems. There are already some applications and manufacturers using FER and gesture recognition. After learning from existing applications and some of the most advantageous technology in A.I., I have built some of my own automatic recognition systems in facial expression and gesture. Furthermore, I have tried to build my own applications to use in real life.

Of all my research, social touch gesture recognition has cost me the most time and required the most detail. The main conclusion is using multiple features and decision-level fusion to achieve a versatility system could be used on both touch gesture dataset. And it would be more suitable to utilise on other kinds of touch gesture recognition dataset than other methods in state-of-the-art. Touch gestures can be widely used in any surface control system like the most popular application in the world, smartphones. Also, some productivity tools like tablets and computers with touch screens have huge demands for automatic recognition of touch gestures.

In social touch gesture research, I have built an automatic recognition system without using deep learning methods. Part of the reason is that deep learning methods may not yield the best results. Part of the reason is that traditional methods save more calculation resources and are faster in calculating. Also, my system has used both datasets. 3D-CNN can achieve good results in the HAART dataset, but cannot be used on the CoST dataset because the frames of the CoST dataset are not the same and 3D-CNN can only used in the same scale of dataset. in my opinion, a system with more versatility can be more easily used in real life applications and can be more easily transformed into industrial applications.

Another hand gesture recognition system is 3D micro-gesture recognition. There are already lots of applications and experiments on vehicle control systems because when one is driving a car, it is hard to have a great scale of movement to control other functions. A micro-gesture controller solves this problem very well. As AR and VR technologies develop, the demand for wearable devices will increase. It is convenient to use micro-gestures as a controller rather than traditional handles or keyboards when subjects manipulate wearable devices.

In the automatic 3D micro-gesture recognition system, I have changed the high-resolution H3D micro-gesture videos to 2D low resolution videos to more easily operate on the data. The results proved that reducing resolution operation did not affect recognition accuracy, and in contrast, it increased performance. In feature extraction, LPQTOP extracts the quantized phase, which has proven more effective on image processing in many cases. Also considering the non-linear characteristic of the feature, non-linear SVM has been chosen as a classifier and achieves a better result than linear SVM. The main conclusion of 3D micro-gesture recognition is to achieve a better result than state-of-the-art by using a low-cost method with considering the characteristics of the features.

Recognition of body gestures for CLBP detection is a very good example of A.I. technology that can benefit mankind. A.I. technology can make great progress in medical fields. CLBP is very hard to detect in our daily lives, and almost everyone at some point will experience some level of CLBP. Thus, detecting CLBP becomes rather important for everyone, especially for elderly people.

In my automatic recognition system for behaviours of CLBP, I used a new TSC classification method and tested its performance. The performance of TSC is better than that of traditional RF, which I have used many times as a good traditional classifier. My system differs from others because it can predict CLBP frame by frame. It uses an HMM to predict the next frame's state, so it is not a real-time system. But when used in real life, and adding a slide window to detect and remember the states in the window, it can be easily modified to a real-time system.

The main conclusion of this body movement recognition system is to detect pain-related body behaviour frame by frame. It would provide more details in research of CLBP, like when the CLBP start and when it is end.

FER has been one of the forefronts of research in A.I. technology. 3D FER is the most popular direction of development in FER, and there are already lots of studies on it. Compared with 2D FER, 3D FER is more accurate and has more promising applications. Furthermore, 3D faces produce better performance in security, and 3D data have stronger noise immunity.

In the BU-4DFE dataset, I verified the effectiveness of dynamic information extraction by BSCM on FER. BSCM has been used in dynamic separation in many situations. I have used it in my 3D FER system, and the performance has increased but is still not as good as many results in the BU-4DFE dataset. In my opinion, the reason is the deep learning method I am using is not as good as theirs. But compared with my system with and without BSCM, it is obvious that BSCM increased the performance of 3D FER recognition.

The main conclusion of 3D FER system is applied latest dynamic separation method to reduce the computing sources and run time. It also would increase the performance of recognition.

Real-time FER is an actual application in cooperation with industry partners. The real-time FER system uses deep learning methods to improve recognition accuracy. Many measures have been applied like face selection and brightness control that have never been used in experimental environments. It is a good example of how requirements in industry situations are different from those in the laboratory.

