University of the
West of England
BRISTOL

2D and 3D Computer Vision Analysis of Gaze, Gender and Age

Wenhao Zhang

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy

Faculty of Environment and Technology

University of the West of England, Bristol

June 2016

TABLE OF CONTENTS

ABSTRACT

Human-Computer Interaction (HCI) has been an active research area for over four decades. Research studies and commercial designs in this area have been largely facilitated by the visual modality which brings diversified functionality and improved usability to HCI interfaces by employing various computer vision techniques. This thesis explores a number of facial cues, such as gender, age and gaze, by performing 2D and 3D based computer vision analysis. The ultimate aim is to create a natural HCI strategy that can fulfil user expectations, augment user satisfaction and enrich user experience by understanding user characteristics and behaviours. To this end, salient features have been extracted and analysed from 2D and 3D face representations; 3D reconstruction algorithms and their compatible real-world imaging systems have been investigated; case study HCI systems have been designed to demonstrate the reliability, robustness, and applicability of the proposed method.

More specifically, an unsupervised approach has been proposed to localise eye centres in images and videos accurately and efficiently. This is achieved by utilisation of two types of geometric features and eye models, complemented by an iris radius constraint and a selective oriented gradient filter specifically tailored to this modular scheme. This approach resolves challenges such as interfering facial edges, undesirable illumination conditions, head poses, and the presence of facial accessories and makeup. Tested on 3 publicly available databases (the BioID database, the GI4E database and the extended Yale Face Database b), and a self-collected database, this method outperforms all the methods in comparison and thus proves to be highly accurate and robust. Based on this approach, a gaze gesture recognition algorithm has been designed to increase the interactivity of HCI systems by encoding eye saccades into a communication channel similar to the role of hand gestures. As well as analysing eye/gaze data that represent user behaviours and reveal user intentions, this thesis also investigates the automatic recognition of user demographics such as gender and age. The Fisher Vector encoding algorithm is employed to construct visual vocabularies as salient features for gender and age classification. Algorithm evaluations on three publicly available databases (the FERET database, the LFW database and the FRCVv2 database) demonstrate the superior performance of the proposed method in both laboratory and unconstrained environments. In order to achieve enhanced robustness, a two-source photometric stereo method has been introduced to recover surface normals such that more invariant 3D facial features become available that can further boost classification accuracy and robustness. A 2D+3D imaging system has been designed for construction of a self-collected dataset including 2D and 3D facial data. Experiments show that utilisation of 3D facial features can increase gender classification rate by up to 6% (based on the self-collected dataset), and can increase age classification rate by up to 12% (based on the Photoface database). Finally, two case study HCI systems, a gaze gesture based map browser and a directed advertising billboard, have been designed by adopting all the proposed algorithms as well as the fully compatible imaging system. Benefits from the proposed algorithms naturally ensure that the case study systems can possess high robustness to head pose variation and illumination variation; and can achieve excellent real-time performance. Overall, the proposed HCI strategy enabled by reliably recognised facial cues can serve to spawn a wide array of innovative systems and to bring HCI to a more natural and intelligent state.

Keywords: *Human-Computer Interaction, eye centre localisation, gaze gesture, gender recognition, age recognition, directed advertising, 3D imaging*

ACKNOWLEDGEMENTS

This research would not have been possible without the kind help and steady support from many people, who help me to fight my fear and anxiety, to seek for joy and self-worth, and to become determined and confident when I am thousands of miles away from my home country.

My supervisory team of Prof. Melvyn Smith, Prof. Lyndon Smith and Dr. Abdul Farooq have been playing such an important role throughout my Ph.D. journey, starting from the very first moment I conceived the intention for taking out a baby step tentatively, and not ending until I can proudly address myself as Dr. Zhang, as I presume with my full trust (in my supervisors and in myself). Their suggestions have always been inspirational while their patience has always been admirable.

My thanks also go to other members of the Centre for Machine Vision team: Laurence Broadbent, who is a source of many inspirations of my research studies, a great help with the imaging system design and data capture experiments, and a generous and sincere friend, needless to say; Dr. Mark Hansen, a more than capable researcher, a patient listener to my tedious troubles and a database of solutions to my academic obstacles; Dr. Gary Atkinson and Dr. Khemraj Emrith, whose advices I have always treasured; Dr. Ian Hales, a proof-reader of my thesis chapters, provided valuable feedback; and many others I have enjoyed having discussions with and having parties with. I have also greatly appreciated the help and support from Aralia Systems Ltd., with regard to the ‘directed advertising’ case study in my thesis.

Apparently, everything I have achieved now is built up on the unconditional love and support from my parents and my extended family. They allow me to pursue my dreams without worries behind, and they have shouldered so many responsibilities emotionally and financially. I have never lacked incentives and motivations for accomplishing my goals, knowing that they selflessly wish me the best, but they never expect anything in return.

Portions of the research in this thesis used the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office, as well as the FRGCv2 database of facial images, collected under the Face Recognition Grand Challenge project.

LIST OF PUBLICATIONS

Journal publications arising from this work are listed as follows:

Zhang, W., Smith, M.L., Smith, L.N., Farooq, A. (2016) ‘Eye Centre Localisation: An Unsupervised Modular Approach’, *Sensor Review*, vol. 36, no. 3, pp. 277-286.

Zhang, W., Smith, M.L., Smith, L.N., Farooq, A. (2016) ‘Gender and Gaze Gesture Recognition for Human-Computer Interaction’, *Computer Vision and Image Understanding*, vol. 149, pp. 33-50.

Zhang, W., Smith, M.L., Smith, L.N., Farooq, A. (2016) ‘Gender recognition from facial images: two or three dimensions?’, *Journal of the Optical Society of America A*, vol. 33, no. 3, pp. 333-344.

Zhang, W., Smith, M.L., Smith, L.N., Farooq, A. (2016) ‘Eye centre localisation and gaze gesture recognition for human computer interaction’, *Journal of the Optical Society of America A*, vol. 33, no. 3, pp. 314-325.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Symbol	Description	First Appearance
D_x and D_y	The x and y component of a displacement vector pointing to an isophote centre	Section 4.1.1
$I(x, y)$	Luminance function	Section 4.1.1
I_x and I_y	First-order derivatives of the luminance function $I(x, y)$ in the x and y directions	Section 4.1.1
I_{xx}, I_{xy} and I_{yy}	Second-order partial derivatives of the luminance function $I(x, y)$ in the x and y directions	Section 4.1.1
$cd(x, y)$	Curvedness	Section 4.1.1
$\boldsymbol{E_a}(x, y)$	The attentive energy map calculated by the first module of the proposed eye centre localisation approach	Section 4.1.1
α	Maximum greyscale in an image	Section 4.1.1
$\boldsymbol{E_{aul}}(x, y)$	The left half of the attentive energy map	Section 4.1.1
$\boldsymbol{E_{aur}}(x, y)$	The right half of the attentive energy map	Section 4.1.1
c_{aul}	The optimal estimation of the left eye centre	Section 4.1.1
m	The maximum row number of a matrix	Section 4.1.1

n	The maximum column number of a matrix	Section 4.1.1
$\boldsymbol{c}^*$	The optimal eye centre, i.e. the eye centre with the highest votes	Section 4.1.2
$\boldsymbol{c}$	An eye centre candidate	Section 4.1.2
$arg\,max$	The argument of the maximum	Section 4.1.2
$\boldsymbol{p}(x, y)$	A pixel that is different from the current eye candidate	Section 4.1.2
$\boldsymbol{d}(x, y)$	The displacement vector connecting $\boldsymbol{c}$ and $\boldsymbol{p}(x, y)$	Section 4.1.2
$\boldsymbol{g}(x, y)$	A gradient vector in an image	Section 4.1.2
$I_{\boldsymbol{c}}$	The intensity value at an isophote centre in a greyscale image	Section 4.1.2
$rw(x, y)$	Significance measure set by the proposed iris radius constraint	Section 4.1.3
$\|\quad\|_2$ or ℓ_2-norm	The Euclidean norm of a vector	Section 4.1.3
$\widehat{D}$	The estimated radius of an iris	Section 4.1.3
σ	The order of a Butterworth low pass filter	Section 4.1.3
ω	The cut-off frequency of a Butterworth low pass filter	Section 4.1.3
$(matrix)^T$	Transpose of a $matrix$	Section 4.1.3

S_x and S_y	The length and width of a sliding window	Section 4.1.4
o_g	Gradient orientation in degrees	Section 4.1.4
tan^{-1}	The inverse tangent function	Section 4.1.4
π	A mathematical constant, the ratio of a circle's circumference to its diameter	Section 4.1.4
k	The number of bins for recording gradient orientations	Section 4.1.4
$\mathscr{s}$	A coefficient that is an integer between 0 and $k-1$	Section 4.1.4
$sw(x,y)$	Weight of a gradient vector adjusted by the SOG filter	Section 4.1.4
$\boldsymbol{E_b}(x,y)$	The attentive energy map calculated by the second module of the proposed eye centre localisation approach	Section 4.1.5
$\boldsymbol{E_f}(x,y)$	The integrated attentive energy map	Section 4.1.5
c_{maxl}	The pixel position that has the maximum energy response in the corresponding attentive energy map	Section 4.1.5
ϵ_f	The width of an eye region in a face image	Section 4.1.5
ϵ	A value relative to ϵ_f, i.e. $\epsilon = 0.3\epsilon_f$	Section 4.1.5
e	Normalised eye centre localisation error	Section 4.2

max/min	The maximum/minimum between two values	Section 4.2
d_{left} and d_{right}	The absolute eye centre localisation error for the left and the right eye, respectively	Section 4.2
P_d	The pupillary distance in pixels	Section 4.2
e_{max}, e_{min} and e_{avg}	The maximum, minimum and average normalised eye centre localisation error	Section 4.2
$\boldsymbol{e_l}(f)$ and $\boldsymbol{e_r}(f)$	Eye centre coordinates for the left and the right eye in image frame f	Section 4.3
$\bar{\boldsymbol{e}}(f)$	The average of the left and the right eye centre coordinates	Section 4.3
$\boldsymbol{e_{lx}}$, $\boldsymbol{e_{ly}}$, $\boldsymbol{e_{rx}}$ and $\boldsymbol{e_{ry}}$	Vectors recording the x and y coordinates of the left and the right eye centres in consecutive image frames	Section 4.3
$\boldsymbol{v_{lx}}$, $\boldsymbol{v_{lx}}$, $\boldsymbol{v_{lx}}$ and $\boldsymbol{v_{lx}}$	The velocity vectors for $\boldsymbol{e_{lx}}$, $\boldsymbol{e_{ly}}$, $\boldsymbol{e_{rx}}$ and $\boldsymbol{e_{ry}}$	Section 4.3
$\overline{\boldsymbol{v}}_{\boldsymbol{l}} = \{\overline{\boldsymbol{v}}_{\boldsymbol{x}}, \overline{\boldsymbol{v}}_{\boldsymbol{y}}\}$	The average velocity vector of the two eyes	Section 4.3
$\boldsymbol{r_g}(f)$	The ratio between movements (measured in pixels) of the left and right eye centres	Section 4.3
log	The logarithm operation	Section 4.3

$\boldsymbol{d_g}(f) = \{d_{gx}(f), d_{gy}(f)\}$	A vector where the signs of its x and y component correspond to the movement directions of an eye	Section 4.3
$\overline{\boldsymbol{v}}_{\boldsymbol{s}}$	The integrated gaze shift vector	Section 4.3
★ and •	The starting and ending position of a gaze gesture	Section 4.3
$p(\boldsymbol{x}\|\lambda)$	A parametric probability density function that can represent a Gaussian Mixture Model (GMM)	Section 5.1
$\boldsymbol{x}$	A D-dimensional data vector	Section 5.1
λ	The collective representation of GMM parameters	Section 5.1
β_i, $\boldsymbol{\mu_i}$ and $\boldsymbol{\delta_i}$	The mixture weights, the mean vector and the covariance matrix of component Gaussian densities	Section 5.1
$g(\boldsymbol{x}\|\boldsymbol{\mu_i}, \boldsymbol{\delta_i})$	Component Gaussian densities	Section 5.1
N	Number of Gaussians densities	Section 5.1
$\boldsymbol{X}$	Descriptors of low-level features extracted from an image	Section 5.1
T	The length of $\boldsymbol{X}$	Section 5.1
∇	The nabla symbol, a vector differential operator	Section 5.1

$\psi_{\lambda}^{\boldsymbol{X}}$	The gradient vector that characterise the samples $\boldsymbol{X}$	Section 5.1
$\mathcal{F}_{\lambda}$	The Fisher information matrix	Section 5.1
$\mathcal{L}_{\lambda}$ and $\mathcal{L}_{\lambda}{}'$	The lower triangular matrix and its conjugate transpose calculated from Cholesky decomposition of $\mathcal{F}_{\lambda}$	Section 5.1
$\Psi_{\lambda}^{\boldsymbol{X}}$	The Fisher Vector of $\boldsymbol{X}$	Section 5.1
$\gamma_t(i)$	The soft assignment of descriptor $\boldsymbol{x_t}$ to the Gaussian component i	Section 5.1
$\mathit{\Psi}_{\mu,i}^{\boldsymbol{X}}$ and $\mathit{\Psi}_{\delta,i}^{\boldsymbol{X}}$	The gradients of Gaussian component i with respect to the mean $\boldsymbol{\mu_i}$ and the covariance $\boldsymbol{\delta_i}$, respectively	Section 5.1
$\boldsymbol{\Phi}$	A Fisher Vector represented by the gradients of Gaussian components	Section 5.1
ps and ss	Size and step of a sliding window for low-level feature sampling	Section 5.2
(px_c, py_c)	Centre position of the cth sampling patch	Section 5.2
pn	The maximum number of sampled patches	Section 5.2
$\vec{\boldsymbol{w}}$ and b	The weight vector and the intercept term that define a hyper-plane for a support vector machine	Section 5.2
$\boldsymbol{W}$	A matrix with its diagonal values as in $\vec{\boldsymbol{w}}$	Section 5.2

$R(p,q)$	A reflectance map	Section 6.1
ϱ	Surface albedo	Section 6.1
$\vec{\boldsymbol{N}}$	A surface normal vector	Section 6.1
$\vec{\boldsymbol{L}}$	A lighting vector defining illumination direction	Section 6.1
z	Surface depth	Section 6.1
$\overrightarrow{\boldsymbol{L_n}}$	A normalised lighting vector	Section 6.1
$\boldsymbol{R_c}$	Connected regions of an arbitrary surface	Section 6.2
$\boldsymbol{g_r}(i,j)$	The recovered surface gradient map	Section 6.2
$\boldsymbol{g_t}(i,j)$	The ground truth for the surface gradient map	Section 6.2
$e_{\vec{N}}$	The average error of surface normals between the reconstruction and the ground truth	Section 6.2

ACRONYMS/ABBREVIATIONS

2D/3D	Two-dimensional/three-dimensional
CCTV	Closed-circuit television
CDF	Cumulative distributed function
CNN	Convolutional neural network
CPU	Central processing unit
DOOH	Digital out-of-home
FERET	Face Recognition Technology
FLD	Fisher linear discriminant
FV	Fisher vector
GMM	Gaussian mixture model
GUI	Graphical user interface
HCI	Human-computer interaction
HD	High-definition
HOG	Histogram of oriented gradients
LBP	Local binary pattern
LED	Light-emitting diode

LFW	Labelled face in the wild
ML	Maximum likelihood
NIR	Near infrared
PCA	Principal component analysis
PS	Photometric stereo
RBF	Radial basis function
RGB	Red, green and blue
RMS	Root-mean-square
SFS	Shape from shading
SIFT	Scale invariant feature transform
SOG	Selective oriented gradient
SVM	Support vector machine
TOF	Time-of-flight

Chapter 1 Research Background and Objectives

1.1 Human-Computer Interaction Principles

This thesis explores vision-based methods that facilitate automatic recognition of gaze, gender and age from facial images; and the overall objective is to enable enhanced Human-Computer Interaction (HCI). The wide applicability and high practicability of this research has been demonstrated, and the common challenges in the studied areas illustrated by the variability of HCI scenarios. With the emergence of personal computing in the late 1970s, the concept of HCI was pushed rapidly and steadily into various aspects of life. A fundamental reason is that the individual role of science (which aims at fundamental research) or engineering (which aims at applied research) was not sufficient for addressing the urgent need for increasing the usability of computer software and operating systems. The potential of the increased accessibility to personal computers and the demand for higher usability of computer platforms called for the practical need for HCI as a synthesis of science and engineering. Differently put, both opportunities and needs contributed to the emergence of HCI as an established discipline. Therefore, HCI arose and has since been facilitated by a number of science and engineering areas, notably cognitive psychology, artificial intelligence, computer science, etc., aiming to increase the usability of computers and to decrease the barrier between the needs of human beings and the perceptive/assistant objectives of computers. As seemingly self-explanatory as it is, the definition of HCI has extended and varied as its exploitation strengthens and diversifies, with the advancement of many supporting technologies. Essentially, HCI studies surround the development and use of computing technologies to interact with people (users) for fulfilling specific tasks. However, HCI, as a cross-disciplinary research area, embraces a scope greater than the implementation of various interfaces through which users can interact with computer devices. It also encompasses HCI theory, design criteria and evaluation means, which are the foundation and assurance of excellent HCI systems and are worthy of an equal amount of study.

The functionality of a HCI system has always been a critical design criterion but it is usually addressed solely from a technical point of view. Its assessment generally surrounds the system itself, dedicated but not restricted to system features, performance, reliability and durability (McNamara and Kirakowski, 2006). However, functionality does not in itself guarantee a credible standard for designing and evaluating HCI systems since a HCI event always involves two parties – the computer and the user – and therefore its evaluation should be subjective and context dependant. When both parties are concerned, usability (Bevan, 2001) is commonly considered to be a complementary criterion as it evaluates the adequacy and efficiency of goal-directed HCI events. High usability entails desirable learnability, flexibility and robustness. For example, a new user should be able to achieve effective interaction with ease, meaning that one would be able to exchange information with a HCI system via multiple channels, as well as receiving sufficient feedback and support from the system while performing goal-directed actions. The ISO 9241 standard (FDIs, 2009) addressed usability measures from the evaluation of effectiveness, efficiency and satisfaction. To put forward a more practical evaluation standard, Shneiderman (1992) elaborated this as: time to learn, speed of performance, rate of errors by users, retention over time, and subjective satisfaction. It can be seen that although these standards deliver similar messages, they have been established with different focuses, mainly due to the dynamic and versatile nature of HCI.

While functionality and usability of a system interface or an algorithm remain as fundamental evaluation criteria, users' external point of view in terms of expectations, satisfaction and experience has become increasingly significant as many innovative HCI technologies arise that aim to create user-centred interaction environments. This is not strictly independent of the former criteria since richer functionality and usability will most likely evoke user satisfaction. However, the premise for a consistent functionality/usability level and a good user experience is that the designer fully comprehends the needs and expectations of target users. In other words, although it has been widely acknowledged that endeavours to boost user experience serve to facilitate HCI design, it is unattainable to maximise user experience without a considerable amount of knowledge about the user. This was discussed by Kim (2015) as the term 'foremost creed', which had been coined by Hansen (1971). Dix et al. (2004) also stressed that a major source of failure repeatedly seen in engineering was 'the narrow optimisation of a design that does not take sufficient amount of contexual factors'. As a matter of fact, contexual factors are meaningless without consideration of human characteristics and behaviours.

Observing a user and analysing user characteristics and behaviours can serve to reveal one's preferences, capabilities, intentions, etc. For instance, the senior generation and the young generation have different mental and physical capabilities. Therefore, age information as prior knowledge can ensure that a suitable interaction modality is delivered to cater to a target's capabilities. Similarly, when the gender of a user is known, a visual interface or system feedback can be themed to suit the user's preferences that root from gender. Aside from demographic information, behavioural data (such as a user's position, gaze direction and facial expression) normally reveal a user's intention or emotion, and therefore forecast one's tendency or uncover one's impressions/reactions. On the one hand, this allows efficient interactions by predicting a user's subsequent actions; on the other hand, this gathers user feedback in a non-intrusive manner, which is of great value to HCI system evaluation. Understanding user demographics and behaviours provides a way to couple human intelligence and machine power, i.e. the integration of superior cognitive capability and computational capacity, which shows high promise in magnifying the functional and assistive role of HCI systems.

Another HCI principle is "naturalness" – a trait which delivers the concept that HCI sessions should be enabled in a similar way to human daily activities. However, designing a real-world HCI system that enables near real-life interaction modes is more challenging than it seems. The reason behind this is threefold: firstly, human perceptive and cognitive capabilities are developed over years; their robustness and adaptability cannot be easily emulated by computer algorithms. Secondly, human beings gather comprehensive information, in real time, through multiple sensors (such as visual sensors, acoustic sensors, haptic sensors, etc.), whilst the implementation of multimodal HCI systems is a complicated process. Thirdly, human sensors excel in precision and efficiency when collecting data from the surroundings. This can barely be matched by most artificial sensors at a reasonable cost.

With these HCI principles introduced, Figure 1.1 provides a summary of HCI development history, interfaces and design/evaluation criteria.

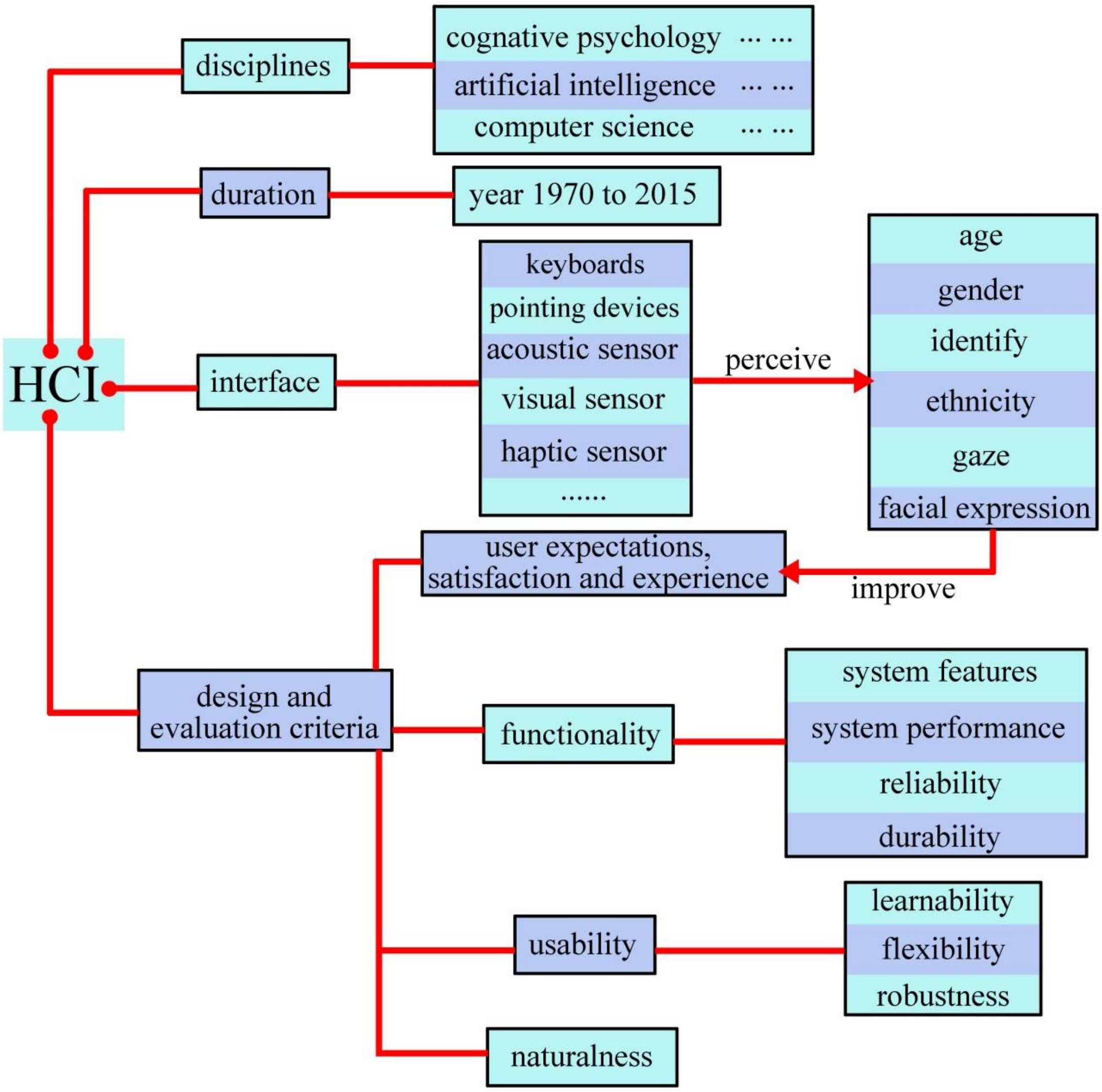


Figure 1.1 A summary of HCI development history, interfaces and design/evaluation criteria

Admittedly, the HCI design principles introduced above are demanding and cannot be easily and fully complied with. They have ruled out many traditional techniques that are entirely reliant on user input, for example, the classic mouse/keyboard interface, meaning that they dictate that HCI system designs should be equipped with a certain level of intelligence enough to perceive and respond to information At the same time, to view HCI from multiple perspectives can facilitate more intelligent HCI systems that permeate modern life.

This research is intended to reflect these HCI design principles in a combination of theoretical studies and practical designs by exploiting novel algorithms, intelligent HCI systems and advanced HCI strategies. Therefore, the outcomes of this research will facilitate designs and implementations of HCI systems that can highly fulfil user expectations and satisfaction, with rich functionality and usability. This thesis is dedicated to arguably the most prevalent HCI modality – vision based HCI – which is enabled by a number of 2D and 3D image acquisition approaches, and image processing and recognition algorithms. Therefore, the subsequent section gives a general introduction to the visual modality while Chapter 2 reviews in more detail representative vision based HCI applications and heated research topics in the visual modality.

1.2 The Visual Modality in Human-Computer Interaction

It is uncontentious that it is extremely challenging to design and implement HCI systems that are in full compliance with the aforementioned principles. Nevertheless, innovative and exploratory HCI systems are being forged that attempt to fulfil the increasingly sophisticated needs of their users. The sophistication intensifies as personal computers present more digital information to people. The way users perceive, process and respond to information has largely altered. For example, a user nowadays would expect a digital system to deliver information with clarity, efficiency and relevancy. One would also demand a user-friendly and multifunctional interface with attractive aesthetic design. In addition, when overwhelmed by an excess amount of information, one would be likely to feel unconcerned and to be unresponsive unless the information is appropriately filtered and highly personalised. To address this, HCI has not only become central to information science theoretically and professionally, but also has stepped into people’s lives, offering multiple communication channels, i.e. modalities. These modalities gain their input via various types of sensors, mimicking human sensors including visual sensors, audio sensors, haptic sensors, etc. These modalities, independently or when combined, give rise to a wide array of HCI systems, referred to as unimodal HCI systems when only a single modality is involved or multimodal systems when a number of modalities are combined.

Among the various types of sensors, the visual sensor and the resulting visual based interaction modality are the most widespread, taking visual signals as input, which include but are not limited to images/videos of faces, bodies and hands. Among them, the main detection and recognition tasks include the following elements:

- user presence and location
- user identity, gender, age and ethnicity
- user facial expression, focus of attention, gesture and body posture

The corresponding research spans the areas of face recognition, gender recognition, age recognition, gaze tracking, gait analysis, gesture recognition and facial expression recognition. The primary goal of all these branches is to sense and perceive human and human activities such that interfaces can be designed to provide services accordingly. Nevertheless, each individual branch has placed its focus differently, therefore a variety of compelling paradigms emerge with specific functionalities. While face recognition reveals an individual’s identity; gender/age recognition aims to gather human demographics for the system to understand human characteristics. Facial expression recognition further estimates human affection states and gathers emotional cues while gaze tracking serves to extract users’ attentive information and to predict their intentions.

Although face recognition has the potential to improve services for consumers, to enable online and offline identity authentication and to facilitate other innovative technologies, it has been controversial in many application domains. Concerns arise in the wake of employment of face recognition software and face databases associated with personal identities in various scenarios ranging from security implementations to entertainment applications. Sceptical users would be reluctant to use HCI systems enabled by face recognition technologies, due to fears that civil liberties and privacies are at a high risk of being compromised (BBC, 2015). Other than the implications for user privacy and personal safety from the side of system users, face recognition technologies need to be properly and strictly regulated by the government and thus their constrained use can be foreseen in the near future. In contrast, recognitions of other human characteristics such as gender, age and ethnicity are targeted to user groups rather than individuals. Therefore, their development in either the research or the commercial domain can expand without distinct concerns over user privacy and government restrictions. In comparison, exploring other identity-neutral facial cues will be less controversial and is likely to witness higher applicability. There is also a high possibility for facial expression analysis to struggle in contributing to HCI. This is mainly due to HCI environments being complex and dynamic. During a HCI session, facial expressions of a user are not necessarily triggered by the interactive system, but can be caused by other stimuli from outside the HCI environment. For example, one may smile to a friend while performing a boring HCI activity. In these conditions, the interpretation of facial expressions is out of context and is therefore meaningless.

By estimating and collecting such human characteristics and behavioural statistics, the visual modality precisely addresses the main HCI principle – to improve user expectations, satisfaction and experience by creating a user-centred HCI environment. This is not a coincidence, but a conspicuous fact, since roughly eighty percent of perception is visual (Moore, 1994) in a natural face-to-face interaction scenario. Therefore, learning about a user can be most conveniently and comprehensively achieved by the visual modality. This further explains why the exploitation of the visual modality has become the most widespread, and has continued to leave many open problems and unexplored opportunities. A few examples can be introduced to reflect the versatility of the visual modality from different perspectives. One example concerns the design of a virtual keyboard which can optically interface with a user via an illuminator and a light sensor (Bair and Kern, 2011). The projected light illuminates the input zone of the virtual keyboard and reflects off a user’s fingers. The reflected light patterns then reveal the finger positions and provide feedback to the compact computer. Although accounting for incremental improvement to classic HCI systems, the design of this system does not escape the boundaries set by the classic mouse/keyboard paradigm, regarding the context, scale and functionality of the application. In contrast to this application with low cognitive capability and intelligence, another example features a number of hand gesture recognition algorithms that can detect, track and recognise static or dynamic hand gestures (Rautaray and Agrawal, 2015). Conscious and unconscious hand gestures convey human emotions and intentions. Automatic recognition of hand gestures can create immersive gaming environments, facilitate augmented reality applications, and benefit pervasive and mobile computing applications. Another recent example has further demonstrated machine intelligence – ‘Pepper’ the robot (Aldebaran softbank group, n.d.). This is the first robot designed to live and converse with humans. Although Pepper is not a functional robot for domestic use, it has the ability to analyse facial expressions and body language through two HD cameras and a 3D depth sensor. The visual modality in this example is combined with other modalities, such as the speech and auditory modality and the haptic modality, to predict human emotions and to react correspondingly.

As well as being one of the main enablers for enhanced intelligence, such as cognitive capability, of HCI systems, the visual modality also announces the potential of HCI systems pervading the lives of human beings, by giving rise to diversified applications. Marketing, for example, is an area that used to receive less benefit from the advancement of HCI. Experiences of passively tuning into advertisements during web browsing based on user location and browsing history (Farahat and Bailey, 2012) are probably more than familiar to the modern generations. However, not until the recent decade has the idea of *digital out-of-home (DOOH) directed/targeted advertising* been introduced, which still remains more of a conceptual model than a full-blown practice. DOOH directed advertising aims to deliver personalised advertisements to consumers when they are outside their homes. This extends the internet-based personalisation such that a greater applicability can be foreseen. Its effectiveness is evidenced by many studies indicating that targeted advertising increases the equilibrium profits of firms and lowers advertising expenditures (Iyer, Soberman and Villas-Boas, 2005; Adams, 2004). The visual modality paves the way for the emergence of this novel advertising mode by allowing intelligent systems to analyse consumer demographics and behaviours, and/or even to stimulate consumer engagements, in public venues such as shopping malls and airports.

Despite the huge variety and potential that the visual modality gains through detection, recognition and tracking tasks, many problems appear to be fundamentally ill-posed due to extrinsic variations such as image noise, variation in lighting condition, camera viewpoint, to name just a few. Further complications arise from the complexity of human appearance, body structure and behaviour, which introduce intrinsic intra-class and inter-class variations (e.g. different facial features within the male group or between male and female groups). These are the main and general factors that impede the seamless duplication of human perception and the transformation from laboratory setups into practical applications.

These complications are challenging and remain unresolved to date. Investigations on 2D images (in colour or greyscale) have struggled to effectively deal with extrinsic and intrinsic variations in many object recognition or classification tasks. Novel imaging techniques are therefore in high demand that can better characterise object textures, structures and shapes. In view of these challenges and complications, some researchers explore 3D imaging technologies, which have recently served to complement or replace 2D based means for image acquisition and analysis. While many 3D imaging systems are expensive and complex, the Microsoft Kinect sensors have created many opportunities for professional study as well as for amateur usage. This is due to its ability to sense depth information, with its wide availability at low cost. Its contributions have been seen in areas such as hand-gesture recognition, body biometric estimation and 3D surface recognition. An introduction to the Kinect sensor and a review of its applications and evaluations can be found in Section 2.4, where other 3D imaging techniques are also reviewed and discussed.

1.3 Motivation

This section illustrates the motivations that shape this research and drive it from one stage to another. The motivations of this research are derive from a review of HCI designs and principles in the preceding section, as well as the impeditive factors identified in this field. They are summarised as follows and can be further reflected by the literature review in the subsequent chapter.

(1) Current HCI systems explore either demographic data or behaviour data but have rarely accounted for both nor made predictions according to the fused knowledge.

Gender, age and gaze have been exploited separately in the visual modality for HCI by various research works. Variations in appearance of a face due to gender and age have a significant impact on face recognition/verification. Predicted age and gender information, as prior knowledge, also provide critical cues for effective HCI as they reveal human characteristics and thereby assist HCI systems with understanding their target users. In the other respect, gaze study and eye analysis implicitly uncover user attention and intention, which can serve as signals that trigger particular HCI activities and alter system behaviours, or as statistics that help evaluate HCI experience. Although human demographics are informative, they are insufficient to form a credible basis for determining HCI system responses. Human behaviours should also be utilised to adjust system responses in a dynamic manner.

More concisely, demographic recognition helps determine and define the *initial interaction state*, while behavioural recognition serves to tune the *transient interaction state* in a dynamic manner. This can be explained by an example of a face-to-face interaction process, illustrated by the flow chart in Figure 1.2.

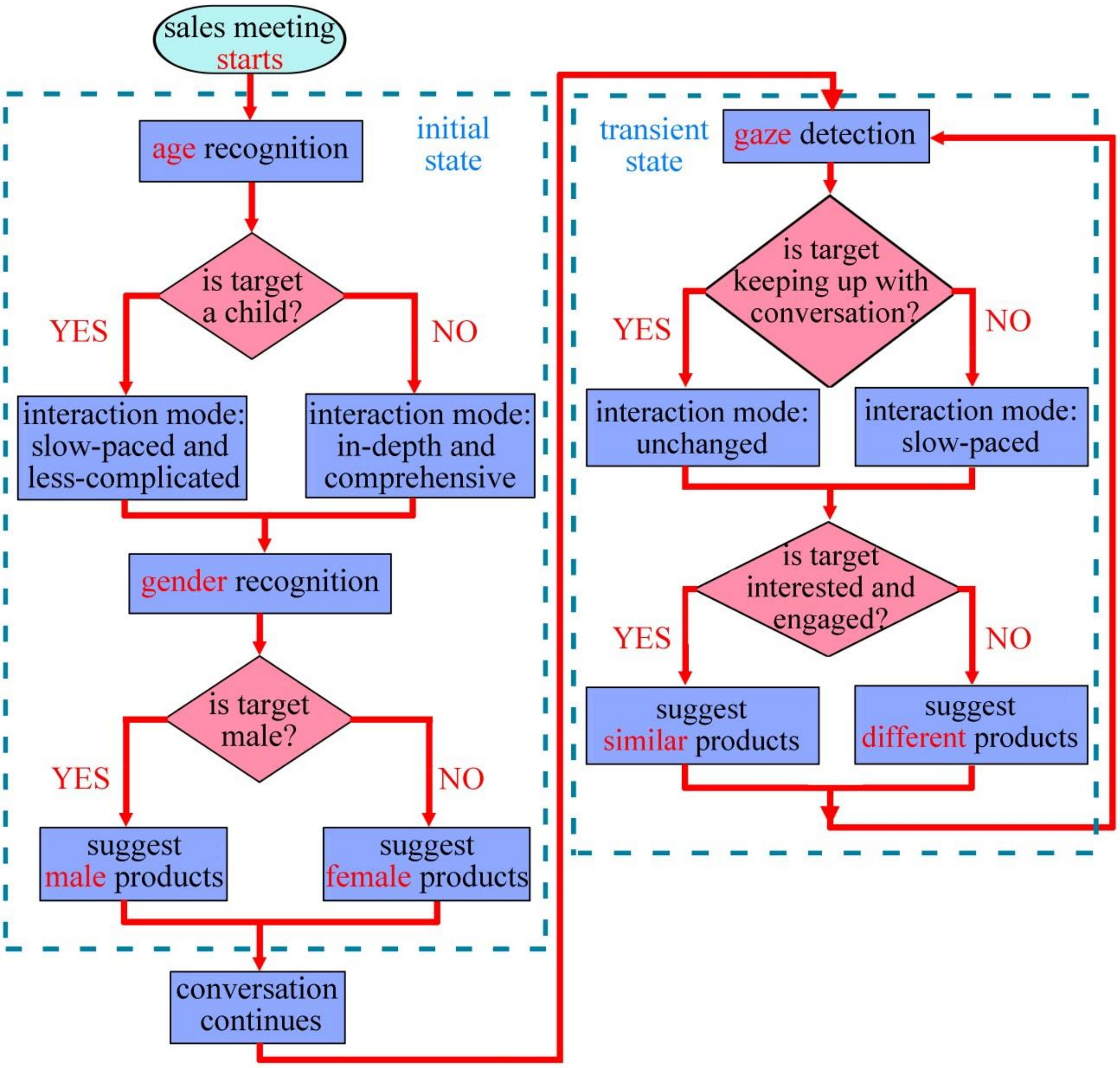

sales meeting starts
initial state
age recognition
is target a child?
YES
NO
interaction mode: slow-paced and less-complicated
interaction mode: in-depth and comprehensive
gender recognition
is target male?
YES
NO
suggest male products
suggest female products
conversation continues
transient state
gaze detection
is target keeping up with conversation?
YES
NO
interaction mode: unchanged
interaction mode: slow-paced
is target interested and engaged?
YES
NO
suggest similar products
suggest different products

Figure 1.2 An example of face-to-face interaction enabled by demographic recognition and behavioural recognition

As Figure 1.2 illustrates, when a salesman is trying to win over a customer during a face-to-face sales meeting, he is likely to adjust his marketing strategy according to the gender and age of the customer, e.g. selling male clothing to a man, lipsticks to a woman, or toys to a child. In addition, when the customer is a child, a slow-paced and less-complicated communication mode is needed; when the customer is an adult, an in-depth and comprehensive communication mode can be employed. This manner is defined by the initial state of the face-to-face interaction and is facilitated by the demographic data available. As the interaction progresses, by gauging the reactions of the customer (gaze directions that can imply the level of attentiveness, for example), the salesman could estimate the level of interest the customer is showing and alter his sales strategy accordingly. For example, a male customer might show greater interest in female clothing over male clothing, and a senior adult might be fascinated by fluffy toys. Concerning the face-to-face interaction manner, an adult with hearing or visual impairments will also appreciate a slow-paced communication mode, different from other adult customers of a similar age.

In these cases, behaviour data allow for timely and effective adjustment to sales strategies, and compensate for the inadequacy of demographic data during dynamic HCI sessions. It is only intuitive to relate this face-to-face sales/marketing strategy to a typical HCI scenario – DOOH directed advertising.

Not being restricted to this specific example, the combination of demographic recognition and behaviour recognition can forge a more personalised and natural HCI strategy, as well as serving users based on *'what they need'* instead of *'who they are'*.

(2) Most HCI systems work under constrained environments and lack precision and robustness in real-world scenarios.

Many classic and contemporary HCI applications and algorithms claim to have achieved good performance on well-controlled public datasets or in laboratory settings, but they degrade severely as they undergo real-world experiments (Sutcliffe, 2006; Oulasvirta, 2009). The major challenges in real-world scenarios are posed by variations in illumination, image noise, viewpoint, occlusion, image misalignment and low-quality images. Although a number of promising methods have been proposed to compensate for undesirable illumination (Singh, Zaveri and Raghuwanshi, 2010), to tolerate head pose variations (Zhang and Gao, 2009), and to enable outdoor applications, the challenges remain *unsolved* in realistic environments. Poor performance regarding reliability and robustness is a main factor that limits visual modality based HCI studies from realising effective and widespread implementations. Therefore, robust systems that can work under *complex and dynamic scenes* are in high demand. As a result, tackling the issues posed by environmental complications has become a necessity for both theoretical studies and industrial works.

(3) Lack of personalised interaction, intelligence and adaptability.

As stated previously, to fulfil user expectations and satisfaction, and to enrich user experience is nowadays the foremost principle that guides and drives a HCI system design. However, the gap is yet to be bridged due to inadequate knowledge about target users. The ability to perceive user characteristics will allow a HCI system to gain escalated intelligence, to engage its users actively with more personalisation, and to adapt system responses to user behaviours at the same time.

(4) Insufficient exploration in 3D imaging for visual detection and recognition.

As feature extraction and selection aim to produce more salient and discriminative information to reflect the uniqueness of a class, 2D features generally and inevitably encounter extrinsic challenges posed by poor illumination conditions, changes in viewpoint and other environmental factors. This explains why much research has reported high performance under laboratory environments but have failed to resolve tasks under real-world conditions. 3D imaging techniques can provide features that are independent of ambient light and allow 3D curvatures, textures and shape representations to contribute to higher accuracy and robustness in object detection and recognition tasks. 3D imaging is receiving an increasing amount of study but is still overshadowed by the domination of 2D based techniques.

Furthermore, existing methods have rarely explored all three major components of 3D based detection and recognition methods: *3D reconstruction algorithms*, *3D* *imaging systems* and *3D recognition algorithms*. As a result, high inconsistency can be seen in most studies in this area, which largely impairs the practicability of these methods. Specifically, many 3D recognition algorithms are evaluated on publicly available 3D databases of high resolution, captured in laboratory environments (Hu, Yan and Shi, 2010; Fagertun et al., 2013). These 3D data are normally obtained by expensive devices and/or complex imaging systems. Deployment of these imaging systems for HCI purposes is often impractical. In addition, the majority of these systems, despite their high cost, cannot operate in real time. Therefore, it can be claimed that a suite of explorations is in high demand that can effectively link together studies of 3D reconstruction algorithms, 3D imaging systems and 3D recognition algorithms. A number of mainstream 3D imaging techniques and their applications are reviewed in Section 2.4.

1.4 Aim and Objectives

The *main aim* of this research is to investigate the potential for a real-world system capable of capturing image data of a person or persons to accurately interpret human characteristics, recognise human responses and predict intentions for HCI. This aim is to be attained by, for the first time, associating demographic and behavioural data captured in a relatively unconstrained environment and making informed interpretations, predictions and interactions. To achieve this aim, the following objectives need to be met:

1) Develop a system (hardware and software) able to capture 2D and 3D user data at high spatial and temporal resolution.

Colour images and 3D images will be collected or reconstructed such that 2D features, 3D features and fused features become available for classification tasks. In-depth analysis from all aspects will be conducted, aiming to extract both coarse and detailed information while maintaining decent real-time performance.

2) Design and implement algorithms for robust and salient feature extraction from 2D and 3D faces.

Salient facial features that are robust against illumination variation, head pose change and other environmental factors could reflect the intrinsic attributes of objects of interest and would facilitate classification tasks. Features are selectively incorporated by analysing 2D and 3D data, which are adopted for purposes including gender classification, age classification, gaze analysis, etc. 2D features have the advantages that they can be conveniently obtained without involving a complex imaging process. They can provide satisfactory face representations, but only when large head poses and illumination variations are not present. 3D features, on the other hand, are robust against these variations, but at a sacrifice of algorithm efficiency and system simplicity. This indicates that 2D features and 3D features are complementary and the incorporation of both types of features has the potential to lead to intelligent vision systems with high accuracy, while maintaining desirable usability.

3) Perform robust analysis and classification of extracted features for classification and recognition of demographic and behavioural data.

Accurate classification of gender and age, together with purposive attention inferred from gaze, could help evaluate the reaction and intention from different groups of users, which could lead to group behaviour prediction. Gaze signals can also be employed as eye gesture input for active user engagement.

4) Develop adaptive models and integrated algorithms, to accurately predict and respond to user characteristics and behaviours by providing personalised feedback, e.g. advertising messages for DOOH directed advertising.

The adaptive model constructed will make predictions and respond to a user according to data available; for example, more adequate and comprehensive data tend to be captured from a user within a short distance to the camera with a frontal face under ideal illumination conditions. In this case, the adaptive model can opt for a most interactive scheme by employing both 2D and 3D data, but otherwise will stay partially dormant. More specifically, when imaging distance increases to such an extent that 3D reconstructions would suffer from noise and 3D based classification would be impaired by low resolution data, only 2D data/features should be employed. A full interaction scheme will be established to engage the user in the interaction.

It should be noted that, although this research is motivated by the needs and unresolved challenges in the area of HCI, it is essentially dedicated to explorations of novel 2D and 3D vision based methods, such as 3D imaging/reconstruction methods and extraction of discriminative facial features, rather than focusing on designs of HCI applications. Therefore, validation of the proposed research for effective HCI is beyond the scope of this research.

1.5 Original Contributions

In general, the following original contributions have been made by the author toward the realisation of this aim and objectives:

1) An unsupervised modular eye centre localisation approach, which seamlessly combines two types of geometric features for global and regional periocular/facial analysis. The method is free from training and robust against illumination and head pose variations to a high degree. It has also been extended into a gaze gesture recognition algorithm for enhancing the functionality and interactivity of HCI systems (see Chapter 4).

2) A generic classification method, i.e. the Fisher Vector encoding method, which is intended for gender and age recognition, and is capable of encoding and classifying both 2D (see Chapter 5) and 3D (see Chapter 7) facial features. To the best of the author’s knowledge, this is the first time that this method has been applied to gender recognition with a comprehensive evaluation of algorithmic parameters.

3) Two complementary two-source PS methods (see Chapter 6), which can reconstruct 3D faces in real time and under unconstrained environments. To the best of the author’s knowledge, this is the first time that a two-source PS method has been successfully implemented on real data rather than simulated data.

4) An advanced HCI strategy (see Chapter 8) that combines demographic recognition (e.g.

age recognition and gender recognition) and behaviour recognition (e.g. gaze recognition).