The main conclusion of real-time FER is to build a whole balanced system to recognise FER in real-time by considering the performance and running time at same time.

8.2 Future works

In the future, the automatic social touch recognition system can be modified to achieve more accurate and faster recognition speed. To achieve the aim of more accuracy, some simple CNNs like a small scale ImageNet can be applied in real-time FER systems. A simple CNN will not delay the speed very much and will increase the recognition accuracy. Another method to achieve the aim is to apply a more effective fusion in post-processing. There are enough features and predicted results, but the fusion method is not effective enough to achieve the theoretical maximum accuracy for all these features.

In 3D micro-gesture recognition system building processing, I only used the 2D information of the H3D micro-gesture dataset. In future work, I am considering using 3D information of the dataset and increasing the recognition accuracy. I will also try to use some deep learning methods because deep learning has been proven to be a very effective way to solve recognition problems. I have used some deep learning methods in FER, and I will try to use them in gesture recognition.

The automatic recognition system for behaviour of CLBP is not very consummated. Two aspects need to be reinforced. One is recognition accuracy, and the other is to modify it to a real-time system. To increase recognition accuracy, more samples need to be collected. With a large sample, deep learning methods can be applied and recognition accuracy should be increased.

In future work of 3D FER, I plan to improve my deep learning algorithm and change the Net I am using. Because the development of deep learning is very fast now, there may be some new tools, including the Net. I will try to use them in my system. Another improvement is the advance of dynamic separation method. There have been lots of other dynamic separation methods, and many of them have never been used in FER areas or gesture recognition. If I identify the right method, it would improve the performance of recognition.

In future work, the real-time FER system can be combined with EEG signals to improve the interactive film. There are also lots of methods I used in other chapters that can improve recognition accuracy, including dynamic separation in pre-processing and HC in post-processing.

Bibliography

[1] Bart W Koes, Maurits van Tulder, Chung-Wei Christine Lin, Luciana G Macedo, James McAuley, and Chris Maher. An updated overview of clinical guidelines for the management of non-specific low back pain in primary care. *European Spine Journal*, 19(12):2075–2094, 2010.

[2] Robert Plutchik. *The emotions*. University Press of America, 1991.

[3] Anna Flagg, Diane Tam, Karon MacLean, and Robert Flagg. Conductive fur sensing for a gesture-aware furry robot. In *Haptics Symposium 2012(HAPTICS)*, pages 99–104. IEEE, 2012.

[4] Yi Liu, Hongying Meng, Mohammad Rafiq Swash, Yona Falinie A Gaus, and Rui Qin. Holoscopic 3d micro-gesture database for wearable device interaction. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 802–807. IEEE, 2018.

[5] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *8th IEEE International Conference on Automatic Face Gesture Recognition, 2008(FG08)*, pages 1–6. IEEE, 2008.

[6] Yang Xiao, Zhiguo Cao, Li Wang, and Tao Li. Local phase quantization plus: A principled method for embedding local phase quantization into fisher vector for blurred image recognition. *Information Sciences*, 420:77–95, 2017.

[7] Stuart J Russell and Peter Norvig. Artificial intelligence: a modern approach (international edition). 2002.

[8] Wenfei Cao, Yao Wang, Jian Sun, Deyu Meng, Can Yang, Andrzej Cichocki, and Zongben Xu. Total variation regularized tensor rpca for background subtraction from compressive measurements. *IEEE Transactions on Image Processing*, 25(9):4075–4090, 2016.

[9] Viet-Cuong Ta, Wafa Johal, Maxime Portaz, Eric Castelli, and Dominique Vaufreydaz. The grenoble system for the social touch challenge at icmi 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 391–398. ACM, 2015.

[10] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

[11] Jiaying Liu, Xiangjie Kong, Feng Xia, Xiaomei Bai, Lei Wang, Qing Qing, and Ivan Lee. Artificial intelligence in the 21st century. *IEEE Access*, PP:1–1, 03 2018.

[12] David Lynton Poole, Alan K Mackworth, and Randy Goebel. *Computational intelligence: a logical approach*, volume 1. Oxford University Press New York, 1998.