5) Development of a 2D+3D imaging system (see Chapter 3). As well as providing a consistent hardware support for the two-source PS method, it serves as a data capture system for algorithm evaluations and validations.

6) Design of two proof-of-concept HCI systems (see Chapter 8) – a gaze gesture based map browser and an intelligent directed advertising billboard. They act as additional validators for the proposed algorithms with regard to accuracy, robustness and efficiency.

1.6 Thesis Structure

The remainder of this thesis is structured as follows:

Chapter 2 first reviews three major modalities in the field of HCI and a number of representative applications spawned by each modality. The trends and barriers for each modality are then conclusively discussed to shed light upon the near future of HCI in terms of theoretical development and practical advancement. The emphasis is placed on the visual modality since it is arguably the most prevalent modality and is the basis of the proposed methodologies in this thesis.

The three major aspects researched by this work – eye/gaze analysis, gender and age recognition and 3D imaging – are subsequently and individually reviewed. On the one hand, this further clarifies the motivation, the significance and the novelty of the works presented by this thesis. On the other hand, this sets a reference such that the proposed methodologies can be evaluated comparatively and objectively.

Chapter 3 introduces the development of a 2D+3D imaging system. This system plays a dual role in the advancement of HCI: to capture facial data for algorithm evaluation in real-world environments; and to serve as an experimental HCI system that can demonstrate the potential of the proposed method from an applied point of view. This chapter also presents a group of data capture experiments specifically designed for the evaluation of the proposed HCI scheme that combines *demographic recognition* with *behavioural recognition*, and incorporates *2D recognition* with *3D recognition*.

Chapter 4 starts with an introduction to a novel unsupervised and modular eye centre localisation approach as a means to accurately and efficiently gather *behavioural* data for HCI. This approach further leads to a gaze gesture recognition method intended to enhance the *interactivity* and *functionality* of HCI systems.

Chapter 5 proposes to employ Fisher Vectors that are encoded from 2D features for gender recognition. The FV encoding method aims to achieve highly accurate, robust and reliable gender predictions as *demographic* data for HCI. This method is further extended to incorporate 3D features and is also applied to age recognition in Chapter 7.

Chapter 6 explores 3D face reconstructions by photometric stereo variations. Two complementary reconstruction methods are proposed and evaluated. While the 2D+3D imaging system (introduced in Chapter 3) can function as the hardware basis of an experimental HCI system, the proposed two-source PS method forms the algorithmic core that is responsible for the incorporation of 3D features in order to achieve higher reliability and robustness.

Chapter 7 combines 2D and 3D features and yields an empowered scheme for gender/age recognition. It discusses the employment of over ten types of 2D and 3D features and proves the superiority of the proposed method with both laboratory datasets and real-world datasets, publicly available datasets and self-collected datasets.

Chapter 8 introduces two case studies facilitated by the proposed method and demonstrates that, when individually employed or when combined, the proposed behavioural recognition method and the demographic recognition method can give rise to diverse intelligent HCI systems that are able to cater to the needs of different groups of users.

Chapter 9 discusses the proposed algorithms, their evaluations, and implementation details, and draws conclusions.

The co-relation and interaction of the proposed methods are further illustrated in Figure 1.3 in order to outline the highlights and to demonstrate the consistency of this research study.

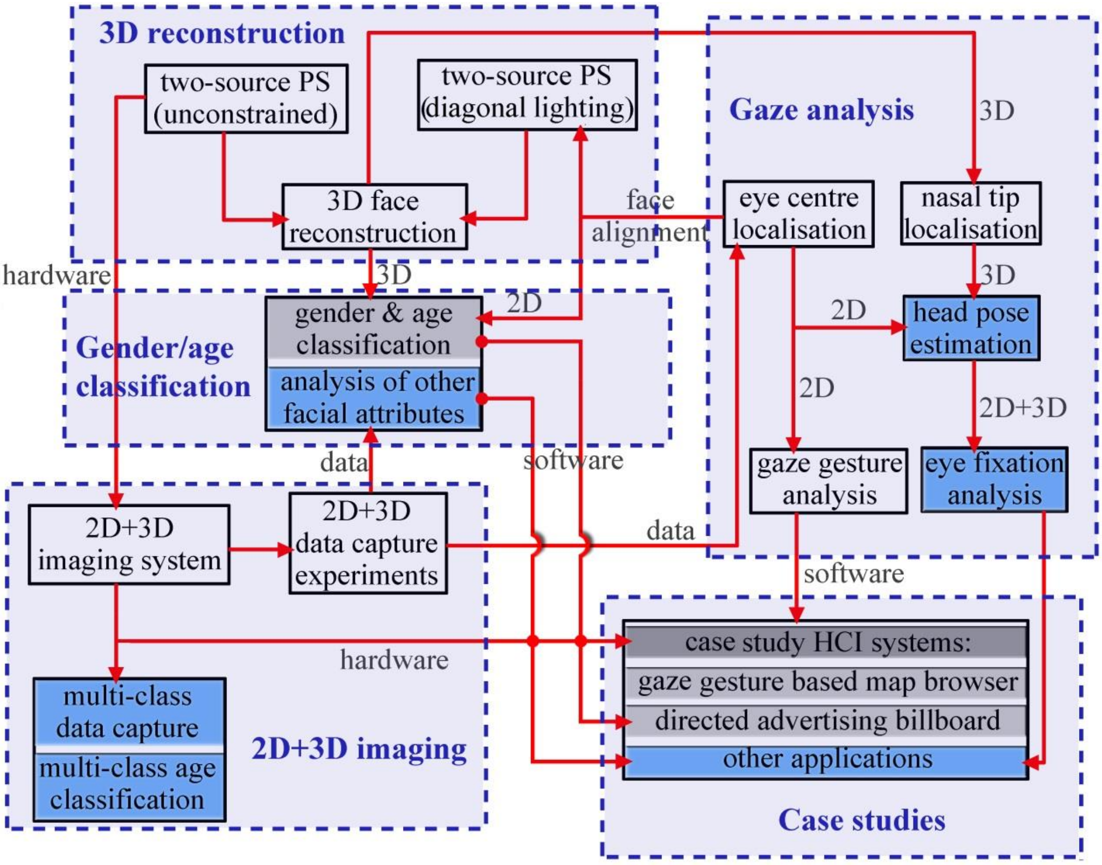


Figure 1.3 A summary of thesis structure, illustrating co-relation of the proposed algorithms, developed systems and conducted experiments. Future works are marked with dark blue backgrounds.

It can be seen that the works carried out by this research are highly co-related and interact with one another such that a suite of proposed algorithms, hardware implementations and case studies is delivered to diminish the isolation between 2D and 3D based visual studies, as well as bridging the gap between theoretical research and practical exploration.

Chapter 2 Literature Review

This chapter first reviews the state of the art of HCI regarding applications enabled by the speech and auditory interfaces, the haptic interfaces and the visual interfaces. It then identifies the trends and barriers of the three major modalities. Going beyond the classic but simplistic mouse/keyboard paradigm and other pointing devices, the focus of this chapter is placed on the visual modality, a highly prevalent and powerful modality embraced by HCI development. It critically introduces a number of representative studies in the area of eye/gaze analysis, gender and age recognition, and 3D imaging, which are the main areas contributed to by this research. With the obstacles identified in the area of HCI, the methodologies presented in subsequent chapters are then targeted at the challenges, in order that a more natural and powerful HCI strategy can be designed to convert the theoretical foundation into practical significance.

2.1 Human-Computer Interaction – Applications, Trends and Barriers

A few decades ago, HCI was enabled by only a few graphical user interfaces (GUIs), fulfilled by the classic mouse/keyboard interaction paradigm. To date, the thriving development of GUIs has appeared to be more appealing, convenient and efficient, but has somewhat failed to escape the boundary of the classic paradigm that seems forever unchanged. Although HCI systems nowadays take more forms than desktop computers and HCI scenarios start to display a trend of being flexible and diversified, in most designs explicit commands from a user remain as the major control and initiation signals for HCI events. More specifically, keyboards and keypads are the primary means of data entry. They are accompanied by the use of pointing devices (e.g. touchscreen, stylus, mouse, etc.) that assist in performing selection, positioning, and other simple interaction tasks. This has caused the most prevalent HCI systems to require supervision from their users to a large extent. Consequently, they are far from being intelligent, personalised and engaging. To overcome these limitations, users are progressively and experimentally enabled to participate in the interaction process through multiple sensor modalities. As illustrated by Hassanpour, Wong and Shahbahrami (2008), traditional interactive interfaces were evolving from inefficient mouse and keyboard devices to more intuitive and motivating user interfaces. These sensor modalities give rise to the *speech and auditory interfaces*, whose ultimate aim is to bring the fictional human-computer chatting scenarios (in a non-robotic manner) into reality. This modality has been considered complementary to the visual modality, or in some cases, even more informative than the visual modality, as it reduces visual overload, reinforces visual messages and conveys additional emotional cues (Peres et al., 2008). Although these fictional scenarios are far from being a desirable reality now, success has been witnessed when low cognitive capability is demanded from a HCI interface. Speech recognition technologies have experienced a worldwide success, evidenced by a number of devices and platforms that have enabled speech-to-text functions, notably Google voice search (Schalkwyk et al., 2010) and Windows speech recognition (Huang et al., 1993). The availability and performance of such technologies are further facilitated by smart mobile devices, which employ natural language user interfaces to provide a bidirectional communication channel between a user and a device. Apart from speech recognition that only places focus on vocabularies, speech emotion recognition (Pan, Shen and Shen, 2012) is also a highly active research topic that belongs to the speech and auditory modality. It can be employed to aid in e-learning by providing effective emotional exchanges. Automatic remote call centres can also benefit from automatically and timely detected customer emotions that can reveal customers’ dissatisfaction. Although low error rates are achievable by a few practical applications in this modality, the limitations should not be overlooked. Except for the low cognitive load of such systems, other impeditive factors include noisy environments, unstable recognition across different users, slow pace of speech output, the ephemeral nature of speech and difficulty in speech scanning and searching (Shneiderman, 1992).

In contrast to speech and auditory interfaces, *haptic interfaces* are enabled by another type of sensory data – the sense of touch. Being one of the most informative senses, haptic sensation combines both tactile and kinaesthetic sensations (Peres et al., 2008). As powerful as it seems to be, only until recently have haptic technologies gained the capabilities of delivering believable stimuli at a reasonable cost, and using human interface devices of a practical size (Brewster and Murray-Smith, 2000). Generally, haptic interfaces consist of sensors that are dedicated to capturing motions and forces exerted by a user, which are then applied to the operator by actuators. A classic example is teleoperation, also referred to as telerobotics, which extends human senses and operations to a remote environment. In various applications, haptic interfaces appear in different physical forms and complexity levels, from wearable devices to non-portable systems, and from single Degree-of-Freedom (DOF) to multi-DOF systems. One representative application is the MAHI arm exoskeleton (Gupta and O'Malley, 2006; Sledd and O'Malley, 2006), designed primarily for rehabilitation and training in virtual environments. It is capable of applying force and feedback to independent human joints due to it being a five-DOF haptic interface.

Apart from this application, Peres et al. identified (2008) that a large number of other interfaces were particularly beneficial to perception of limb movement and position, skilled performance of tasks with high precision and efficiency, virtual training in safe and repeatable environments. Nevertheless, one cannot overlook the barriers that hinder the applicability of the haptic modality. The primary limitation is the high cost for implementations of interfaces with high fidelity. This is mainly due to the requirements of high-resolution sensors, high DOF devices, powerful actuators and a complex mechanism that can effectively link together all these components.

This is where the visual modality exhibits its superiority. Despite all the advantages of haptic interfaces that have been illustrated, they can be replaced by *visual interfaces* in many cases due to their similar capabilities of performing structural encoding/decoding and describing object positions and orientations in a 3D space (Ballesteros and Heller, 2006). Arguably, the visual modality spreads a wider spectrum than this. The visual modality manifests a two-fold significance. On the one hand, visual sensors are capable of gathering visual signals, which are then processed to enhance the cognitive abilities (e.g. automatic object recognition) of vison based systems; on the other hand, visual interfaces can create immersive HCI environments by means of computer graphics (e.g. virtual reality). The former role of the visual modality focuses on computer perception and cognition while the latter role leans towards user perception and experience. The complementary link between its two roles provides the visual modality great potential for revolutionising HCI. Consequently, the visual modality has spawned many applications with different focuses. As an example of gaze tracking vision systems, smart solutions are starting to become available that monitor the gaze direction of a driver in order to identify driver distraction/drowsiness and provide timely alerts (Tawari, Chen and Trived, 2014). These driver assistance systems, capable of detecting and acting on driver inattentiveness, are of great value to road safety. Another application concerns a gesture recognition device – a sterile browsing tool, ‘Gestix’, which has been designed for doctor-computer interaction. This assistive technology provides doctors the sterility needed in an operation room where radiology images can be browsed in a contactless manner. A doctor’s hand is tracked by a segmentation algorithm using colour model back-projection and motion cues from image frames (Wachs et al., 2008). Regarding expression recognition, an intelligent tutoring system, namely a Learning Companion, is proposed to predict when a learner might be frustrated, initiate interaction depending on the user’s affective state and provide support accordingly (Kapoor, Burleson and Picard, 2007).

When these modalities are integrated, applications with richer functionality and enhanced intelligence can be forged. A recent application concerns a type of service robot that has been put into use in Bank of Communications (GBTIMES, 2015). This robot, named ‘Jiaojiao’ and acting as the bank manager in over 30 cities, can greet customers, introduce banking services, and answer inquiries. This is achieved by the incorporation of a number of enabling technologies, such as intelligent voice interaction, smart imaging, semantic recognition and biometric recognition (e.g. face recognition, gender recognition and age recognition). Nevertheless, the reliability of these emerging innovative technologies is not desirable enough, meaning that they are more appreciated for their ‘entertaining’ roles than their designed functionalities. In addition, inexpensive solutions need to be sought before these applications can find their ways to a wider range of HCI scenarios.

2.2 Eye/Gaze Analysis

Eye/gaze analysis is receiving an increasing amount of attention in HCI through utilization of the visual modality. Compared to gender and age recognition, it reveals more personal behavioural information by estimating the attention and intention of an individual. With eye/gaze analysis, a HCI system can not only observe its user passively, but it allows its user to take control of the system actively by responding to eye movement. Therefore it excels in remote and contactless interaction and provides an ideal channel for elderly people and those with motor disabilities to access HCI systems.

According to the features extracted, eye centre localisation methods fall into two main categories: inherent feature based methods and additive feature based methods. An additive feature based method actively projects near-infrared illumination toward the eyes, and results in reflections on the corneas, which are referred to as ‘glints’ in the literature (Zhu and Ji, 2005). Being highly reliant on dedicated devices, this method essentially alters the primary task of eye centre detection into corneal reflection detection as a simplified detection task. A passive inherent feature based method is more generalizable since it employs characteristic features from the eye region itself without the need for active illumination, and therefore becomes the method explored in this paper. The approach can be further divided into three categories: 1) geometric or morphological election methods that utilise gradient, isophote or curvature features to estimate the eye centre that comply with geometric or morphological constraints, 2) machine learning based methods where distinct features are extracted to train a classifier to search for the eye region that best matches the model representation, or 3) hybrid methods which normally follow a multi-stage scheme that comprises the previously summarised. While several methods have achieved interesting results, they have also exhibited their respective limitations.

One geometric feature based method (Timm and Barth, 2011) localises eye centres by means of gradients. In this approach, the iris centre obtains the maximised value in an objective function that peaks at the centre of a circular object. This method has achieved high accuracy since its eye model is capable of dealing with deformation of a circular pupil/iris contour in an image, which is likely to be caused by image noise, head pose, pupil position and undesirable illumination. In addition, it operates in real time and maintains relatively high accuracy on low resolution images. However, its performance declines in the presence of strong gradients from eyelids, eyebrows, shadows and occluded pupils that overshadow iris contours. This is known to be a common problem suffered by most eye centre localisation methods and remains unsolved to date. Another unsupervised method using geometric features investigated the Self-Similarity Space, where image regions that can maintain particular characteristics under geometric transformations receive high self-similarity scores (Leo et al., 2014). This eye model is derived from the relative rotational invariance of a pupil/iris region. As a result, the extraction of the pupil/iris region directly affects the computation of self-similarity scores as the inclusion of eyebrows and other interfering sharp edges can produce high self-similarity scores that surpass those from the pupil/iris. Regarding machine learning based methods, for all algorithms that utilise extracted features to train a model (Niu et al., 2006) it holds that the training data are of critical influence on the performance of the algorithms (Zhu and Ramanan, 2012). More specifically, variations posed by illumination and head rotation have a huge impact on the accuracy and robustness of most algorithms. These are the main factors that prevent many algorithms from finding their way to real-world implementations. Inspired by Fisher Linear Discriminant (FLD) (Duda, Hart and Stork, 2012), Kroon, Hanjalic and Maas (2008) designed a linear filter trained by the image patches extracted from normalized face images. This method not only considers the high response from the filtered image, but also examines a rectangular neighbourhood around the estimated eye centre positions. This is based on the observation that a pupil in an image is formed by a collection of dark pixels within a small region. Another machine learning based method (Niu et al., 2006) focuses on the design of a novel classifier rather than the extraction of representative features. This method introduces a 2D cascade AdaBoost classifier that combines bootstrapping positive samples and bootstrapping negative samples (Viola and Jones, 2001). The final localisation of an eye is achieved by the fusion of multiple classifiers.

In general, unsupervised methods have the advantage that they are independent of training data. Therefore, they are less likely to be biased toward a certain type of environmental setting. As a result, they can be more adaptive to dynamic environments. For unsupervised methods, the models that characterise the eye regions are vital in determining the accuracy and robustness of the resulting algorithms.

Many eye centre localisation methods are designed for only frontal faces and thus deteriorate with the presence of head rotations and/or eye movements typical of an unconstrained real-world application. Asadifard and Shanbezadeh (2010) employed a cumulative distributed function (CDF) for adaptive centre of pupil detection on frontal face images. Their approach firstly extracts the top-left and top-right quarters of a face image as the regions of interest and then filters each region of interest with a CDF. An absolute threshold is defined for the filtering process given the fact that the pixels in the pupil region are darker than the rest of the eye region. This method only accounts for frontal faces; and the eye model in this method only considers the intensity values of an eye region in a greyscale image. Only when a complete pupil can be extracted by means of erosion would this method give an accurate estimation of the eye centre. However, under realistic scenes, specularities on a pupil or on a pair of spectacles will split the pupil into several disconnected regions while self-cast shadows would easily change the values calculated by a CDF. Another study on frontal faces (Türkan, Pardas and Cetin, 2007) explored edge projections for eye localisation. With a face image available, their method firstly defines a rough horizontal position for the eye region according to facial anthropometric relations. After the eye band is cropped, it gathers eye candidate points that are extracted by a high-pass filter of a wavelet transform. A Support Vector Machine (SVM) based classifier (Chang and Lin, 2011) is then used to estimate the probability value for every eye candidate. This type of method normally requires that all face images are perfectly aligned so that the facial geometry agrees with facial anthropometric relations as the prior knowledge. Any misalignment will cause inconsistency to the features and thus lead to poor results.

Although recent studies have shown promising results in accurately localising eye centres, the estimation error increases at relatively long distance and is affected by both shadows and specularities. Nevertheless, the practical value of eye centre localisation methods has been exploited by researchers to boost HCI experience through various means, for example, via gaze gestures. As opposed to absolute eye fixation points, gaze gestures are sequences of relative eye positions that reflect eye gaze shifts in the spatial-temporal domain. Although gaze gestures are not intended for estimation of absolute gaze positions, they are favoured for being free of calibration and robust to variations caused by head movement and user-camera distance. A similar paradigm to gaze gesture based HCI employs hand gesture, which is however inconvenient to users with motor disabilities.

Drewes, Luca and Schmidt (2007) carried out a study on eye-gaze interaction for mobile phone use following two methods, the standard dwell-time based method and the gaze gesture method. Proposing to implement an eye tracker on a mobile phone platform, they further designed a number of gaze gestures which, upon recognition, can trigger certain actions such as scrolling up and down a phone book, or opening and closing an internet browser. This study concludes that gaze gesture is robust to head movement since it only captures relative eye movement rather than absolute eye fixation points. Calibration is also unnecessary and this therefore makes eye gesture more suitable for real-world applications. The two interaction methods are further compared by a recent study (Hyrskykari, Istance and Vickers, 2012) where the participants were using either gaze gestures or dwell icons in the context of a 3D immersive game. At the end of the experiment, they evaluated the task completion time, selection error and missed gestures or clicks so as to compare the two types of command input method. They suggested that "gaze gestures are not only a feasible means of issuing commands in the course of game play, but they also exhibited performance that was at least as good as or better than dwell selections". Another study (Rozado, Rodriguez and Varona, 2012) has achieved gaze gesture recognition for HCI under more general circumstances. It employs the hierarchical temporal memory pattern recognition algorithm to recognise predefined gaze gesture patterns. 98% accuracy is achieved for ten different intentional gaze gesture patterns. Some other works on gaze gestures dedicated to HCI have similar limitations. Firstly, they all depend on active NIR lighting for eye centre localisation. Secondly, the eye centre localisation algorithms only work at short distances.

In summary, the attention that eye/gaze analysis receives has never faded despite the variety of observed facial features. A facility for eye centre localisation and exploitation of its practical application offers huge potential for HCI applications. The challenges of this research area largely arise from existing limitations, e.g. poor illumination conditions that create shadows and specularities around the eye region. Further complications arise from changes in head pose, eye movement, long distance scenarios, and dependency on dedicated devices or complex system structures.

Apart from the studies reviewed in this section, a detailed summary of ten state-of-the-art methods is provided in Section 4.2, with a comparison with the proposed eye centre localisation method.

2.3 Gender and Age Recognition

This section reviews a number of 2D and 3D methods for gender recognition and age recognition. In addition, ten state-of-the-art gender recognition approaches are further reviewed in Table 5.1 with a detailed comparison with the proposed gender recognition method in terms of features/classifiers, recognition accuracy, database, validation method and limitation(s). As most studies consider gender and age recognition as classification tasks where a gender label or an age label is classified as belonging to a particular category, the terms ‘gender/age recognition’ and ‘gender/age classification’ are normally used interchangeably.

Gender recognition from facial images is a challenging task in that a face exhibits a broad range of intra-class variations due to diverse facial attributes or dynamic environmental factors. The former complications mainly include age, ethnicity and makeup while the latter include illumination condition, head pose, facial occlusion and camera quality. Most high-performance gender recognition methods involve machine learning and follow four stages: face detection, facial image pre-processing, feature extraction and classification (Ng, Tay and Goi, 2012).

For the face detection stage, the Viola-Jones face detector (Viola and Jones, 2004) has been widely adopted due to its ease of implementation and relatively high accuracy. It is essentially a face detector that employs Haar-like features, a classifier learning with AdaBoost and a cascade structure (Viola and Jones, 2001). It can operate in real time and has contributed to the implementations of many practical applications in the last decade.

For the image pre-processing stage, normalisation, i.e. contrast and brightness adjustment; image resizing and face alignment are commonly considered useful despite their varied implementation details. Among them, face alignment has been reported to be able to guarantee an increase in the classification accuracy by a number of studies. For example, in a research (Mäkinen and Raisamo, 2008) evaluating a number of gender classification methods, it was concluded that SVMs outperformed other classification methods with 86.54% accuracy on 36×36 aligned images, and that higher accuracy could be achieved by improving the implementation of the automatic alignment methods. Another research (Mäkinen and Raisamo, 2008) with regard to gender classification illustrated that face alignment brought an increase to the classification accuracy for various methods including use of neural networks, SVM, and AdaBoost. Different pre-processing methods are experimented with in the proposed method with the results reported in Section 5.3.

For the feature extraction and selection stage, a wide range of features have been experimented with and evaluated in the literature, 2D or 3D, densely extracted or sparsely detected, driven by the critical need for higher robustness and discriminability. They include intensity values from greyscale images (Moghaddam and Yang, 2000), LBP (Shan, 2012; I. Ullah, 2012), facial strips (Lee, Huang and Huang, 2010), Haar-like features (Viola and Jones, 2001), SIFT features (Wang et al., 2010), etc. Depending on the type of features extracted, one or more face descriptors per image are obtained. These descriptors are commonly drawn to characterise facial texture, geometry or topology whose representations seek to obtain high robustness to intra-class variations (e.g. invariance to facial expression, head pose, illumination, etc.).

For the classification stage, SVMs and neural networks have been the most popular classifiers. SVMs with different kernels were investigated in (Moghaddam and Yang, 2000) and convolutional neural networks (CNNs) were adopted by Tivive and Bouzerdoum (2006) and Phung and Bouzerdoum (2007) as gender classifiers.

Following the four major stages, a number of approaches have reported relatively high classification rates on publicly available datasets. However, most of them have only been evaluated on controlled database and the reported high classification rates are conditional to certain data arrangement and validation methods (see Table 5.1 for a summary).

For example, a decision-fusion based method was presented by Alexandre (2010) that utilised multiple SVMs to classify greyscale values, local binary patterns and histograms of edge directions as features. The three types of features were extracted from images of various sizes. Finally all classification results were integrated to make the final decision by means of majority voting, leading to 99.07% accuracy. However this result was obtained from a small subset of the FERET database and their validation method was not sophisticated enough to reflect the performance of their approach objectively. In addition, only controlled databases were used for training and testing in this research so that the applicability of this approach to dynamic environments remained unevaluated. Similarly, Lee, Huang and Huang (2010) employed ten regression functions to conduct region-based classifications and fed the vector of classification results into an SVM to generate the final decision. Despite the 98.8% accuracy with the FERET database they reported, they didn’t illustrate their evaluation method and the split of training and testing data. The reappearance of the same subject in both the training and testing data might account for the high accuracy they obtained. A face alignment scheme is compulsory to their approach, the absence of which leads to a 6% drop in the classification rate, bringing 98.8% down to 92.8%. This is an inherent limitation of conventional region-based approaches where defined facial regions have to be perfectly aligned. Hu, Yan and Shi (2010) proposed a fusion-based method for gender recognition by integrating different facial regions using the ‘matcher weighing fusion’ method. The facial landmarks for segmenting the face into its sub-regions are detected by a profile-based method and a curvature based method. Interestingly they prove experimentally that the fusion of multiple facial sub-regions is superior to the complete face region alone and that the upper face contains more discrimination ability regarding gender classification.

While features obtained from colour or greyscale images are still attracting huge research focus, the trend has now been directed toward the study of 3D face features. One of the studies (Hu, Yan and Shi, 2010) achieved high accuracy by integrating multiple facial regions segmented by a selection of facial landmarks. 3D features extracted from these regions were then classified by a SVM. This study evaluated the contributions of individual facial regions and concluded that the upper facial region contained the highest gender discriminability. Another study (Fagertun, Andersen and Paulsen, 2012) investigated “gender strength”, which was aimed at replacing the conventional binary gender labels by introducing a continuous gender class variable. Cognitive tests were performed on 3D face scans to calculate “gender strength”. Since 3D features have shown promising results, they have also been combined with 2D features such that the merits from both types of features can be inherited. Huynh, Min and Dugelay (2012) introduced a type of LBP based feature descriptor to encode 3D facial features for gender classification. A combination scheme that makes use of both depth images and greyscale images was proposed, which showed enhanced classification accuracy on both high and low resolution data. Xia et al. (2013) also proposed a fusion method for gender recognition that combined the shape and texture features extracted from 3D meshes and greyscale images. The fused feature proved to outperform individual feature types.

Based on the individual works reviewed for gender recognition, conclusions can be drawn from the literature regarding the preferences and trends in gender classification. 1) SVMs and neural networks are the most popular classifiers. 2) LBP and its variations are the most popular features. 3) Most studies use the FERET database as the standard evaluation database. 4) Most works are carried out under well-controlled environments while real-world implementations and evaluations lack exploitation. 5) Most works incorporate face alignment in the pre-processing stage. 6) Most works employ the five-fold cross validation for accuracy estimation. 7) Although 3D features have received an increasing amount of study, the corresponding 3D imaging systems and implementations are not keeping the pace with the algorithms. Therefore, 2D methods remain the mainstream for gender recognition.

Apart from gender label of an individual playing a significant role in natural face-to-face interaction scenarios, an age label provides an additional demographic cue that assists with the formation of successful interaction/communication strategies. Therefore, automatic age recognition generally resembles gender recognition and should be valued as much as the latter. However, age recognition is a more demanding task in that age labels normally go beyond two classes, as opposed to gender labels. Consequently, research on automatic age recognition appears to be more limited and less explored relative to that on gender recognition, in terms of both the breadth and outcomes of studies. To facilitate HCI, efforts are mainly seen in the speech modality and the visual modality for higher age recognition accuracy. One example (Siegert et al., 2012) concerns the speech modality which is adopted to identify gender and age of an individual simultaneously at the initial stage of a HCI session. In this example, a hierarchical classification strategy is designed to chain an age classifier and a gender classifier together such that prior age information is fed to the gender classifier, and vice versa. Although this strategy outperforms classical strategies where only one classifier is trained to resolve a four-class (young male, young female, senior male and senior female) classification problem, the weighted average classification rate is only 69%, which is far from accurate and reliable enough for useful applications. It should also be noted that most studies classify the age of a subject as belonging to a certain age range, e.g. between

40 years old and 60 years old, rather than an ambitious attempt to predict the exact age label. This is due to the complexity of the task itself, limited age labels of publicly available databases, inadequate reliability of existing age recognition methods and complications in algorithm evaluation due to overwhelming age categories. For example, a few studies have defined age recognition as a seven-class problem: Gallagher and Chen (2009) introduced contextual features that could represent face structures for seven age groups rather than for individuals. When they combined contextual features with appearance features, the highest classification rate achieved was 42.9%. While this result seemed far from accurate, they conducted a ‘less stringent evaluation’ by considering age predictions that fell under the two neighbouring age groups of the ground truth as accurate. This evaluation method brought the classification rate up to 78.1%. By introducing this error tolerance zone, the evaluation became less convincing, and yet the results still seemed unreliable. As the LBP features are believed to be an excellent descriptor for facial textures (e.g. facial wrinkles and furrows) that are closely related to appearance caused by aging, LBP and its variants are considered effective for age recognition. In a similar evaluation setting, Shan (2010) employed the boosted LBP features and a SVM for age recognition on a real-life face database. This method achieved a classification rate of 50.3%, with 87.1% for the ‘less stringent evaluation’. Another study (Ylioinas, Hadid and Pietikainen, 2012) experimented with the encoded LBP features and compared different parametric settings. This type of feature with the optimal parametric setting achieved a 51.7% classification rate, with 88.7% for the ‘less stringent evaluation’. In these studies, the division of age groups produced seven age groups: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65 and 66+ (years old). The highest classification rate in the rigid evaluation is only around 50%. Although the classification rate in the less rigid evaluation is relatively high, the average error for the predicted age can be extremely large. For example, when a 20-year-old (who belongs to the age category 20-36) is classified into the neighbouring category 37-65, this is somewhat oddly deemed an accurate prediction by this evaluation method. When an 80-year-old (who belongs to the age category 66+) is classified into the neighbouring category 37-65, this prediction is also deemed accurate. This means that the case where a 20-year-old and an 80-year-old

are classified into the same age group is considered an accurate prediction by this evaluation method, causing the results to be unreliable.

More recently, a number of studies and applications have shown interesting results. An application ‘How-Old.net’ (Microsoft, n.d.) has been developed to estimate age labels from facial images. While the evaluation results of this application have not been revealed, the surge of attention drawn by his commercial development is accompanied by criticisms from its users for its unstable predicted results. In addition, in the medical field, a research study (Chen et al., 2015) examines 3D facial morphologies in an attempt to provide insight into the human aging process. It suggested that 3D facial features could serve as “more reliable aging biomarkers than blood profiles” and could “better reflect the general health status than chronological age”.

Despite the large number of attempts that investigate various types of features, classifiers and modalities for age classification, the accomplishments to date have not been particularly satisfactory. When operating on multiple age classes, age recognition fails to generate accurate and reliable results, as shown by the literature reviewed. This means that age classification algorithms need to perform reliably on binary classes before they are extended to resolve a multi-class classification task.

2.4 3D Imaging Techniques

As introduced in the preceding chapter, the recent trend of computer vision is weakening the domination of 2D image (e.g. colour or greyscale) analysis. 3D imaging systems and 3D acquisition/reconstruction algorithms are replacing or updating established 2D techniques, though their practicability and accessibility are still far from being desirable.

For many decades, what people have taken for granted in human vision systems is what people struggle to acquire and process with computer vision systems. This leads to the exploration of the 3D vision that can best represent scenes and objects with three dimensions. Once transformed from the 3D world-space to the 2D imaging-plane, the spatial properties of original scenes can no longer be preserved and easily analysed, and thereafter the focus is shifted to colour channels or greyscale values.

In the broad visual modality, 3D imaging techniques have been applied to wide-ranging and diverse application domains. Aside from general object detection and recognition tasks, those intended for HCI include face recognition, facial expression recognition, gesture recognition, body detection, action recognition, etc. Being able to recover spatial properties of objects and environments, 3D features have been reported to surpass 2D features in relation to their reliability and robustness. This phenomenon can be intuitively inferred since 3D features are independent of pose and ambient illumination conditions. Moreover, the variation caused by the camera viewpoint can also be compensated through the rotation and translation of the 3D coordinate system.

Commonly, 3D features are obtained via various means including structured-light (Scharstein and Szeliski, 2003) 3D scanning, 3D laser scanning (Blanz and Vetter, 2003), stereoscopic systems and photometric stereo (PS). The principles of these 3D imaging techniques are briefly stated as follows:

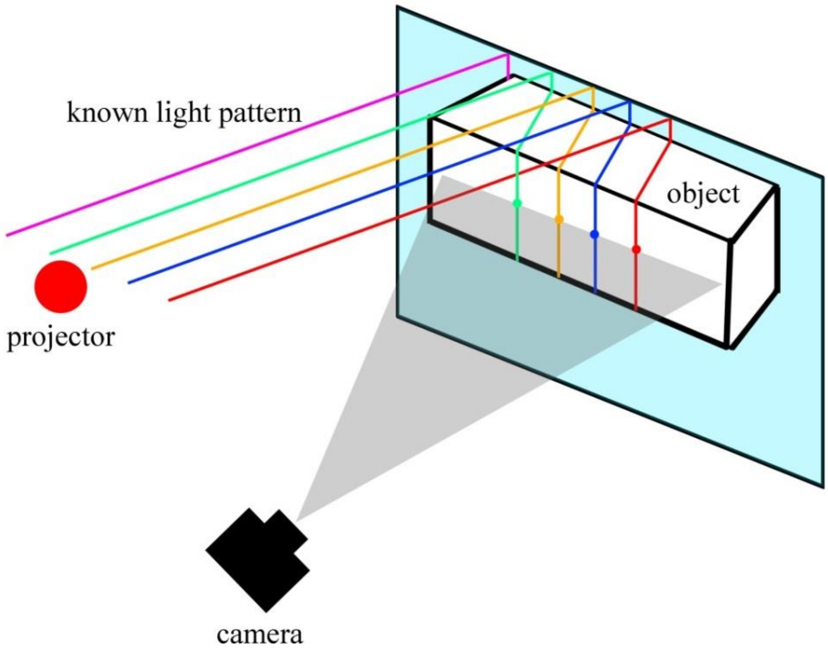


Figure 2.1 The principle behind structured-light 3D scanners, which is based on the distortion of a known light pattern observed from a certain viewpoint.

A *structured-light 3D scanner* employs projected light patterns and a camera system to reconstruct a surface shape (see Figure 2.1). The known light patterns deform upon arriving at a surface, which allow surface depth to be calculated. The structured light principle has been famously employed by the Microsoft Kinect sensor (V1) as a low-cost solution to 3D imaging. The accuracy, resolution and other performance indicators of the Kinect sensor are separately reviewed in this section.

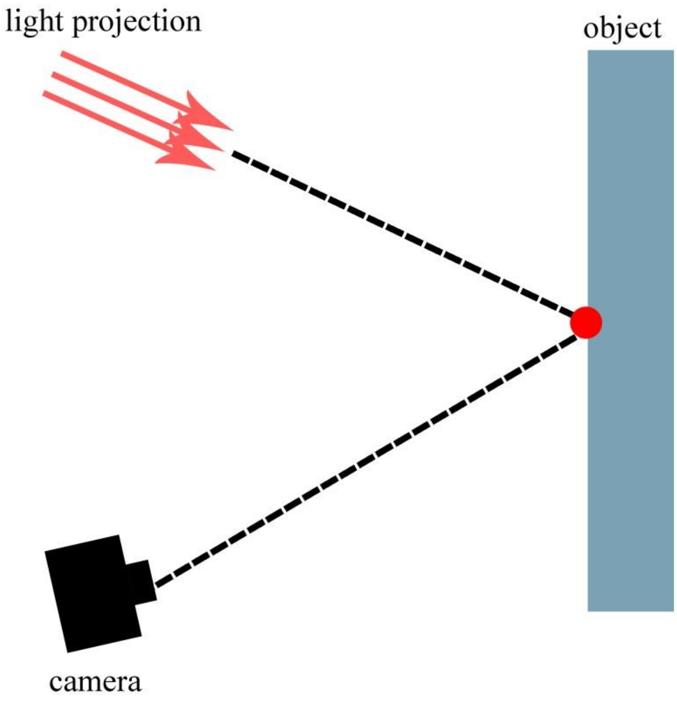


Figure 2.2 The principle behind 3D laser scanners: detection of reflected light from a projector.

3D laser scanners commonly appear in two forms – the *time-of-flight 3D laser scanners* and *triangulation based 3D laser scanners*. Both types of scanners emit a pulse of light towards the target object and detect the reflected light (see Figure 2.2). While the former type calculates the travel distance of the light by measuring the travel time of the light, the latter type employs a camera to locate the laser dot on the target object and solves the trigonometry problem involving the triangle uniquely determined by the laser dot, the camera and the laser emitter. Following a similar principle, LIDAR can provide accurate and high-resolution 3D data. It is commonly adopted for remote sensing (e.g. atmospheric sensing and imaging of buildings) but normally incurs a high cost.

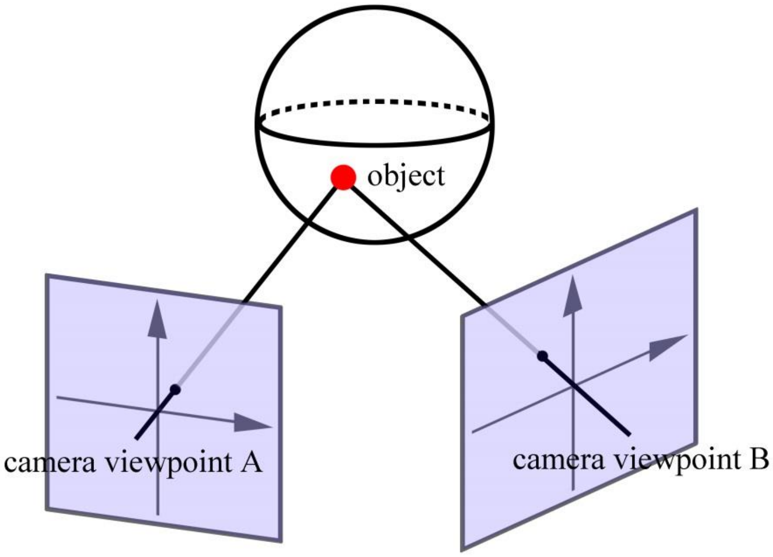


Figure 2.3 The principle behind stereoscopic vision systems. When two or more images of an object point are captured, the 3D position of the point is where the two projection rays intersect.

Stereoscopic systems are inspired by the human vision system where two eyes naturally enable binocular vision. This type of system captures a pair of images with two cameras at slightly different viewpoints and fuses them to derive depth information (see Figure 2.3). However, the computation of the disparity between the two images is not so straightforward since it gives rise to what is commonly known as *the correspondence problem*, which is responsible for causing prolonged 3D reconstruction time. As an extension of the two-view stereo vision, multi-view stereo vision is also subject to the correspondence problem during the triangulation process. In addition, when images from more camera viewpoints are required, a stereo vision system will be more complex and expensive.

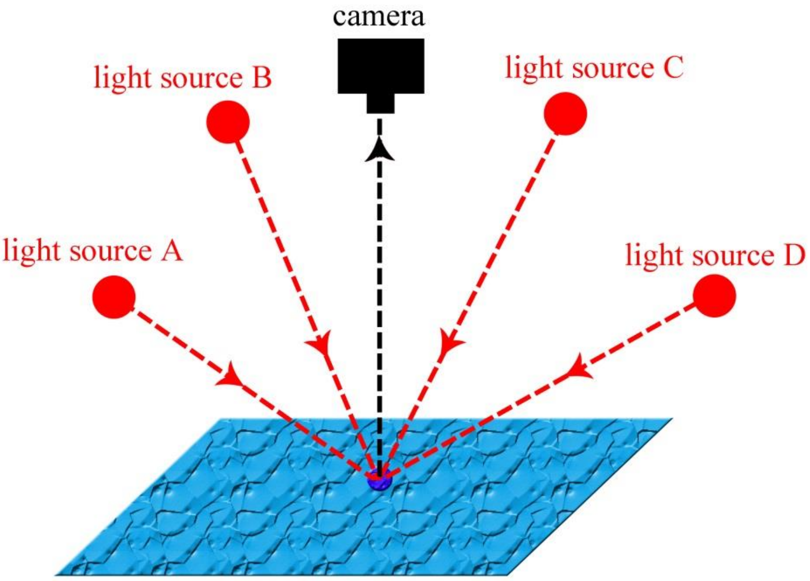


Figure 2.4 The principle of photometric stereo which employs a single camera to capture multiple images of a surface illuminated by multiple light sources.

Photometric Stereo is a technique which, rather than calculating a depth image or a point cloud, recovers a surface normal field of an object illuminated from different directions while the viewing direction is held constant (see Figure 2.4). This technique is fundamentally based on the fact that the fraction of the incident illumination reflected in a particular direction is dependent on the surface orientation. Therefore, when the directions of incident illumination are known and the radiance values are recorded, the surface orientation can then be derived.

The *Microsoft Kinect sensor (V1)* is a device launched in 2010 and has since become one of the most favoured enabling technologies for 3D sensing. The Kinect V1 employs an infrared projector and an infrared camera for estimation of depth maps, which is achieved by capturing and analysing the dotted IR patterns emitted by the infrared projector. Although being fundamentally based on the structured light principle, the Kinect V1 also combines the depth from focus technique and the depth from stereo (parallax) technique to improve the accuracy of depth map calculation. The newest generation of this sensor, the Kinect V2, is based on the time-of-flight principle, which brings improvements to its field of view, image resolution and other features. As a result, it contributes to professional studies but also puts 3D imaging within the reach of amateurs for regular use.

One of its prevalent applicable areas concerns recognition or tracking of human body and body parts. This is commonly achieved by representing a human body by a number of joints which form a skeletal model. Recognition and tracking of the skeletal model can be then enabled by unsupervised or supervised classification methods (Alexiadis et al., 2011). This has further contributed to the advancement of HCI (Lai, Konrad and Ishwar, 2012), human activity analysis (Reddy and Chattopadhyay, 2014) and physical rehabilitation (Chang et al., 2012). Other applicable areas receive benefits from the combination of 2D and 3D data by use of the Kinect sensor. Existing and potential applications in this regard include indoor 3D mapping (Henry et al., 2012), visual enhancement (Hu et al., 2013) and tele-immersive conferencing (Zhang, 2012).

2.4.1 Advantages and limitations of Photometric Stereo

While the Kinect V1 and V2 as a low-cost solution to 3D imaging can be deemed accurate in some scenarios, a few limitations can be seen. Firstly, the horizontal field of view of the Kinect V1 is 57 degrees (70 degrees for the Kinect V2), which is lower than normal webcams; its nominal depth range is between 0.8 metres and 3.5 metres. This indicates that there exists a relatively large ‘blind zone’ which cannot be ‘observed’ by the Kinect. Although other software frameworks can deliver depth values up to over 9 metres, the decreased resolution and accuracy as well as increased noise level makes it much less reliable beyond a 4-metre distance. Secondly, its nominal spatial sampling step (distance between two adjacent sample points) at a 2-metre distance is around 3 mm, with nominal depth sampling step being 10 mm, which results in sparse sampling points. At a 6-metre distance, the nominal depth sampling step is further increased to 100 mm. This is due to the low resolution of IR depth image (Amon, Fuhrmann and Graf, 2014) captured by the Kinect sensors (320×240 for the Kinect V1 and 512×424 for the Kinect V2). Thirdly, the Kinect (V1) performs depth estimation by calculating the distance between an object to the camera-laser plane, rather than the actual distance between the object to the image sensor. Therefore, when an object appears in the marginal areas of its filed-of-view and at short distance, the error incurred by this approximation may not be negligible. In practice, depth values measured at 2 metres can deviate by as much as 40 mm. Similar results regarding the evaluation of the accuracy and resolution of the Kinect can be found in other works which conducted theoretical and/or experimental error analysis (Khoshelham, 2011; Khoshelham and Elberink, 2012). In comparison, the PS technique allows for 3D reconstructions of high spatial and depth resolution and thus can serve to reveal detailed 3D textures. Its performance is not severely impaired by increased imaging distance which causes the Kinect reconstructions to be sparse and inaccurate. As well as outperforming the Kinect devices, PS systems exhibit more advantages when compared to many other 3D imaging systems.

While some 3D imaging systems are mechanically complex and expensive (e.g. a structured light 3D scanner) and other are computationally expensive (e.g. a stereoscopic system), *photometric vision systems* manifest better feasibility for real-world applications in that 1) they require only inexpensive and simple settings, 2) they are capable of performing real-time 3D reconstruction (Malzbender et al., 2006), 3) they calculate surface normal data that are object-centred while most other 3D imaging techniques obtain image-centred data, and that 4) they provide superior reconstruction results that reveal 3D textures. The only limitation of the PS technique is the need for active illumination. But this has often been viewed as a merit in practice since active illumination allows the PS technique to be applicable to dark environments. Therefore, a variation of the PS technique has been adopted by this research for 3D face reconstruction on account of its various advantages and benefits.

The PS technique has been favoured by many researchers and scholars for a couple of decades. It was originally introduced by Woodham (1980) who proposed a method that could deal with predominantly Lambertian surfaces illuminated by distant point light sources. This method proved that a minimum of three reflectance map contours, i.e. three PS images, could uniquely determine surface orientation of an object. Thereafter, some works have extended this method to more general conditions, for example, in the presence of specularities and/or shadows by use of four light sources (Coleman and Jain, 1982; Solomon and Ikeuchi, 1996; Barsky and Petrou, 2003). Other PS variations explored the cases where the light sources cannot be treated as point sources (Farooq et al., 2005; Lee et al., 2005; Smith, 1999; Smith and Smith, 2005), or those where the target Lambertian object is in motion (Lim et al., 2005). Not only has the PS technique developed theoretically, its embodiments as imaging systems or practical implementations have also validated its persistent advancement.