[13] K Meena and R Sivakumar. *Human-Computer Interaction*. PHI Learning Pvt. Ltd., 2014.

[14] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[15] Alex Alland. *Evolution and human behaviour: an introduction to Darwinian anthropology*. Routledge, 2012.

[16] David MP Jacoby, Edward J Brooks, Darren P Croft, and David W Sims. Developing a deeper understanding of animal movements and spatial dynamics through novel application of network analyses. *Methods in Ecology and Evolution*, 3(3):574–583, 2012.

[17] William Ashby. *Design for a brain: The origin of adaptive behaviour*. Springer Science & Business Media, 2013.

[18] Clayton Hickey and Wieske van Zoest. Reward-associated stimuli capture the eyes in spite of strategic attentional set. *Vision Research*, 92:67–74, 2013.

[19] Beatrice De Gelder. Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3):242, 2006.

[20] Jonathan Chang, Karon MacLean, and Steve Yohanan. Gesture recognition in the haptic creature. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pages 385–391. Springer, 2010.

[21] Steve Yohanan and Karon E MacLean. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics*, 4(2):163–180, 2012.

[22] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong. " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *Lrec*, 2004.

[23] Mark Onsager. Understanding the importance of non-verbal communication. *Body Language Dictionary*, 2014.

[24] Yuanyuan Gu, Xiaoqin Mai, and Yue-jia Luo. Do bodily expressions compete with facial expressions? time course of integration of emotional signals from the face and the body. *PLoS One*, 8(7):e66762, 2013.

[25] ME Kret, Swann Pichon, Julie Grèzes, and Béatrice de Gelder. Similarities and differences in perceiving threat from dynamic faces and bodies. an fmri study. *Neuroimage*, 54(2):1755–1762, 2011.

[26] Carl Therrien. Inspecting video game historiography through critical lens: Etymology of the first-person shooter genre. *Game Studies*, 15(2), 2015.

[27] Seth D Baum, Ben Goertzel, and Ted G Goertzel. How long until human-level ai? results from an expert assessment. *Technological Forecasting and Social Change*, 78(1):185–195, 2011.

[28] Saad Albawi, Fatih Yetkin, and Kerem Altun. Social touch gesture recognition using deep neural networks.

[29] Nan Zhou and Jun Du. Recognition of social touch gestures using 3d convolutional neural networks. In *Chinese Conference on Pattern Recognition*, pages 164–173. Springer, 2016.

[30] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[31] Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[32] Matthias Rehm, Nikolaus Bee, and Elisabeth André. Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1*, pages 13–22. British Computer Society, 2008.

[33] Vladimir I Pavlovic, Rajeev Sharma, and Thomas S Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):677–695, 1997.

[34] Craig Villamor, Dan Willis, and Luke Wroblewski. Touch gesture reference guide. *Touch Gesture Reference Guide*, 2010.

[35] Merel M Jung, Xi Laura Cang, Mannes Poel, and Karon E MacLean. Touch challenge'15: Recognizing social touch gestures. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 387–390. ACM, 2015.

[36] John Greer Elias, Wayne Carl Westerman, and Myra Mary Haggerty. Multi-touch gesture dictionary, November 23 2010. US Patent 7,840,912.

[37] Bill Buxton. Multitouch overview, 2008.

[38] B Stumpe. A new principle for xy touch screen. 1977.

[39] Geoff Walker. A review of technologies for sensing contact location on the surface of a display. *Journal of the Society for Information Display*, 20(8):413–440, 2012.

[40] Dimitri Kanevsky, Roberto Sicconi, and Mahesh Viswanathan. Touch gesture based interface for motor vehicle, November 13 2007. US Patent 7,295,904.

[41] Dylan TX Zhou, Tiger TG Zhou, and Andrew HB Zhou. Wearable augmented reality eyeglass communication device including mobile phone and mobile computing via virtual touch screen gesture control and neuron command, October 6 2015. US Patent 9,153,074.

[42] Heather Knight, Robert Toscano, Walter D Stiehl, Angela Chang, Yi Wang, and Cynthia Breazeal. Real-time social touch gesture recognition for sensate robots. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3715–3720. IEEE, 2009.