For example, a custom-made four-source PS device was designed by Hansen et al., 2010, with hardware configuration that can be easily deployed in commercial settings. This PS device (Figure 2.5, used with permission of the author) mainly consists of 4 visible light sources, a high speed camera and an ultrasound proximity sensor.

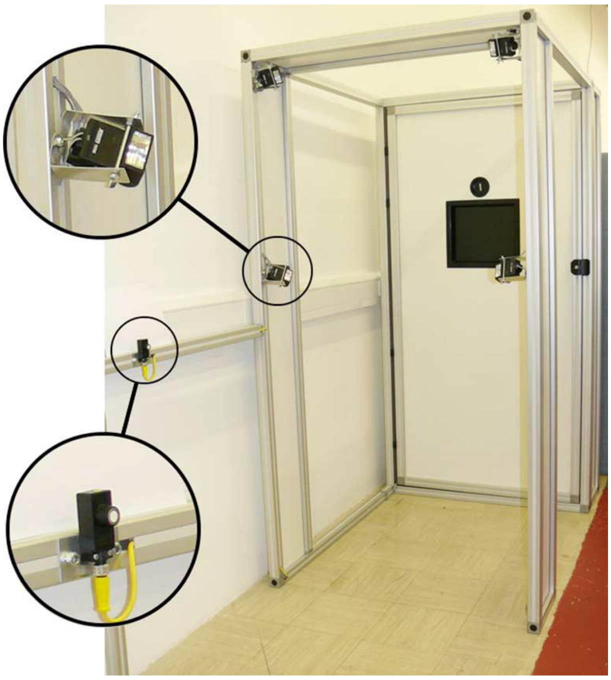


Figure 2.5 The Photoface capture device. One of the light sources and the ultrasound trigger are shown in the enlarged areas (Zafeiriou et al., 2013).

During an experiment, when an individual walks through the archway and triggers the ultrasound proximity sensor, the light sources will be lit up alternatingly and allow a set of four PS images to be captured by the camera. PS reconstruction is then performed for surface normal estimation. This hardware setting has an alternative configuration, in which the visible light sources are replaced by near infrared (NIR) lights. As well as being invisible to human eyes and thus less intrusive, the employment of NIR lights also brings higher reconstruction accuracy compared to visible lights. This device, along with the reconstruction method, has been applied to face recognition and verification using 3D features, yielding promising results.

Apart from this system designed for face recognition, the PS method is also beneficial to military and security applications. By extracting 3D features that are free from the confusion of textured camouflage, a strategy based on the PS was proposed to detect and recognise concealed objects (Sun et al., 2009). Security systems can thus be designed such that the presence of suspicious or dangerous concealed objects in mass transport environments can be revealed.

PS techniques are superior in capturing detailed high-frequency 3D textures and are less affected by image noise compared to triangulation based techniques. The superiority of surface normals recovered by the PS method (and its variations) is evidenced by a number of works that compared a range of data formats and well-known face recognition algorithms.

For example, instead of following the convention that depth maps and point clouds are frequently used for 3D face representations, Gökberk, İrfanoğlu and Akarun (2006) employed additional 3D facial features including surface normals, shape index values and other representations based on 2D depth images. In their experiments on the 3D-RMA dataset, the surface normal features obtained the highest accuracy (97.72%), while point cloud features achieved 92.95% accuracy and depth image based feature performed much worse. Similarly, Hansen (2012) evaluated different face recognition methods on both 2D and 3D face representations including texture (2D), LBP (2D), depth map (3D), surface normal (3D) and shape index (3D). According to his findings, surface normals and shape index features achieved the highest accuracy on average, i.e. 95.73% and 95.96%, respectively. Depth map as a type of 3D feature achieved the worst results, i.e. 73.34% on average. These works imply that surface normals recovered by the PS method are better 3D face representations compared to depth maps obtained by other approaches reviewed in this section. In addition, PS methods normally require only one camera for image capture, simplifying the calibration process and allowing for high efficiency. These are the main reasons that photometric approaches have initiated a broad research field with great application prospect, as they promise to bring cutting-edge capabilities of computer vision into practical use.

Chapter 3 Development of a 2D+3D Imaging system

As reflected by the review of HCI studies in the literature, the essence of HCI designs resides in the comprehension of user characteristics and behaviours, which reveal user expectations and can be utilised to improve user satisfaction and experience. This particularly addresses the role of human involvement in the HCI process and conveys the demand that HCI theories have to be reflected and implemented from an applied perspective. To this end, a 2D+3D imaging system has been developed to accompany the algorithms and methodologies proposed by this research. The design of this system is intended 1) to gather 2D and 3D facial data in real-world environments for algorithm evaluations; 2) to provide a platform that integrates the proposed algorithms – gender recognition, age recognition, eye centre localisation, gaze gesture recognition and PS 3D reconstruction; and 3) to act as an experimental HCI system that demonstrates the applicability of the proposed algorithms and HCI strategies.

3.1 System Structure

The structure of the 2D+3D imaging system is shown in Figure 3.1. It is intended to gather PS facial data as well as colour images in real-world environments. The design of the system consists of 1) a high-definition (HD) 47-inch display, 2) a webcam (referred to as camera 1 in the rest of the thesis) operating at 640×480 resolution, 3) two near-infrared (NIR) LEDs (SFH4232 with 850 nm wavelength) for PS illumination, 4) a Point Grey

GS3-U3-41C6NIR-C camera (referred to as camera 2 in the rest of the thesis) operating at 2048×800 resolution, with a 850nm +/-5nm NIR bandpass filter 5) a PC in the cabinet for data storage and processing, and 6) a control unit that synchronises NIR LEDs with the cameras. NIR LEDs with the wavelength of 850 nm have been employed in that they are widely available off the shelf, and that the camera employed has a strong spectral sensitivity to such a wavelength. The cameras are 2.1 metres from the floor and the NIR illuminators are both 0.75 metres from the cameras. Among other experiments for determining the optimal light positions, this arrangement produced the smallest shadowed facial area around the nose while providing even illumination to the face.

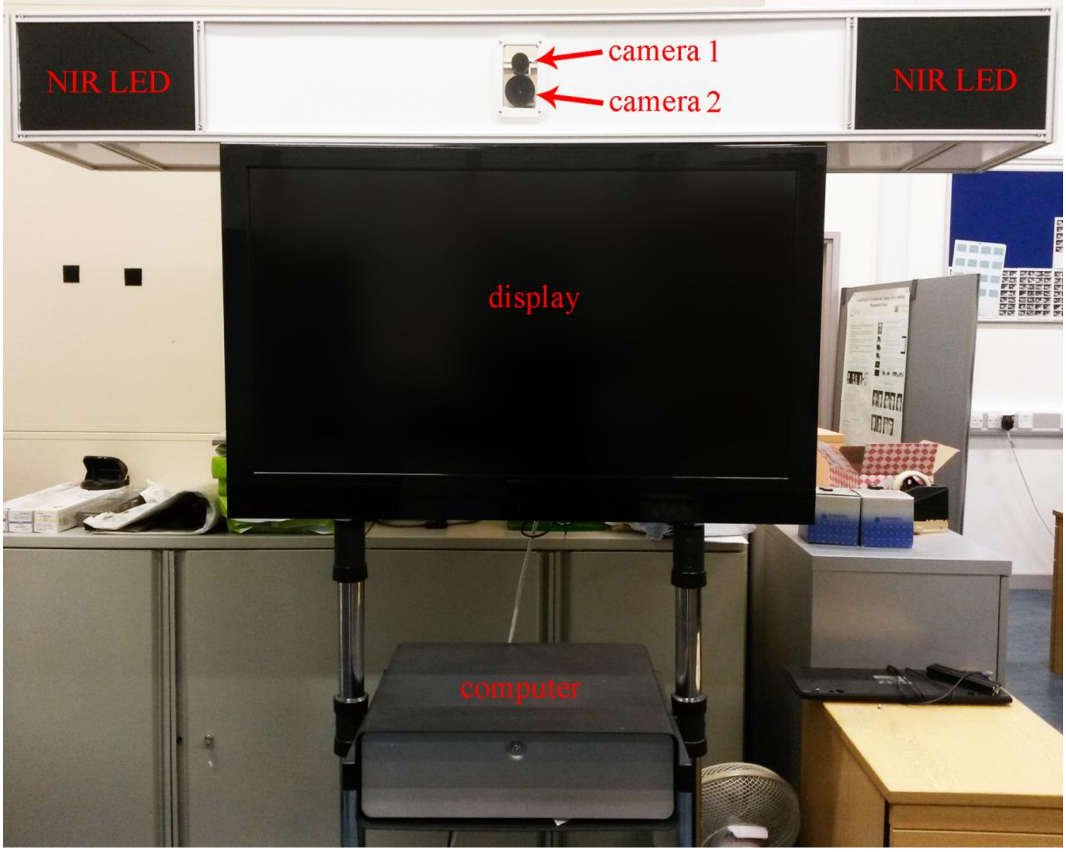


Figure 3.1 System structure for the 2D+3D imaging system

The two NIR LEDs are concealed within a housing and behind black acrylic covers, which offer high levels of NIR light transmission. This design is discreet and ensures the safe use of NIR illuminators.

3.2 Data Capture Experiments

The data capture experiments are designed to explore two tasks: 1) gender recognition evaluation and 2) eye centre localisation evaluation. The 2D+3D imaging system was employed at this stage for data capture. Therefore, in the rest of the thesis where the data capture experiments are concerned, the two terms '2D+3D imaging system' and 'data capture system' are used interchangeably.

The data capture process is described as follows:

1) The data capture system was firstly placed at a university public kitchen area with the presence of only artificial light sources. A total number of 45 volunteers participated in this experiment in 2 different recording sessions.

2) The system was then placed at a university library foyer where lighting conditions were affected by both lamps installed in the foyer and sunshine through the window. A total number of 127 volunteers participated in this experiment in four different recording sessions.

3) In the overall 6 experimental sessions (which provided sufficient data with adequate variations), every volunteer was asked to stand at one metre away from the display and look at the display where a green dot appeared at seven different locations on the screen in a sequence (see Figure 3.2). This pattern of green dots was employed to cause head/eye movement of different magnitudes/orientations such that the data captured could incorporate more variations and therefore better represent real-world scenarios. Different arrangements of this pattern should have a similar effect. Every volunteer was asked to look at the green dot in a natural manner. This incurred head rotations and/or eye movements. The NIR LEDs and the camera were synchronised so that, for every position of the green dot, the two NIR LEDs illuminated alternatingly while camera 1 captured colour images of a volunteer and camera 2 captured PS image sets (two NIR images in each set) of the volunteer. It should be noted that since camera 2 was covered by a NIR filter, the ambient light posed negligible impact on camera 2; while the NIR LEDs had

minimal impact on camera 1 due to an infrared cut-off filter physically attached to this camera.

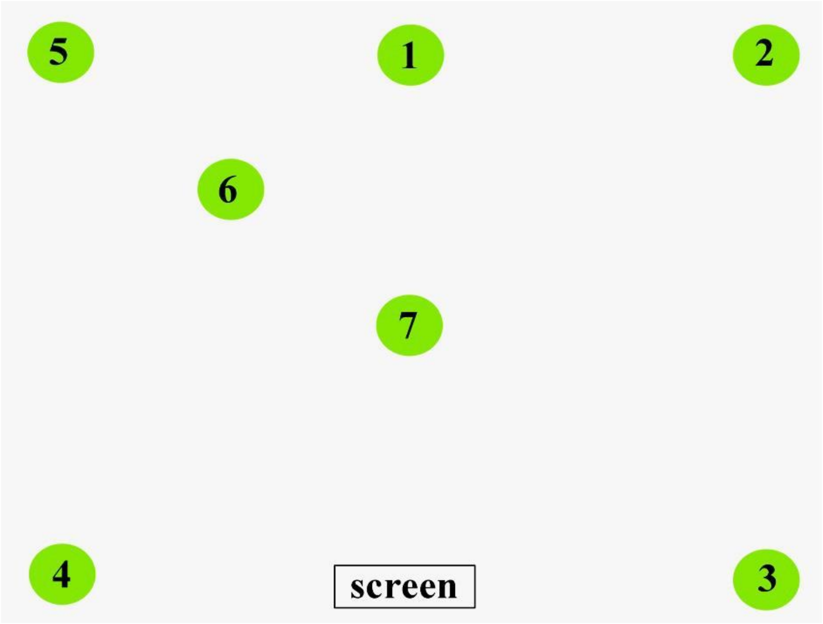


Figure 3.2 Calibration signs for the data capture experiments. The numbers within the dots represent the order in which these dots appear.

4) The overall 6 recording sessions gathered image data of 90 male subjects and 82 female subjects. The image data contain Caucasian, Asian and African faces, with an age range from 18 to 58 years. The image data (from 172 subjects) were employed for eye centre localisation evaluations, while the image data of 75 female subjects and an equal number of male subjects (150 faces overall) for gender recognition evaluations (a few image sets were stripped where image frames only contained partial faces due to the subjects standing at incorrect locations while being recorded).

5) The eye centre positions and the gender labels for all the facial images were then manually labelled as the ground truth for algorithm validation.

A group of images captured for one subject is shown in Figure 3.3.

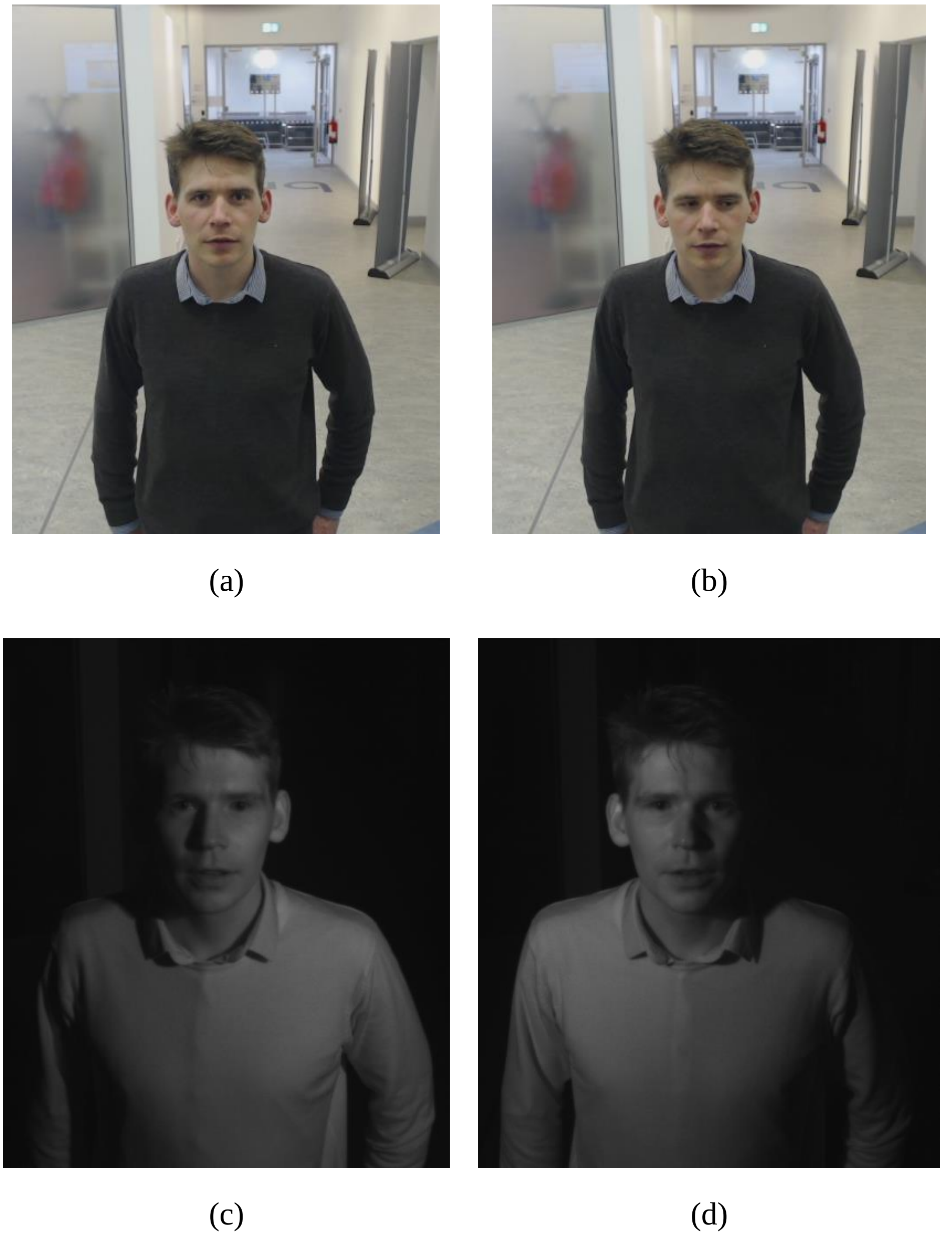


Figure 3.3 A group of representative images for one subject captured during the experiments. Backgrounds in the images are partially removed for aid with visualisation. The colour images were captured by camera 1 while the NIR images were captured by camera 2. (a) A colour image where the subject was looking at dot ‘1’. (b) A colour image where the subject was looking at dot ‘4’. (c) A NIR image where the subject was illuminated by the top-left light source. (d) A NIR image where the subject was illuminated by the top-right light source.

Other than being able to gather data, this system offers a testbed for significant potential in the advancement of HCI. This is enabled by the employment of the two-source PS method that facilitates the design of a system with hardware simplicity as well as desirable real-time performance. Development of this data capture system into two types of HCI applications is introduced in Chapter 8.

Chapter 4 Eye Centre Localisation and Gaze Gesture Recognition

As stated previously, behavioural data can be gathered and interpreted by the visual modality to uncover user attention and to predict user intentions.

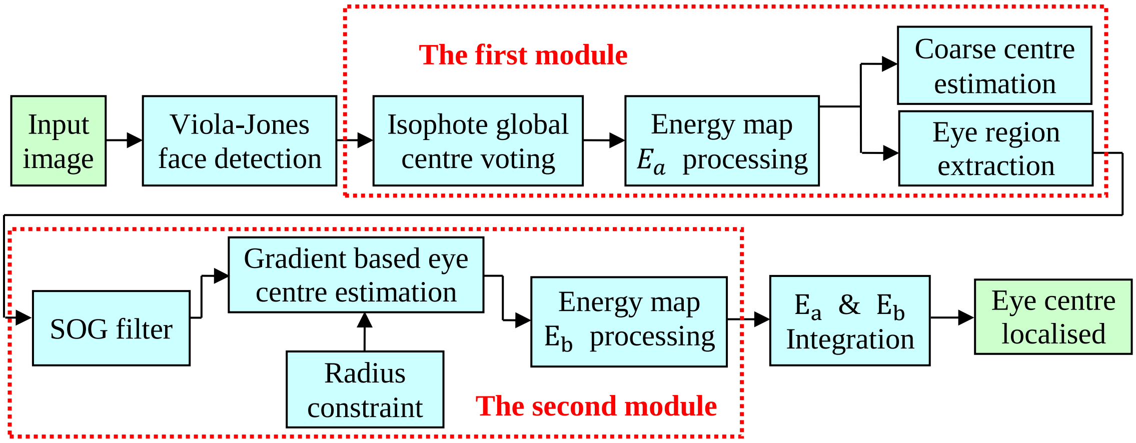
As a complementary respect to demographic data which attempt to answer the question “who is the user”, behaviour analysis responds to the fact that a HCI process is not forever unchanged. On the contrary, a HCI session always appears to be dynamic due to user behaviours that are constantly changing. Demographic data can only address the initial state of the interaction process by predicting a user’s gender, age, ethnicity, or identity, but they cannot adapt to any action initiated by a user. Therefore, probing into behavioural data is the only way to accommodate dynamic HCI sessions.

Eye/gaze analysis reveals the focus of attention of a user over time. Its study offers rich and significant cues to assist with human behaviour analysis in HCI sessions. To this end, this chapter proposes an accurate, reliable and efficient eye centre localisation algorithm and a gaze gesture recognition algorithm. The former can be employed by a HCI system as a passive means to monitor user attentiveness, while the latter can be activated by a user as an active means to input commands into a HCI system and trigger certain HCI events.

In Chapter 8, the eye centre localisation algorithm and the gaze gesture recognition algorithm are utilised to enable two case study HCI systems – a gaze gesture based map browser and an intelligent directed advertising billboard. Therefore, the accuracy, efficiency and robustness of the proposed algorithms can be further assessed in real-world scenarios.

4.1 Eye Centre Localisation – an Unsupervised Modular Approach

This section introduces a hybrid method that can perform accurate and efficient localisation of eye centres in low-resolution images in real time. The algorithm is summarised in Figure 4.1 as an overview of the eye centre localisation chain.



The first module
Input image
Viola-Jones face detection
Isophote global centre voting
Energy map E_a processing
Coarse centre estimation
Eye region extraction
SOG filter
Gradient based eye centre estimation
Radius constraint
Energy map E_b processing
The second module
E_a & E_b Integration
Eye centre localised

Figure 4.1 An overview of the eye centre localisation algorithm chain

The algorithm includes two modalities. The first module performs a global estimation of the eye centres over a face image and extracts the corresponding eye regions. Results from the first module are fed into the second module as prior knowledge and lead to a regional and more precise estimation of the eye centres. The two energy maps generated by the two modalities are fused to produce the final estimation of the eye centres.

4.1.1 Isophote-based Global Centre Voting and Eye Detection

Human eyes can be characterised as radially symmetric patterns which can be represented by contours of equal intensity values in an image, i.e. isophotes (Lichtenauer, Hendriks and Reinders, 2005). Due to the large contrast between the iris and the sclera as well as that between the iris and the pupil, the isophotes that follow the edges of the iris and the pupil reflect the geometric properties of the eye (see Figure 4.2). Therefore the centres of these isophotes will be able to represent the estimated eye centres.

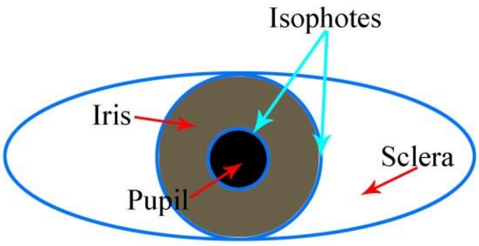


Figure 4.2 The structure of an eye and its isophotes

Valenti and Gevers (2008) designed an isophote-based algorithm for eye centre localisation. It enables pixels in an eye region to vote for the isophote centres they belong to. They calculated the displacement vector pointing from a pixel to its isophote centre following equation (4.1):

$$\{D_x, D_y\} = -\frac{\{I_x, I_y\}\left(I_x^2 + I_y^2\right)}{I_y^2 I_{xx} - 2I_x I_{xy} I_y + I_x^2 I_{yy}} \tag{4.1}$$

where I_x and I_y are first-order derivatives of the luminance function $I(x, y)$ in the x and y directions. I_{xx}, I_{xy} and I_{yy} are the second-order partial derivatives of the luminance function in the x and y directions. The importance of each vote is indicated by the curvedness of the isophote since the iris and pupil edges that are circular obtain high curvedness values as opposed to flat isophotes. Koenderink and Doorn (1992) calculated the curvedness following:

$$cd(x,y) = \sqrt{I_{xx}^2 + 2 \times I_{xy}^2 + I_{yy}^2} \tag{4.2}$$

The brightness of the isophote centres is also considered in the voting process based on the fact that the pupil is normally darker than the iris and the sclera. Therefore, an energy map $\boldsymbol{E_a}(x,y)$ is constructed that collects all the votes to reflect the eye centre position following equation (4.3):

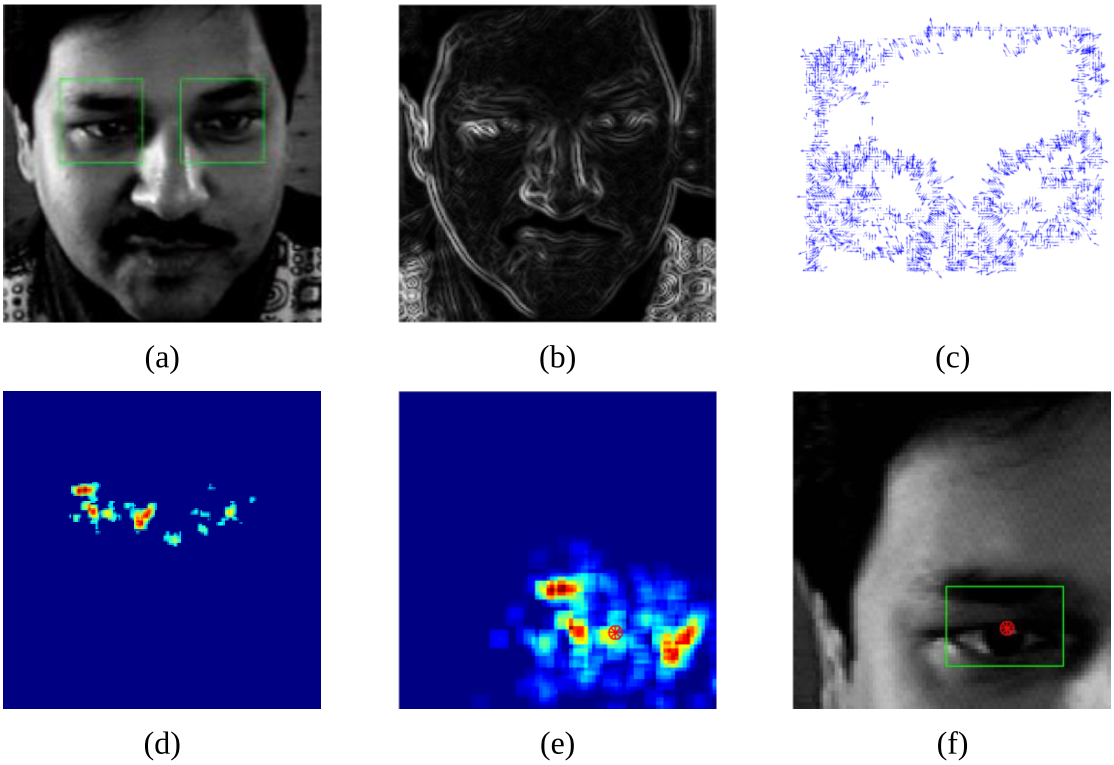
$$\boldsymbol{E_a}(x + D_x, y + D_y) = [\alpha - I(x + D_x, y + D_y)] \times cd(x,y) \tag{4.3}$$

where α is the maximum greyscale in the image ($\alpha = 255$ in the experiments). The α term works to assign a higher score to a darker pixel, while the $cd(x,y)$ term assigns a higher score to a pixel with higher curvedness. Though isophotes have been employed by a number of methods, they have only been extracted from the eye regions which are either cropped according to anthropometric relations (which are interrupted by head rotations) or found by an eye detector (which largely increases the complexity of the algorithm). The proposed method is different, in that it extracts isophote features for the *whole face* and constructs a global energy map $\boldsymbol{E_a}(x,y)$. Energy points that are below 30% of the maximum value are removed. The remaining energy points therefore become the new eye centre candidates that are fed to the second module for further analysis. $\boldsymbol{E_a}(x,y)$ is then split into the left half $\boldsymbol{E_{aul}}(x,y)$ and the right half $\boldsymbol{E_{aur}}(x,y)$, corresponding to the left and right half of the face, where the mouth region (the lower half of the energy map) is simply removed since it is unlikely to concern any eye region information regardless of normal head rotations. The energy centre, i.e. the first moment divided by the total energy, is further calculated, which is selected instead as the optimal eye centres. Taking $\boldsymbol{E_{aul}}(x,y)$ as an example, this can be formulated as equation (4.4):

$$\{\mathrm{cx}_{aul}, \mathrm{cy}_{aul}\} = \frac{\sum_{x=1}^{m} \sum_{y=1}^{n} \{x,y\} \cdot \boldsymbol{E_{aul}}(x,y)}{\sum_{x=1}^{m} \sum_{y=1}^{n} \boldsymbol{E_{aul}}(x,y)} \tag{4.4}$$

where $c_{aul} = \{\mathrm{cx}_{aul}, \mathrm{cy}_{aul}\}$ is the optimal estimation of the left eye centre, m and n

are the maximum row and column number in E_{aul}. The eye region to be analysed by the second module is then selected which centres at the optimal eye centre estimation (its width being 1/10 of the face size and its height being 1/15 of the face size). As a result, the proposed method does not require an eye detector and is robust to head rotations since global isophotes are investigated. This process is shown in Figure 4.3.


(a) (b) (c)

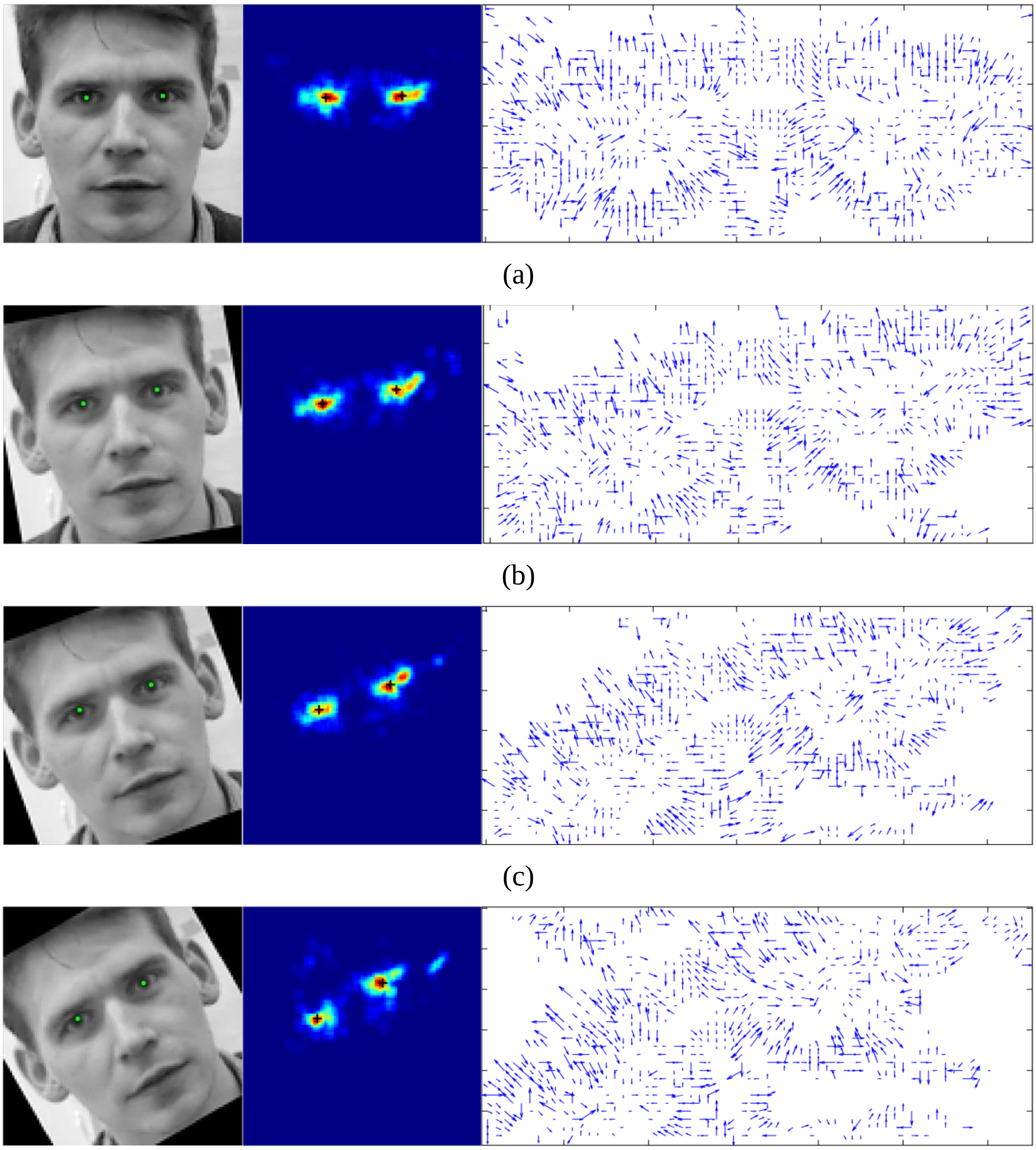
(d) (e) (f)

Figure 4.3 An example of isophote based eye centre localisation (a) A face image from the BioID database (BioID Technology Research, n.d.). (b) The curvedness image. (c) The displacement vectors for the upper face. (d) The energy map for the upper face. (e) The energy map for the left eye. (f) The eye centre localised and the eye region extracted.

The incorporation of this module brings two major benefits:

1) The global analysis of isophote features from a detected face region rather than local eye regions is robust against severe in-plane head rotations. This further reinforces the rotational invariance of isophote features. An example can be found in Figure 4.4 where excellent rotational invariance is demonstrated on a face image manually rotated by 10 degrees, 20 degrees and 30 degrees. This image was obtained during the data capture

experiment introduced in the preceding chapter. The detected face area is 105 × 105 pixels.



(a)

(b)

(c)

(d)

Figure 4.4 A demonstration of the first module applied to a face image with 10-degree, 20-degree and 30-degree in-plane head rotations. From left to right, each subfigure displays the detected/rotated raw face image, the corresponding energy map, and the displacement vector field for the upper face.

As can be seen from Figure 4.4, the first module of the proposed eye centre localisation method is not bounded by anthropometric measurements of frontal faces. The high-energy clusters in the energy maps can adapt to head rotations without the need for a trained eye classification model.

2) This module is unsupervised and non-iterative, and therefore can efficiently prepare eye centre candidates for further analysis by the second module. This can be reflected statistically as in subsection 4.4.1.

4.1.2 Gradient-based Eye Centre Estimation

The first module performs the initial approximate eye centre estimation, filtering eye centre candidates and selecting local eye regions for the second module. Based on an objective function (equations (4.5) and (4.6)), a radius constraint is further introduced and a Selective Oriented Gradient (SOG) filter is designed to re-estimate the eye centre positions with enhanced accuracy and reliability.

The radially symmetric patterns of the eye generate isophotes around the iris and pupil edges that can effectively vote for the eye centre. When simply modelled as circular objects, the iris and pupil can produce gradient features that give an accurate estimation of the eye centre. This is based on the idea that the prominent gradient vectors on the circular iris/pupil boundary should agree with the radial directions and therefore the dot product of each gradient vector with its corresponding radial vector is maximised. Following the same eye structure as in Figure 4.2, this can be illustrated by Figure 4.5 where the red arrows represent the gradient vectors and the green arrows represent the corresponding radial vectors.

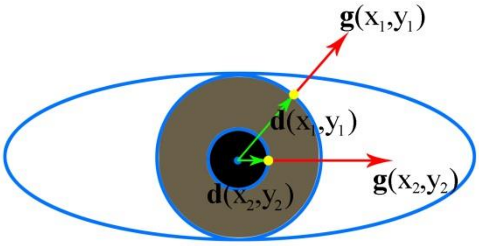


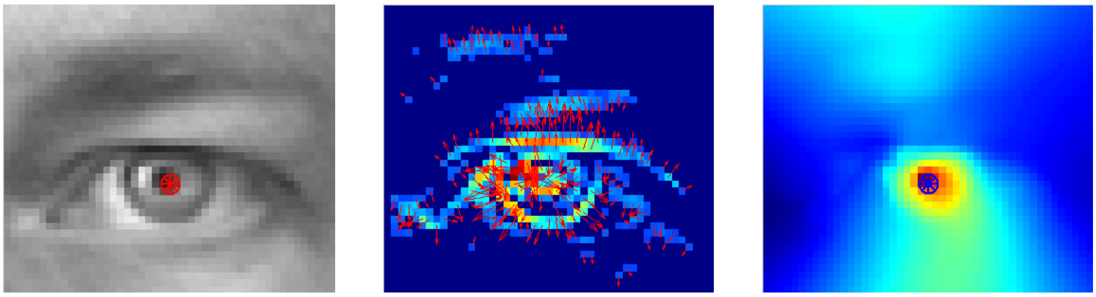
Figure 4.5 The eye model for the gradient-based module

Timm and Barth (2011) formulated this model as an objective function:

$$\boldsymbol{c}^{*} = \underset{\boldsymbol{c}}{\arg max} \left\{ \frac{1}{m \cdot n} \sum_{x=1}^{m} \sum_{y=1}^{n} [\boldsymbol{d}^{\boldsymbol{T}}(x,y) \cdot \boldsymbol{g}(x,y)]^2 \right\} \tag{4.5}$$

$$\boldsymbol{d}(x,y) = \frac{\boldsymbol{p}(x,y) - \boldsymbol{c}}{\|\boldsymbol{p}(x,y) - \boldsymbol{c}\|_2}, \quad \forall x \forall y \colon \|\boldsymbol{d}(x,y)\|_2 = 1, \ \|\boldsymbol{g}(x,y)\|_2 = 1 \tag{4.6}$$

where $\boldsymbol{c}$ is the centre candidates, $\boldsymbol{c}^{*}$ is the optimal centre, N is the number of pixels in the eye region to be analysed, $\boldsymbol{d}(x,y)$ is the displacement vector connecting a centre candidate $\boldsymbol{c}$ and $\boldsymbol{p}(x,y)$ which is any pixel different from $\boldsymbol{c}$, $\boldsymbol{g}(x,y)$ is the gradient vector on the edge point, $I_{\boldsymbol{c}}$ is the intensity value at an isophote centre and $m \, and \, n$ have the same definition as in the preceding subsection. The displacement vectors and gradient vectors are normalised to unit vectors. Similarly to the first module (equation (4.3)), the $[\alpha - I_c(x,y)]$ term is added to this equation in order to assign high scores to dark pupil pixels. This objective function is also modified such that the direction of the gradient vector is only considered if it is reverse to the displacement vector based on the fact that the pupil is always darker than its neighbouring regions and thus generates outward gradients. A sample implementation of this approach on an eye image is illustrated in Figure 4.6.



(a) (b) (c)

Figure 4.6 An example of gradient based eye centre localisation. (a) An eye image. (b) The display of gradient magnitude where the gradient directions are represented by arrows. The gradients with magnitude below 70% of the maximum are removed. (c) The resulting energy map for eye centre candidates

Note that, for the sake of demonstration, the eye region in Figure 4.6 was manually cropped such that it could be demonstrated independently of the first module. However, in all the evaluation experiments, the eye regions were much smaller and were automatically cropped by the first module. This module was also tested on image frames from a video where challenging conditions exemplified facial accessory, head pose, extreme pupil position, half-closed eyes and eye/face occlusion. These results can be found in Figure 4.7.

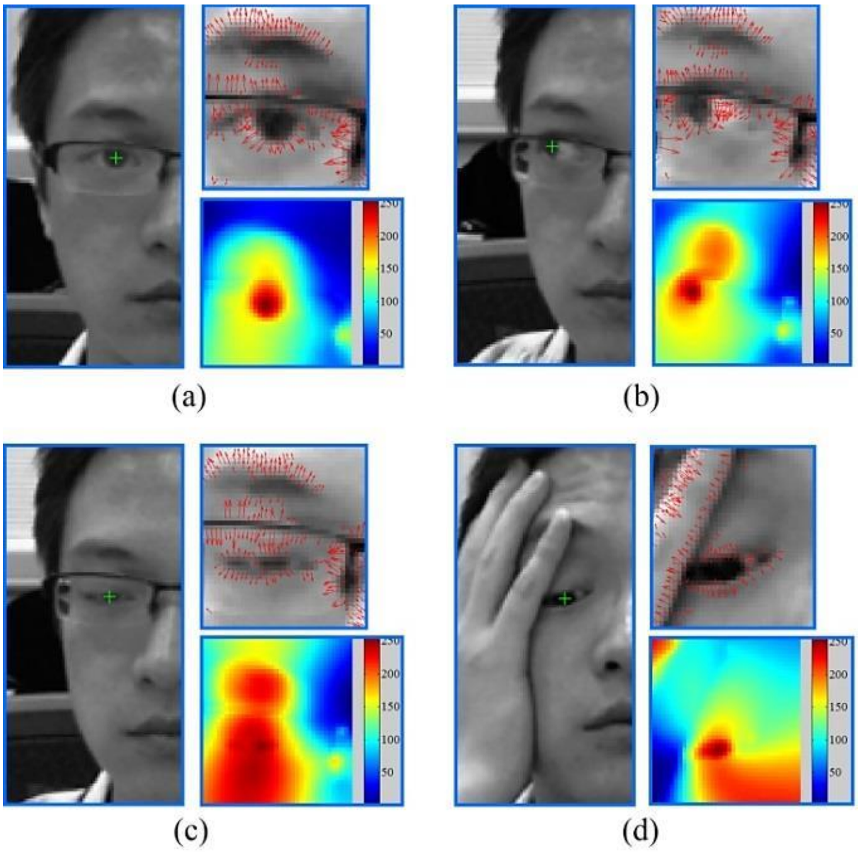


Figure 4.7 Sample results generated by the second module. Challenges include: (a) facial accessory: a pair of black-framed glasses; (b) a rotated face with an extreme pupil position; (c) a half-closed eye; (d) a partially occluded face/eye.

As can be seen from this demonstration, the proposed approach can overcome challenges that can cause the circle Hough transform technique to fail. These challenges, such as deformed circular contours, segmented contours and image noise, are also common obstacles to other geometric feature based approaches, as reviewed in Chapter 2. While this method appears sufficiently robust against deformable pupil/iris patterns, it has a number of inherent limitations that would cause error or even failure in the estimation. First of all, the objective function is formulated given that the pupil and iris are circular objects. When the edges around eye corners and shadows exceed the pupil and the iris in the circularity measure, they will cause the eye centre estimates to be located on themselves rather than the centre of the iris/pupil. Secondly, the gradients on the eyelid, eye corners and eyebrows will also participate in the eye centre estimation. These are defined and referred to as ‘non-effective gradients’ in the rest of the thesis, as opposed to ‘effective gradients’ which are located on the edges of the iris and pupil. In more severe cases where makeup and shadows are prominent, the iris and the pupil, by contrast, generate weak ‘effective gradients’ such that the energy map is prone to erroneous energy response. This will further escalate the error of the estimation.

To resolve the above problems shared by most methods that utilise geometric features for eye centre localisation, a radius constraint and a SOG filter are designed that effectively deal with the circularity measure and problems posed by eyebrows and eyelids.

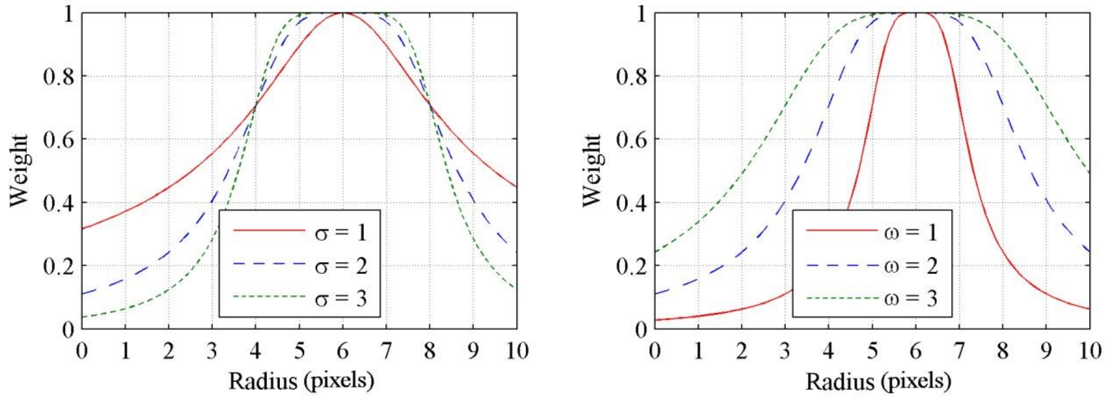
4.1.3 Iris Radius Constraint

A radius constraint is introduced such that the Euclidean norms of the displacement vectors, which are related to the estimated iris radius, have more influence on the calculation of eye centre localisation. This is based on the assumption that the shadows and the eyebrow segments have random radius values, while the iris radii are more constant relative to the size of a face.

This provides a way to differentiate circular clusters of various radii and to determine their weights in the energy map accumulation. The function for the significance measure emulates the frequency response of a Butterworth low pass filter:

$$rw(x,y) = \sqrt{\frac{1}{1+\left(\frac{\|\boldsymbol{d}(x,y)\|_2 - \widehat{D}}{\omega}\right)^{2\sigma}}} \tag{4.7}$$

where $\|\boldsymbol{d}(x,y)\|_2$ is the ℓ_2-norm of the displacement vector without being normalised to a unit vector. $\widehat{D}$ is the estimated radius of the iris, whose value can be set according to the size of the face in an image. σ and ω correspond to the order and the cut-off frequency of the filter. Such a frequency response was chosen instead of a Gaussian function since it would provide more freedom in controlling the flatness band and the roll-off rate of the curve. This means that the iris radius constraint can be more accurately tuned. The curves corresponding to varying σ and ω following equation (4.7) are shown in Figure 4.8. It should be noted that in each subfigure only one parameter is variable while the other remains constant.



(a) (b)

Figure 4.8 Iris radius constraint with different parameters. (a) Curves with varying $\boldsymbol{\sigma}$ ($\boldsymbol{\sigma = 1, 2, 3}$) and constant $\boldsymbol{\omega}$ ($\boldsymbol{\omega = 2}$). (b) Curves with varying $\boldsymbol{\omega}$ ($\boldsymbol{\omega = 1, 2, 3}$) and constant $\boldsymbol{\sigma}$ ($\boldsymbol{\sigma = 2}$).

The radius weight function is maximally flat around the estimated centre $\widehat{D}$ and drops rapidly when the radius is out of the flatness band whose range is controlled by σ. The roll-off rate is controlled by ω, indicating the decreasing rate in weight. Increasing ω while decreasing σ will enhance the rigidity of the constraint which could be assumed for circumstances where strong shadows are present.

As a result, the new objective function becomes:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\arg max} \left\{ \frac{1}{m \cdot n} \sum_{x=1}^{m} \sum_{y=1}^{n} rw(x,y) \cdot [\alpha - I_c(x,y)] \cdot [\boldsymbol{d^T}(x,y) \cdot \boldsymbol{g}(x,y)]^2 \right\} \quad (4.8)$$

This objective function allows adjustable coefficients that effectively alleviate the problematic issues posed by edges around eyelids, eye corners and shadows.

4.1.4 Selective Oriented Gradient Filter

With the inspiration drawn from the histogram of oriented gradients (HOG) feature descriptor (Dalal and Triggs, 2005), A Selective Oriented Gradient (SOG) filter is introduced that discriminates gradients of rapid change in orientation from those of less change. This novel SOG filter is specifically design and introduced into the modular eye centre localisation scheme so that it is perfectly tailored to reinforce the two main modalities despite its versatile applicability.