[43] Angela Mahr, Christoph Endres, Christian Müller, and Tanja Schneeberger. Determining human-centered parameters of ergonomic micro-gesture interaction for drivers using the theater approach. In *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 151–158. ACM, 2011.

[44] Benoît Martin. Virhkey: a virtual hyperbolic keyboard with gesture interaction and visual feedback for mobile devices. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pages 99–106. ACM, 2005.

[45] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)*, 35(4):142, 2016.

[46] Leonardo Angelini, Francesco Carrino, Stefano Carrino, Maurizio Caon, Denis Lalanne, Omar Abou Khaled, and Elena Mugellini. Opportunistic synergy: a classifier fusion engine for micro-gesture recognition. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 30–37. ACM, 2013.

[47] Robert Neßelrath, Mohammad Mehdi Moniri, and Michael Feld. Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions. In *2016 12th International Conference on Intelligent Environments(IE)*, pages 190–193. IEEE, 2016.

[48] Sonja Rümelin, Chadly Marouane, and Andreas Butz. Free-hand pointing for identification and interaction with distant objects. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 40–47. ACM, 2013.

[49] Renate Hauslschmid, Benjamin Menrad, and Andreas Butz. Freehand vs. micro gestures in the car: Driving performance and user experience. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 159–160. IEEE, 2015.

[50] Weizhe Zhang, Weidong Zhang, and Jie Shao. Classification of holoscopic 3d micro-gesture images and videos. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG2018)*, pages 815–818. IEEE, 2018.

[51] Tao Lei, Xiaohong Jia, Yuxiao Zhang, Yanning Zhang, Xuhui Su, and Shigang Liu. Holoscopic 3d micro-gesture recognition based on fast preprocessing and deep learning techniques. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 795–801. IEEE, 2018.

[52] Ray L Birdwhistell. *Kinesics and context: Essays on body motion communication*. University of Pennsylvania press, 2010.

[53] Ray L Birdwhistell. *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. Department of State, Foreign Service Institute, 1952.

[54] Jeffrey J Jacobsen and Stephen A Pombo. Wireless hands-free computing headset with detachable accessories controllable by motion, body gesture and/or vocal commands, October 7 2014. US Patent 8,855,719.

[55] Liwei Chan, Chi-Hao Hsieh, Yi-Ling Chen, Shuo Yang, Da-Yuan Huang, Rong-Hao Liang, and Bing-Yu Chen. Cyclops: Wearable and single-piece full-body gesture input devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3001–3009. ACM, 2015.

[56] Wei-Po Lee, Che Kaoli, and Jhih-Yuan Huang. A smart tv system with body-gesture control, tag-based rating and context-aware recommendation. *Knowledge-Based Systems*, 56:167–178, 2014.

[57] Johan WS Vlaeyen, Ank MJ Kole-Snijders, Ruben GB Boeren, and H Van Eek. Fear of movement/(re) injury in chronic low back pain and its relation to behavioral performance. *Pain*, 62(3):363–372, 1995.

[58] Min SH Aung, A Singh, SL Lim, A CdC Williams, P Watson, and Nadia Bianchi-Berthouze. Automatic recognition of protective behaviour in chronic pain rehabilitation. ACM, 2013.

[59] Min Aung, Sebastian Kaltwang, Nick Tyler, Paul Watson, Amanda Williams, Maja Pantic, Nadia Berthouze, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, et al. The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. *IEEE transactions on affective computing*, (1):1–1, 2016.

[60] Temitayo A Olugbade, MS Aung, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Amanda C Williams. Bi-modal detection of painful reaching for chronic pain rehabilitation systems. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 455–458. ACM, 2014.

[61] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349, 2003.

[62] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[63] Christa Harstall and M Ospina. How prevalent is chronic pain. *Pain clinical updates*, 11(2):1–4, 2003.

[64] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, pages 45–60, 1999.

[65] Chris J Main and Chris C Spanswick. *Pain management: an interdisciplinary approach*. Elsevier Health Sciences, 2000.