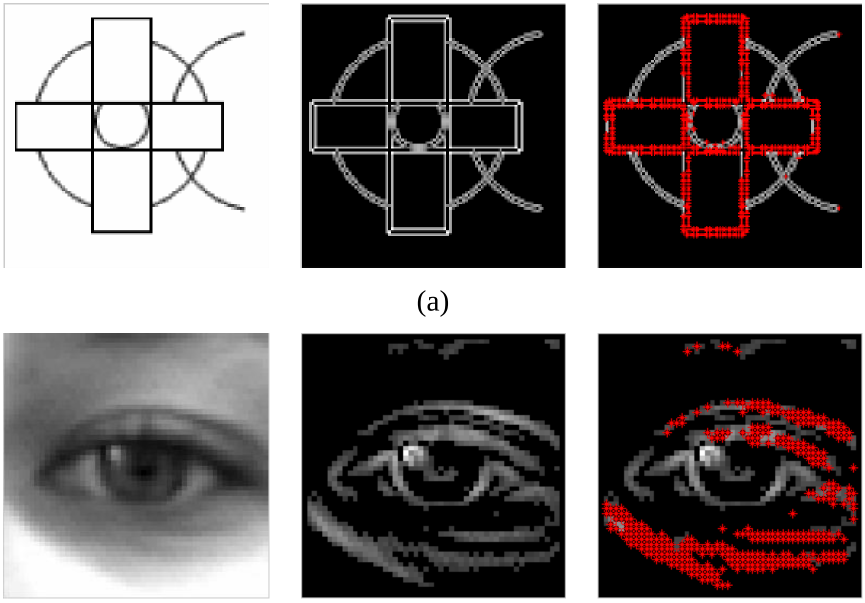
The basic idea takes the form of a statistical analysis of gradient orientations within a window centred at a pixel position. For each $S_x \times S_y$ window centred at a particular pixel location, the gradients in x and y directions are calculated whose orientations follow:

$$o_g = \tan^{-1}\left(\frac{I_y}{I_x}\right)\cdot\frac{180^\circ}{\pi} \tag{4.9}$$

The gradient orientations are then accumulated into $k\ (k < 360)$ orientation bins, where each bin contains the count of the orientations from $\mathcal{s}\cdot\frac{360^\circ}{k}$ to $(\mathcal{s}+1)\cdot\frac{360^\circ}{k}$ $(0 \leq \mathcal{s} \leq k-1)$ within the window. If the count recorded in a bin exceeds a threshold, meaning that there exist in the window many pixels of a particular gradient orientation, these pixels that accumulate the bin are very unlikely to be located on a circular contour. Therefore, these pixels should have their gradient vector halved, i.e. their weights reduced, so that their votes in the objective function will turn into low energy points in the energy map. As a result, the objective function becomes:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\arg max}\left\{\frac{1}{m\cdot n}\sum\nolimits_{x=1}^{m}\sum\nolimits_{y=1}^{n} sw(x,y)\cdot rw(x,y)\cdot[\alpha - I(x,y)]\cdot[\boldsymbol{d}^{\boldsymbol{T}}(x,y)\cdot\boldsymbol{g}(x,y)]^2\right\} \tag{4.10}$$

where $sw(x,y)$ is the weight of a gradient adjusted by the SOG filter. The threshold for the counts is determined by an absolute value (6% of the total pixel number within the window) as well as a value (40%) relative to the number of pixels with non-zero gradients within the window. As a result, the pixels that maintain similar gradient orientations to their neighbours will have their weights reduced and they are referred to as ‘impaired pixels’ in the rest of the thesis. When a curve has low curvature, it comprises more ‘impaired pixels’. Therefore the SOG filter can be used for general curvature discrimination tasks. It has the advantage that it does not require an explicit function for the curve and also it is effective in dealing with curves that form irregular shapes. It should be noted that the SOG filter will not attempt to detect all ‘impaired pixels’ in an eye region. It is only designed to further lower the votes to false eye candidates as a complement to the second module. In fact, even a small number of detected ‘impaired pixels’ can sufficiently increase the difference between votes to true candidates and false candidates. This also means that the SOG filter is not sensitive to the threshold defined above, since the threshold only affects the number of detected ‘impaired pixels’, but will not cause the SOG filter to be erroneous. Figure 4.9 demonstrates the effectiveness of a SOG filter applied to an image containing irregular curves and an image of an eye region.

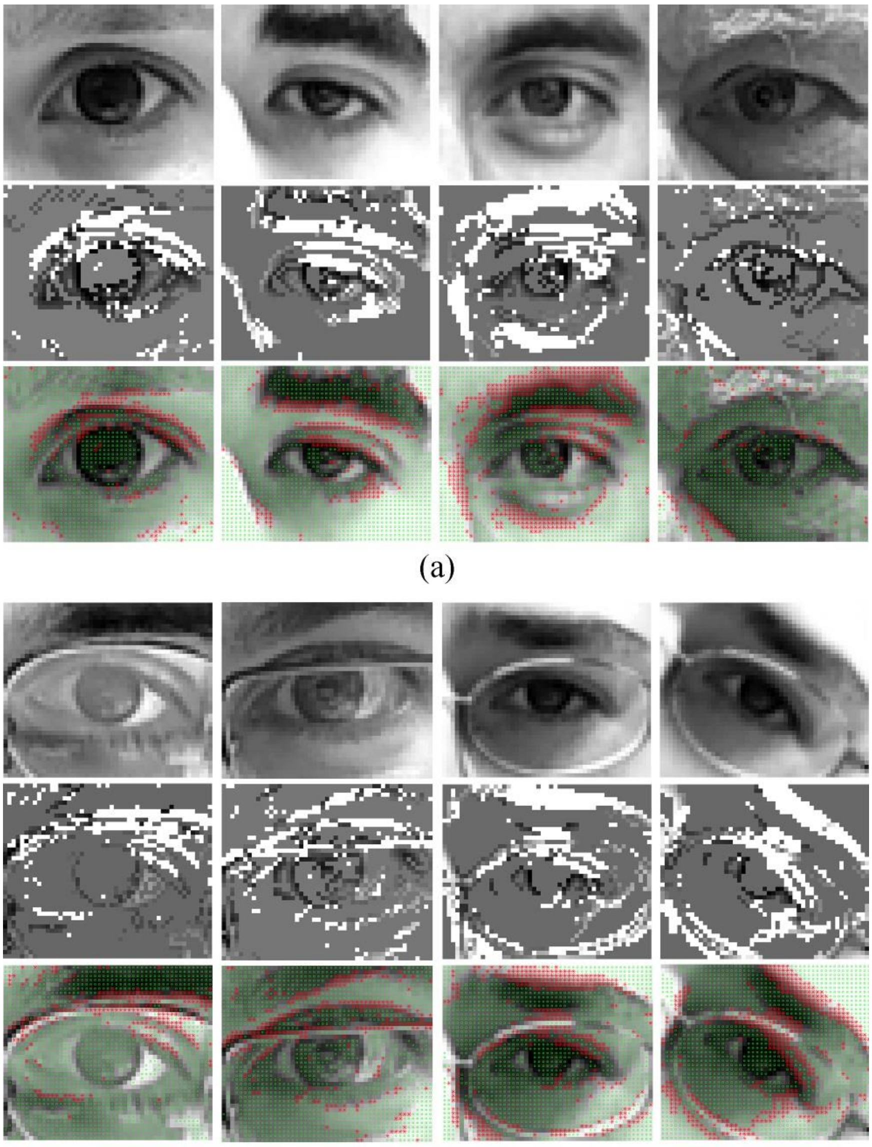


(a)

(b)

Figure 4.9 Gradient filtering using a SOG filter. (a) An example of curved shape detection using a SOG filter where the ‘impaired pixels’ are detected and removed. (b) An example where a SOG filter is applied to an eye image. The magnitudes of gradients are computed where the edges on the eye pouch and eye lid are successfully detected and removed.

It is shown in Figure 4.9 that the SOG filter has successfully distinguished curves with low and high curvatures and that it is effective in dealing with intersected and occluded curves or curve segments. In the eye image, the gradients in the eyelid and shadowed eye pouches are detected as ‘impaired pixels’ whose weights are to be reduced in the accumulation of the energy map while the gradients around the iris and the pupil maintain their original weights. To demonstrate its effectiveness, a SOG filter has also been applied to the GI4E database (Villanueva et al., 2013). Results from a few representative eye images are shown in Figure 4.10.



(a)

(b)

Figure 4.10 Representative examples of a SOG filter detecting interfering gradients. The top row in each subfigure displays eye regions; The middle row displays interfering gradients detected by the SOG filter, displayed in greyscale, i.e. those with weak magnitude (e.g. on the skin area) are marked in 128-grey and those with less orientation variation (e.g. on the eyelid) are marked in white; The bottom row displays interfering gradients detected by the SOG filter, displayed in colour, i.e. those with weak magnitude (e.g. on the skin area) are marked in green and those with less orientation variation (e.g. on the eyelid) are marked in red. (a) Results on eye images without glasses. (b) Results on eye images with different glass frames.

This approach allows the magnitude and orientation of gradients to serve independently in obtaining representative features. It resolves the challenges brought by shadows, facial makeup and edges on the eyelids, eyebrows and other facial parts outside the iris that are most interfering in geometric feature based eye centre localisation approaches.

4.1.5 Energy Map Integration

In the final stage, the two energy maps $\boldsymbol{E_a}(x,y)$ and $\boldsymbol{E_b}(x,y)$ from the first and the second module are integrated into $\boldsymbol{E_f}(x,y)$ so that they both contribute to the election of the eye centre. It is critical, prior to the integration, to determine the confidence of each module, to estimate the complexity of the eye image, and thus to determine their weights in the fusion mechanism.

The left eye region is taken as an example to illustrate the fusion mechanism. If the equivalent centroid c_{aul} calculated by equation (4.4) is close to the pixel position c_{maxl} that has the maximum value in the first energy map $\boldsymbol{E_{aul}}(x,y)$, c_{aul} is considered confident since the isophote centre and the equivalent centroid coincide. In this case, more 'effective gradients' are present, allowing the second module to be more robust and precise. The two modalities are then utilised and fused following equation (4.11). When c_{aul} and c_{maxl} disagree and have a large Euclidean distance, the first energy map will have high energy clusters sparsely distributed, potentially caused by severe shadows and specularities. The second module will be influenced by 'impaired pixels' and produce erroneous centre estimates. Therefore only the equivalent centroids c_{aul} and c_{aur} from the first module are selected to be the final eye centres.

$$\boldsymbol{E_f}(x,y) = \frac{1}{\|c_{aul} - c_{maxl}\|_2} \cdot \boldsymbol{E_a}(x,y) + \boldsymbol{E_b}(x,y) \tag{4.11}$$

where ϵ takes a value relative to the width of the eye region ϵ_f and $0 < \|c_{aul} - c_{maxl}\|_2 \leq \epsilon$. In the experiments, $\epsilon = 0.3\epsilon_f$ pixels. The maximum response in the final energy map will represent the estimated eye centre. The estimate for the final right eye centre follows the same procedure.

4.2 Eye Centre Localisation Experiments and Results

Three publicly available databases were tested in the experiments: the BioID database (BioID Technology Research, n.d.), the GI4E database (Villanueva et al., 2013) and the extended Yale Face Database b (Georghiades, Belhumeur and Kriegman, 2001). The BioID database consists of 1520 images. It has been popular in the literature for the evaluation of other eye centre localisation algorithms. The variations in the database include illumination, face scale, head pose and the presence of glasses. The GI4E database is known for containing images of 103 subjects with 12 different gaze directions. Applied to these two databases, the proposed algorithm can be tested against the others in the literature in resolving challenges introduced by the wide range of variations. The extended Yale Face Database b is captured under extremely challenging lighting conditions and also contains various head poses. Since the proposed method is geometric feature/model based and is unsupervised, it does not attempt to deal with images where pupil/iris regions are invisible. Therefore, a subset of the extended Yale Face Database b is selected where the absolute azimuth and elevation angles are no larger than 40 degrees such that the periocular regions are not completely shadowed.

The relative error measure proposed by Jesorsky, Kirchberg and Frischholz (2001) is used to evaluate the accuracy of the proposed algorithm. It firstly calculates the absolute error which is the Euclidian distance between the centre estimates and the ground truth provided by the database and then normalises the Euclidian distance relative to the pupillary distance. This is formulated by equation (4.12):

$$e = \frac{max\left(d_{left}, d_{right}\right)}{P_d} \tag{4.12}$$

where d_{left} and d_{right} are the absolute errors for the eye pair, and P_d is the pupillary distance. The maximum of d_{left} and d_{right} after normalisation is defined as 'max normalised error' e_{max}. Additionally, the accuracy curve for the minimum normalised error e_{min} and the average normalised error e_{avg} are calculated. A relative distance of

$e = 0.25$ corresponds to half the width of an eye. The evaluations on the BioID database are shown in Figure 4.11.

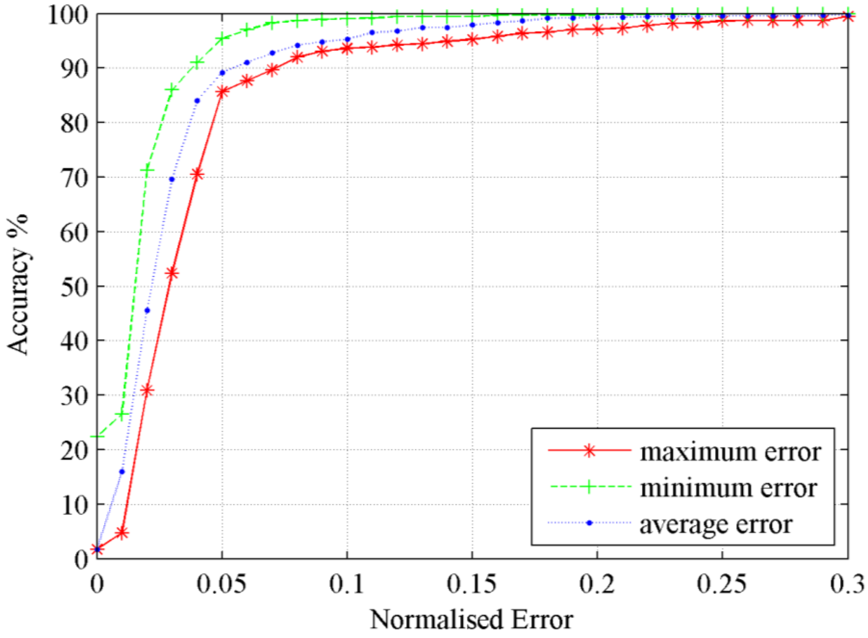


Figure 4.11 Accuracy curve of the proposed method on the BioID database

The proposed algorithm is further compared with ten state-of-the-art methods in the literature, summarised in Table 4.1.

Table 4.1 A comparison of the proposed eye centre localisation method with ten state-of-the-art methods in the literature

Method	Accuracy under minimum and maximum normalised error						Score
	$e_{max} \le 0.05$	$e_{min} \le 0.05$	$e_{max} \le 0.10$	$e_{min} \le 0.10$	$e_{max} \le 0.25$	$e_{min} \le 0.25$	
the proposed method	85.66%	95.46%	93.68%	**99.06%**	**99.21%**	99.93%	**6**
(Leo et al., 2014)	80.67%	\	87.31%	\	93.86%*	\	**0**
(Valenti and Gevers, 2012)	**86.09%**	96.07%	91.67%	97.87%	97.87%	100%*	**3**
(Timm and Barth, 2011)	82.50%	93.50%*	93.40%	98.50%*	98.00%	100%*	**1**
(Asadifard and Shanbezadeh, 2010)	47.00%	\	86.00%	\	96.00%	\	**0**
(Kroon, Hanjalic and Maas, 2008)	65.00%	\	87.00%	\	98.80%	\	**1**
(Valenti and Gevers, 2008)	84.10%	**96.28%**	90.85%	97.94%	98.49%	100%*	**2**
(Campadelli, Lanzarotti and Lipori, 2006)	62.00%	\	85.20%	\	96.10%	\	**0**
(Niu et al., 2006)	75.10%*	\	93.00%	\	96.30%*	\	**0**
(Hamouz et al., 2005)	58.00%*	\	76.00%*	\	90.80%*	\	**0**
(Cristinacce, Cootes and Scott, 2004)	57.00%*	\	**96.00%***	\	97.10%*	\	**2**

The proposed method gains the best results for the accuracy measure $e_{min} \le 0.10$ as well as $e_{max} \le 0.25$, and the second best for $e_{max} \le 0.05$ and $e_{max} \le 0.10$. Except for the accuracy measure for $e_{min} \leq 0.25$ where very similar results are achieved, a score of 2 is assigned to every first rank and a score of 1 is assigned to every second rank. The proposed method gains a total score of 6, outperforming all the other methods when comparing the classification accuracy.

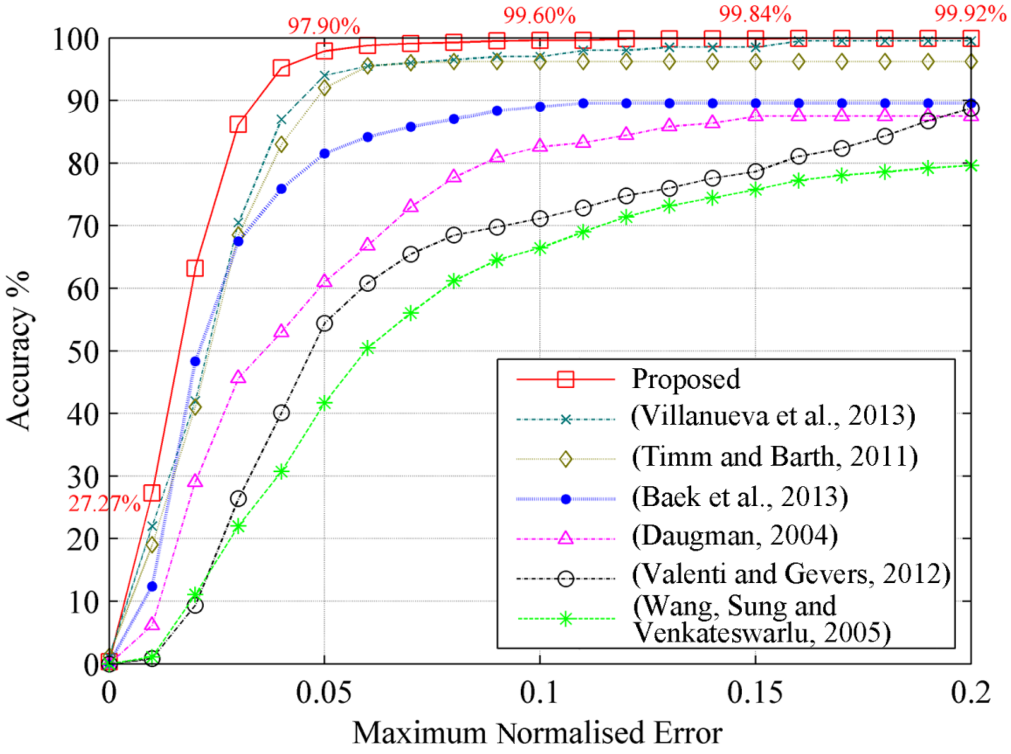
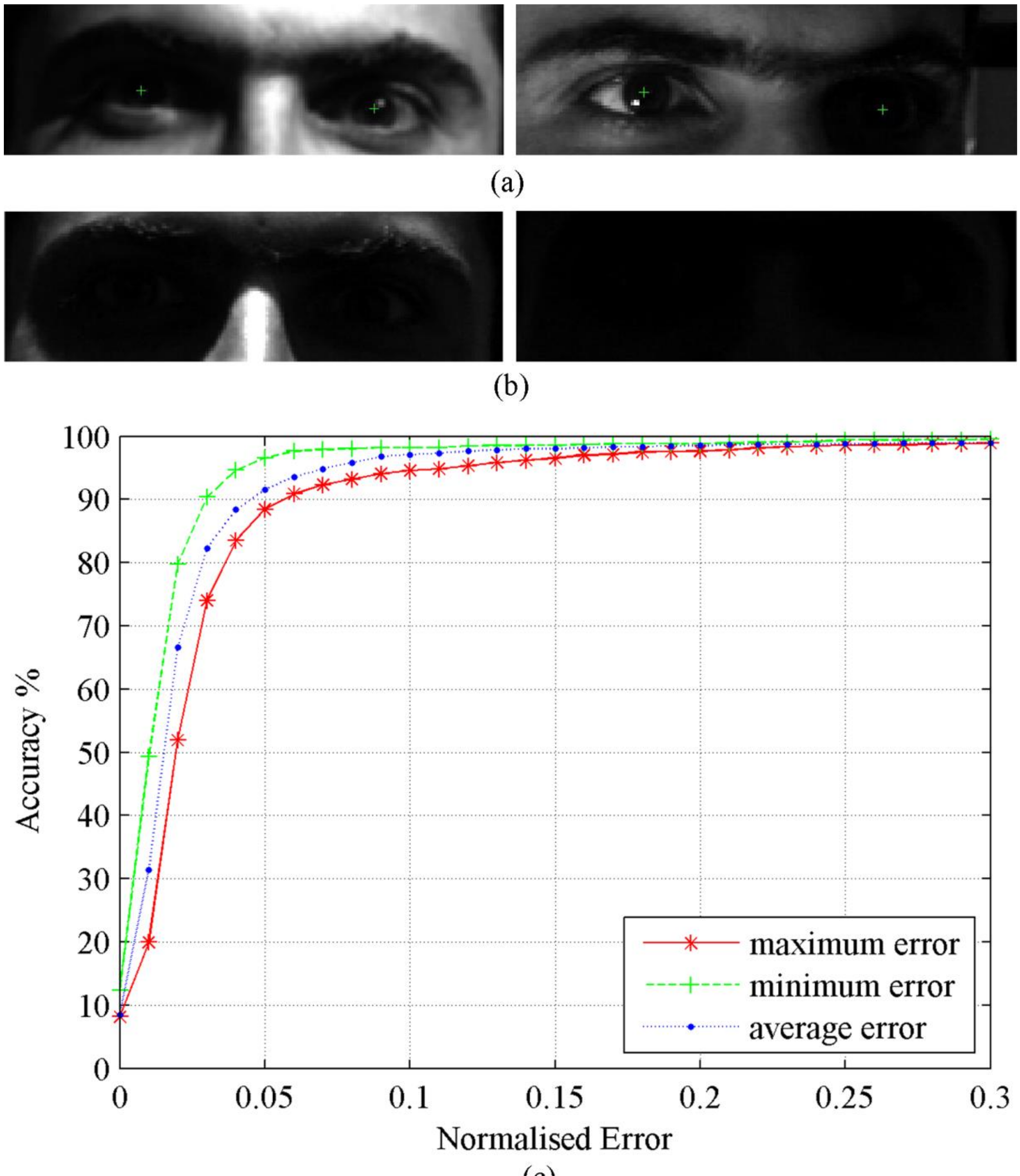


Figure 4.12 Accuracy curves of the proposed method on the GI4E database, in comparison with six other methods.

Further evaluations on the GI4E database are compared to six other methods, as shown in Figure 4.12. Outperforming all the other methods in comparison, the proposed method proves to be robust against eye movement by achieving 97.9% accuracy for $e_{max} \leq 0.05$.

Tests on a subset of the extended Yale Face Database b also reflect that the proposed method can maintain high accuracy on eight different head poses and challenging lighting conditions as long as pupil/iris regions are not entirely shadowed. These results can be found in Figure 4.13.