[66] Roger Chou, Judith A Turner, Emily B Devine, Ryan N Hansen, Sean D Sullivan, Ian Blazina, Tracy Dana, Christina Bougatsos, and Richard A Deyo. The effectiveness and risks of long-term opioid therapy for chronic pain: a systematic review for a national institutes of health pathways to prevention workshop. *Annals of internal medicine*, 162(4):276–286, 2015.

[67] David Tauben. Nonopioid medications for pain. *Physical Medicine and Rehabilitation Clinics*, 26(2):219–248, 2015.

[68] P Welsch, C Sommer, M Schiltenwolf, and W Häuser. Opioids in chronic noncancer pain–are opioids superior to nonopioid analgesics? 2015.

[69] Vladimir Hachinski, Costantino Iadecola, Ron C Petersen, Monique M Breteler, David L Nyenhuis, Sandra E Black, William J Powers, Charles DeCarli, Jose G Merino, Raj N Kalaria, et al. National institute of neurological disorders and stroke–canadian stroke network vascular cognitive impairment harmonization standards. *Stroke*, 37(9):2220–2241, 2006.

[70] Laxmaiah Manchikanti, Vijay Singh, Sukdeb Datta, Steven P Cohen, and Joshua A Hirsch. Comprehensive review of epidemiology, scope, and impact of spinal pain. *Pain physician*, 12(4):E35–70, 2009.

[71] CL Lisetti. Affective computing, 1998.

[72] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[73] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[74] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[75] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[76] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[77] DC Turk and A Okifuji. Pain terms and taxonomies of pain in: Loeser jd, ed. bonicas management of pain, 2001.

[78] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[79] A Freitas-Magalhães. Facial expression of emotion. 2012.

[80] Carl-Herman Hjortsjö. *Man's face and mimic language*. Studen litteratur, 1969.

[81] Marco Del Giudice and Livia Colle. Differences between children and adults in the recognition of enjoyment smiles. *Developmental psychology*, 43(3):796, 2007.

[82] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[83] Jun Wang, Lijun Yin, Xiaozhou Wei, and Yi Sun. 3d facial expression recognition based on primitive surface feature distribution. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1399–1406. IEEE, 2006.

[84] Lijun Yin, Xiaozhou Wei, Peter Longo, and Abhinesh Bhuvanesh. Analyzing facial expressions using intensity-variant 3d data for human computer interaction. In *2006 18th International Conference on Pattern Recognition (ICPR2006)*, volume 1, pages 1248–1251. IEEE, 2006.

[85] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on Automatic face and gesture recognition (FGR 2006)*, pages 211–216. IEEE, 2006.

[86] Yi Sun, Michael Reale, and Lijun Yin. Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. In *8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08)*, pages 1–8. IEEE, 2008.

[87] Hoda Mohammadzade and Dimitrios Hatzinakos. Projection into expression subspaces for face recognition from single sample per person. *IEEE Transactions on Affective Computing*, 4(1):69–82, 2013.

[88] Stephen Moore and Richard Bowden. Local binary patterns for multi-view facial expression recognition. *Computer vision and image understanding*, 115(4):541–558, 2011.

[89] Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE transactions on cybernetics*, 45(8):1499–1510, 2015.

[90] Mingli Song, Dacheng Tao, Zicheng Liu, Xuelong Li, and Mengchu Zhou. Image ratio features for facial expression recognition application. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(3):779–788, 2010.

[91] Ligang Zhang and Dian Tjondronegoro. Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, 2(4):219–229, 2011.

[92] Cynthia M Whissell. The dictionary of affect in language. In *The measurement of emotions*, pages 113–131. Elsevier, 1989.

[93] Jinni Harrigan, Robert Rosenthal, and Klaus Scherer. *New handbook of methods in nonverbal behavior research.* Oxford University Press, 2008.

[94] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[95] A Lanitis, CJ Taylor, TF Cootes, and T Ahmed. Automatic interpretation of human faces and hand gestures using flexible models. In *International Workshop on Automatic Face-and Gesture-Recognition*. Citeseer, 1995.

[96] Irwin Sobel. History and definition of the sobel operator. *Retrieved from the World Wide Web*, 2014.

[97] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[98] Hongying Meng and Nick Pears. Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, 30(12):1049–1058, 2009.

[99] Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):699–714, 2005.

[100] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing lower face action units for facial expression analysis. In *Proceedings fourth ieee international conference on Automatic face and gesture recognition*, pages 484–490. IEEE, 2000.

[101] Chi Ho Chan, Josef Kittler, Norman Poh, Timo Ahonen, and Matti Pietikäinen. (multiscale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 633–640. IEEE, 2009.

[102] Hongying Meng, Nick Pears, Michael Freeman, and Chris Bailey. Motion history histograms for human action recognition. In *Embedded Computer Vision*, pages 139–162. Springer, 2009.

[103] Charles Van Loan. *Computational frameworks for the fast Fourier transform*, volume 10. Siam, 1992.

[104] W Morven Gentleman and Gordon Sande. Fast fourier transforms: for fun and profit. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 563–578. ACM, 1966.

[105] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[106] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[107] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[108] Timothy F Cootes, Chris J Taylor, et al. Statistical models of appearance for computer vision, 2004.

[109] Rhodri H Davies, Carole J Twining, P Daniel Allen, Tim F Cootes, and Christopher J Taylor. Shape discrimination in the hippocampus using an mdl model. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 38–50. Springer, 2003.

[110] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[111] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[112] Christian Martin, Uwe Werner, and Horst-Michael Gross. A real-time facial expression recognition system based on active appearance models using gray images

and edge images. In *8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08)*, pages 1–6. IEEE, 2008.

[113] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009.

[114] Akshay Asthana, Jason Saragih, Michael Wagner, and Roland Goecke. Evaluating aam fitting methods for facial expression recognition. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, pages 1–8. Citeseer, 2009.

[115] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006.

[116] Ian T Jolliffe. Mathematical and statistical properties of population principal components. *Principal Component Analysis*, pages 10–28, 2002.

[117] Aleix M Martínez and Avinash C Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):228–233, 2001.

[118] Sung-Kwun Oh, Sung-Hoon Yoo, and Witold Pedrycz. Design of face recognition algorithm using pca-lda combined for hybrid data pre-processing and polynomial-based rbf neural networks: Design and its application. *Expert Systems with Applications*, 40(5):1451–1466, 2013.

[119] Benjamin Auffarth, Maite López, and Jesús Cerquides. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *Industrial Conference on Data Mining*, pages 248–262. Springer, 2010.

[120] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[121] Tin Kam Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE, 1995.

[122] Iñigo Barandiaran. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 1998.

[123] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[124] Federico Castanedo. A review of data fusion techniques. *The Scientific World Journal*, 2013, 2013.

[125] AD Craig. Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, 13(4):500–505, 2003.

[126] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[127] Yona Falinie A Gaus, Temitayo Olugbade, Asim Jan, Rui Qin, Jingxin Liu, Fan Zhang, Hongying Meng, and Nadia Bianchi-Berthouze. Social touch gesture recognition using random forest and boosting on distinct feature sets. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 399–406. ACM, 2015.

[128] Merel M Jung, Mannes Poel, Ronald Poppe, and Dirk KJ Heylen. Automatic recognition of touch gestures in the corpus of social touch. *Journal on multimodal user interfaces*, 11(1):81–96, 2017.

[129] Lito Kriara, Matthew Alsup, Giorgio Corbellini, Matthew Trotter, Joshua D Griffin, and Stefan Mangold. Rfid shakables: Pairing radio-frequency identification tags with the help of gesture recognition. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 327–332. ACM, 2013.

[130] Yongpan Zou, Jiang Xiao, Jinsong Han, Kaishun Wu, Yun Li, and Lionel M Ni. Grfid: A device-free rfid-based gesture recognition system. *IEEE Transactions on Mobile Computing*, 16(2):381–393, 2017.

[131] Hongyi Liu and Lihui Wang. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 2017.

[132] Jun Wan, Guodong Guo, and Stan Z Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1626–1639, 2016.

[133] Dong-Chen He and Li Wang. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, 28(4):509–512, 1990.

[134] Riccardo Mattivi and Ling Shao. Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns*, pages 740–747. Springer, 2009.

[135] Timur R Almaev and Michel F Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361. IEEE, 2013.

[136] Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh, and Soo-Young Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434. ACM, 2015.

[137] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[138] Dana Hughes, Alon Krauthammer, and Nikolaus Correll. Recognizing social touch gestures using recurrent and convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2315–2321. IEEE, 2017.

[139] Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 314–321. IEEE, 2011.

[140] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 878–883. IEEE, 2011.

[141] Juhani Päivärinta, Esa Rahtu, and Janne Heikkilä. Volume local phase quantization for blur-insensitive dynamic texture classification. In *Scandinavian Conference on Image Analysis*, pages 360–369. Springer, 2011.

[142] Garima Sharma, Shreyank Jyoti, and Abhinav Dhall. Hybrid neural networks based approach for holoscopic micro-gesture recognition in images and videos. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 808–814. IEEE, 2018.

[143] Samy S Abu Naser and Rami M AlDahdooh. Lower back pain expert system diagnosis and treatment. *Journal of Multidisciplinary Engineering Science Studies (JMESS)*, 2(4):441–446, 2016.

[144] Daniel Steffens, Chris G Maher, Leani SM Pereira, Matthew L Stevens, Vinicius C Oliveira, Meredith Chapple, Luci F Teixeira-Salmela, and Mark J Hancock. Prevention of low back pain: a systematic review and meta-analysis. *JAMA internal medicine*, 176(2):199–208, 2016.

[145] Damian Hoy, Christopher Bain, Gail Williams, Lyn March, Peter Brooks, Fiona Blyth, Anthony Woolf, Theo Vos, and Rachelle Buchbinder. A systematic review

of the global prevalence of low back pain. *Arthritis & Rheumatism*, 64(6):2028–2037, 2012.

[146] Francis J Keefe and Andrew R Block. Development of an observation method for assessing pain behavior in chronic low back pain patients. *Behavior Therapy*, 1982.

[147] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*, pages 368–377. Springer, 2012.

[148] Bernardino Romera-Paredes, Min SH Aung, Massimiliano Pontil, Nadia Bianchi-Berthouze, Amanda C de C Williams, and Paul Watson. Transfer learning to account for idiosyncrasy in face and body expressions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[149] Johan WS Vlaeyen and Steven J Linton. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. *Pain*, 85(3):317–332, 2000.

[150] Dennis C Turk and Akiko Okifuji. Psychological factors in chronic pain: Evolution and revolution. *Journal of consulting and clinical psychology*, 70(3):678, 2002.

[151] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 57–64. IEEE, 2011.

[152] Hongying Meng and Nadia Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics*, 44(3):315–328, 2014.

[153] Temitayo A Olugbade, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Amanda C Williams. Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 243–249. IEEE, 2015.

[154] Wenfei Cao, Kaidong Wang, Guodong Han, Jing Yao, and Andrzej Cichocki. A robust pca approach with noise structure learning and spatial–spectral low-rank modeling for hyperspectral image restoration. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (99):1–17, 2018.

[155] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, 2016.

[156] Weijian Li, Di Huang, Huibin Li, and Yunhong Wang. Automatic 4d facial expression recognition using dynamic geometrical image network. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 24–30. IEEE, 2018.

[157] Yi Sun, Xiaochen Chen, Matthew Rosato, and Lijun Yin. Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(3):461–474, 2010.

[158] KarenPalmer. Riot. `http://karenpalmer.uk/portfolio/riot/`. Website.

[159] Pratik Gala, Raj Shah, Vineet Shah, Yash Shah, and Mrs Sarika Rane. Moodyplayer: A music player based on facial expression recognition. 2018.

[160] Qi-rong Mao, Xin-yu Pan, Yong-zhao Zhan, and Xiang-jun Shen. Using kinect for real-time emotion recognition via facial expressions. *Frontiers of Information Technology & Electronic Engineering*, 16(4):272–282, 2015.

[161] Inchul Song, Hyun-Jun Kim, and Paul Barom Jeon. Deep learning for real-time robust facial expression recognition on a smartphone. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*, pages 564–567. IEEE, 2014.

[162] Myunghoon Suk and Balakrishnan Prabhakaran. Real-time mobile facial expression recognition system-a case study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 132–137, 2014.