(c)

Figure 4.13 Evaluation result on a subset of the extended Yale Face Database b. (a) Representative eye regions in the subset with challenging illumination (when the azimuth and elevation angles are no larger than 40 degrees). (b) Representative eye regions not included in the subset (when the azimuth and elevation angles are larger than 40 degrees). (c) Accuracy curves for different error measures.

The high accuracy of the proposed method has been demonstrated through experiments on the three publicly available databases. In the preceding section, the robustness of the proposed method has also been clarified through the theoretical illustration of the two complementary modalities, the iris radius constraint and the SOG filter. To further support the previous theoretical illustration from an applied point of view, a second set of experiments was conducted that focused on resolving low resolution data, head pose variation and poor illumination condition. This set of experiments proves that the proposed algorithm is capable of maintaining high accuracies for data captured out of the laboratory environments, and that it exhibits high robustness.

The facial data employed by this set of experiments are collected by the 2D+3D imaging system introduced in the preceding chapter. Therefore, the data can be deemed very challenging due to the real-world settings and the diverse variations.

The data are divided into 3 groups for algorithm evaluations on 1) frontal faces with centred pupil positions (‘Group A’ data), 2) faces with head rotations and/or eye movements (‘Group B’ data), and 3) faces illuminated by insufficient and uneven lighting (‘Group C’ data).

One representative image for each group is shown in Figure 4.14.

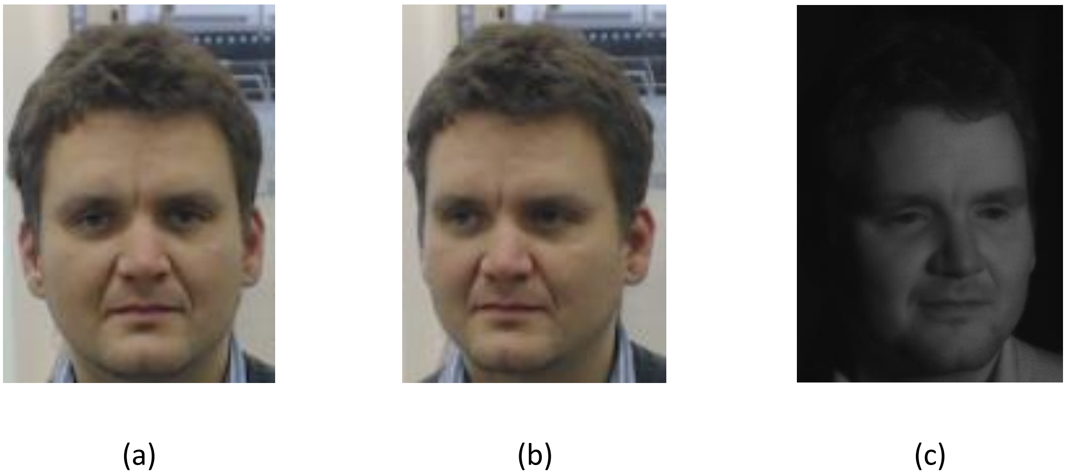


Figure 4.14 A group of representative facial images gathered by the data capture experiments introduced in Section 3.2. Only the face regions are shown in the three sample images. (a) A sample image from 'Group A' data where all faces are frontal. (b) A sample image from 'Group B' data where head pose variation and/or eye movement may be present. (c) A sample image from 'Group C' data, captured under uneven and insufficient illumination condition.

The proposed eye centre localisation algorithm was tested on all three groups of data. Figure 4.15 shows a number of examples of accurately and inaccurately localised eye centres in this set of experiments.

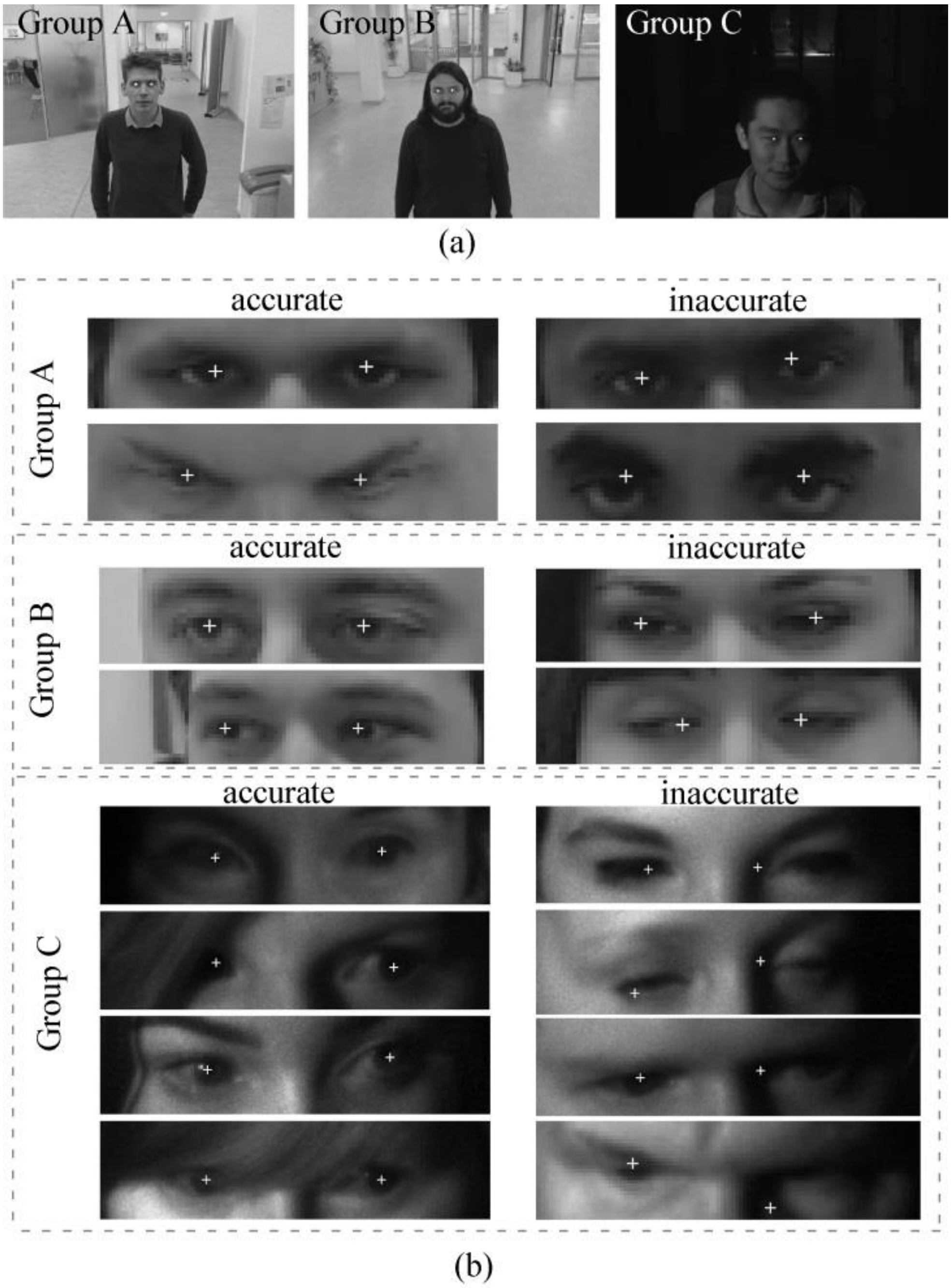
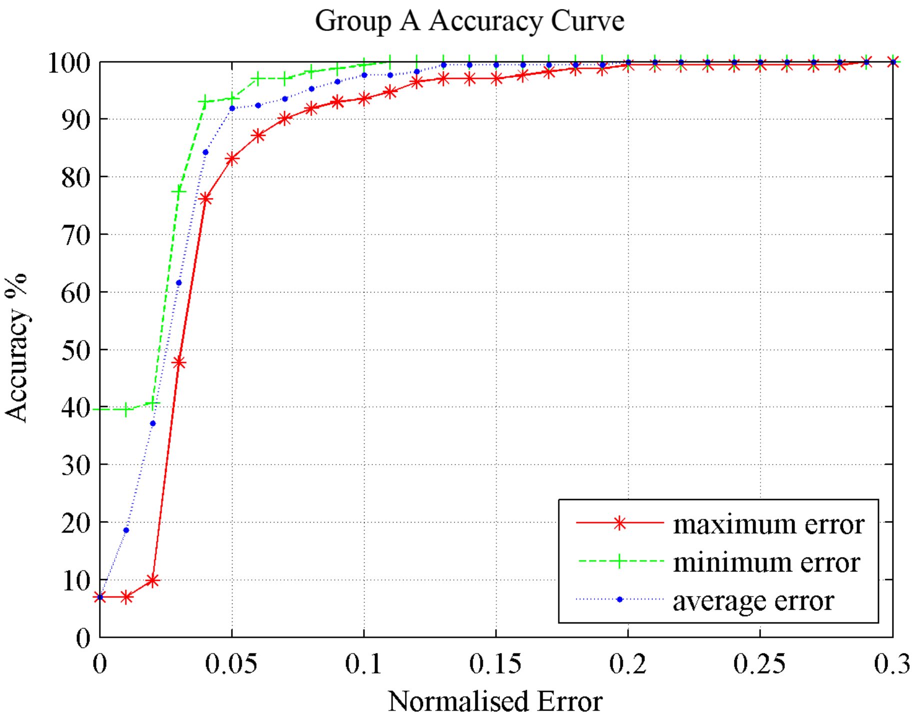
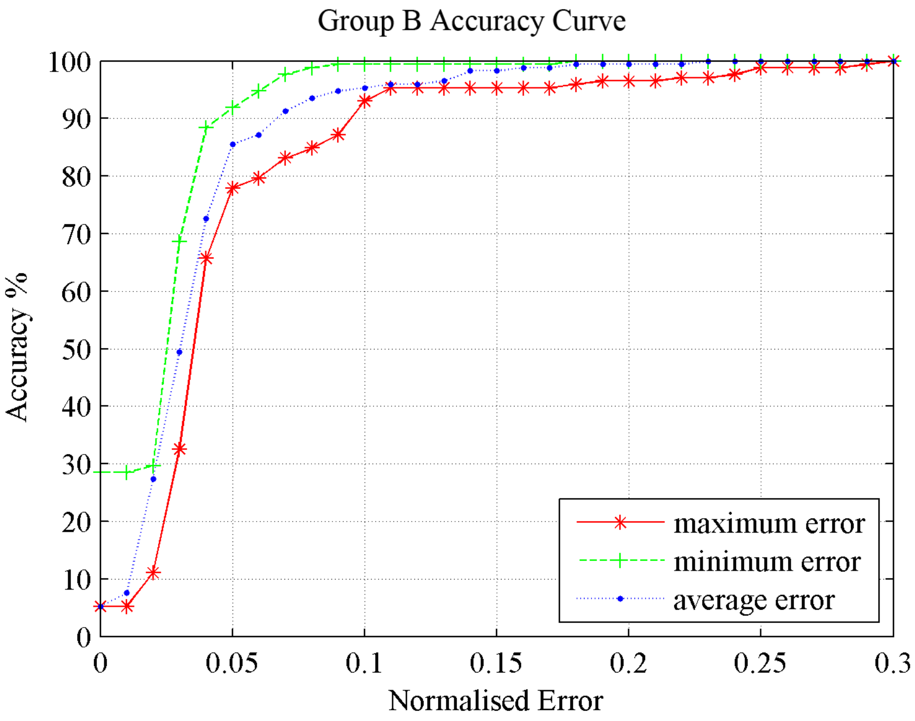


Figure 4.15 Representative examples of the self-collected database and eye centre localisation results. (a) Sample images from Group A, B and C self-collected data. (b) Representative examples of accurately and inaccurately localised eye centres.

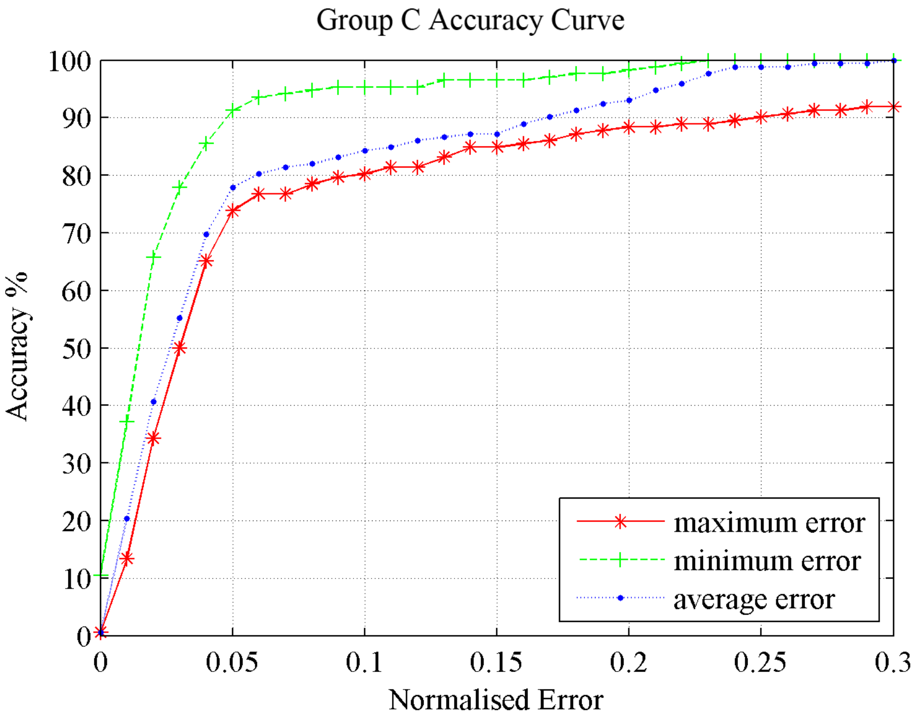
It can be seen from Figure 4.15 that the proposed algorithm is able to function at relatively long camera-subject distance (e.g. a one-metre distance with 640×480 image size, or equivalently a 2-metre distance with 1280×960 image size), and to deal with images captured under poor illumination conditions where severe shadows and specularities are present. Only when the iris and pupil regions are completely shadowed or occluded will they cause false identifications of eye centres. In Figure 4.16, the eye centre localisation results in the form of accuracy curves are further summarised for the three groups of the 172 volunteers, i.e. a total number of 516 images.



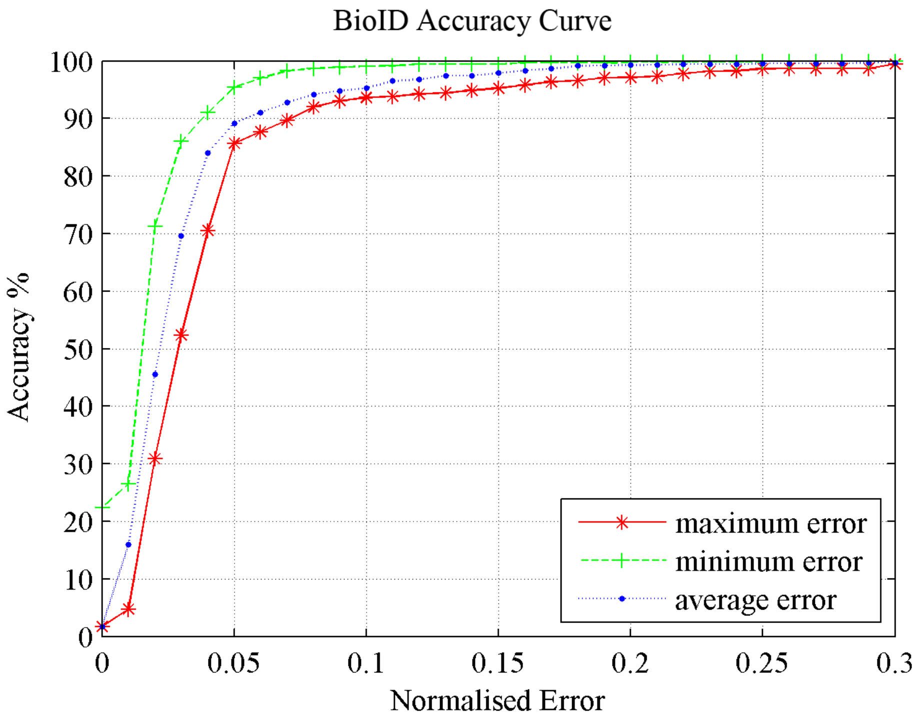
(a)


Group B Accuracy Curve
Accuracy %
Normalised Error
0
10
20
30
40
50
60
70
80
90
100
0
0.05
0.1
0.15
0.2
0.25
0.3
maximum error
minimum error
average error

(b)


Group C Accuracy Curve
Accuracy %
Normalised Error
0
10
20
30
40
50
60
70
80
90
100
0
0.05
0.1
0.15
0.2
0.25
0.3
maximum error
minimum error
average error

(c)



(d)

Figure 4.16 Accuracy curves for (a) 'Group A' data, (b) 'Group B' data, and (c) 'Group C' data. (d) Accuracy curve for the BioID database as a reference.

Figure 4.16 (a) shows that the proposed algorithm maintains high accuracy with 'Group A' data, i.e. frontal face images capture in real-world environments where main challenges are posed by self-cast shadows. Similar results can be seen in Figure 4.16 (b) where further complications include natural head poses and eye movements that cause deformation and occlusion of eye regions. All the curves representing the maximum error, the minimum error and the average error under realistic scenes are comparable to those generated from the BioID database (Figure 4.16 (d)). Figure 4.16 (c) corresponds to the experiment where the NIR illuminator deliberately created poorly illuminated images and severe self-cast shadows, resulting in at least one pupil and/or iris in every image being obscure or even invisible. This accounts for the declined accuracy shown by the curve of the maximum error. Nevertheless, the robustness of the proposed method is validated by the curve of the minimum error, which indicates that even under poor illumination conditions, at least one eye centre in every image is localised with extremely high accuracy (above 90% accuracy with normalised error of 0.05, and 100% accuracy with normalised error of 0.25).

As the proposed eye centre localisation algorithm exhibits high accuracy and robustness under real-world scenarios, a gaze gesture recognition algorithm is further designed to utilise the methodology to boost the HCI experience.

4.3 Gaze Gesture Recognition

Gaze gestures are predefined sequences of eye movements which hold great potential in HCI. They record both intentional and unintentional eye saccades that can reflect user attentiveness and intentions, respectively. In this research, the focus is placed on intentional eye saccades due to their potential to enhance the interactivity of HCI systems. The value of unintentional eye saccades are also indicated by demonstrations of attentive energy maps which accumulate eye centre positions over time.

Commonly, eye centre coordinates in each image frame are estimated and recorded as basic elements that form a gaze gesture sequence. Gaze gestures have been proposed in the literature for disability assistance and other HCI purposes (Wobbrock et al., 2008; Rozado et al., 2012). The realisation of remote control of a HCI system via gaze gestures is non-invasive, low-cost and efficient, involving four main stages: accurate eye centre localisation in the spatial-temporal domain, eye movement encoding (Drewes, Luca and Schmidt, 2007), gaze gesture recognition (Drewes and Schmidt, 2007) and HCI event activation.

For gaze gesture recognition, the algorithm introduced in the preceding section is employed for accurate and robust eye centre localisation in the first stage. Two eye centre positions $\boldsymbol{e_l}(f) = \{e_{lx}(f), e_{ly}(f)\}$ and $\boldsymbol{e_r}(f) = \left\{e_{rx}(f), e_{ry}(f)\right\}$ are estimated for each frame, where f is the frame number and the x and y notations stand for the horizontal and vertical components. The mean value of them is then calculated as $\bar{\boldsymbol{e}}(f) = \{\bar{e}_x(f), \bar{e}_y(f)\}$**.** The x and y coordinates of the left and the right eye centres in every frame are then recorded as four vectors: $\boldsymbol{e_{lx}}$, $\boldsymbol{e_{ly}}$, $\boldsymbol{e_{rx}}$ and $\boldsymbol{e_{ry}}$. Their velocities (first order derivatives) can then be calculated as: $\boldsymbol{v_{lx}} = \boldsymbol{e_{lx}}'$, $\boldsymbol{v_{ly}} = \boldsymbol{e_{ly}}'$, $\boldsymbol{v_{rx}} = \boldsymbol{e_{rx}}'$ and $\boldsymbol{v_{ry}} = \boldsymbol{e_{ry}}'$. The average velocity vector of the two eyes is then $\overline{\boldsymbol{v}}_{\boldsymbol{lx}} = \{\overline{\boldsymbol{v}}_{\boldsymbol{x}}, \overline{\boldsymbol{v}}_{\boldsymbol{y}}\} = \{\frac{(\boldsymbol{v_{lx}}+\boldsymbol{v_{rx}})}{\boldsymbol{2}}, \frac{(\boldsymbol{v_{ly}}+\boldsymbol{v_{ry}})}{\boldsymbol{2}}\}$.

A threshold (4.5 % in the experiments) is then set to remove any small values in the four velocity vectors, which might be caused by unintentional saccadic movements. It should be noted that the threshold is normalised by the pupillary distance ($P_d(f) = ||\boldsymbol{e_l}(f) - \boldsymbol{e_r}(f)||$) so that it is independent of user-to-camera distance. Additionally, the movements of the two eyes are compared with regard to their magnitudes and directions according to equation (4.13) and equation (4.14). These two equations account for the fact that the two eyes in natural behaviours always move together, i.e. the left and the right eye move toward the same direction and shift by similar amount.

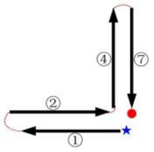
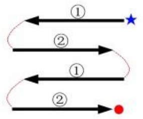
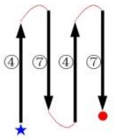
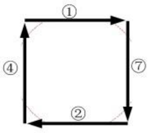
$$\boldsymbol{r}_{\boldsymbol{g}}(f) = \begin{cases} \log_{10}\left(\dfrac{\left|\left|\boldsymbol{v_{lx}}(f) + \boldsymbol{v_{ly}}(f)\right|\right|}{\left|\left|\boldsymbol{v_{rx}}(f) + \boldsymbol{v_{ry}}(f)\right|\right|}\right), & if \begin{matrix} ||\boldsymbol{v_{lx}}(f) + \boldsymbol{v_{ly}}(f)|| \geq 1 \\ ||\boldsymbol{v_{rx}}(f) + \boldsymbol{v_{ry}}(\boldsymbol{f})|| \geq 1 \end{matrix} \\ 0 \quad , & otherwise \end{cases} \tag{4.13}$$

$$\boldsymbol{d_g}(f) = \{d_{gx}(f), d_{gy}(f)\} = \{\boldsymbol{v_{lx}}(f) \cdot \boldsymbol{v_{rx}}(f), \boldsymbol{v_{ly}}(f) \cdot \boldsymbol{v_{ry}}(f)\} \tag{4.14}$$

When both $d_{gx}(f)$ and $d_{gy}(f)$ take positive values, the directions of eye movements are considered consistent. Ideally, when the two eyes shift by the same amount, $\boldsymbol{r}_{\boldsymbol{g}}(f)$ should have a value of 0. To allow for a margin of error, a threshold of 0.6 (found empirically) is set instead such that the magnitudes of eye movements are deemed consistent when the absolute value of R_g falls below the threshold. Only when these two conditions are satisfied are the eye centre positions updated by those from the subsequent frame.

Let a positive value in $\overline{\boldsymbol{v}}_{\boldsymbol{x}}$ be denoted by ‘1’ and a negative value by ‘2’, and let a positive value in $\overline{\boldsymbol{v}}_{\boldsymbol{y}}$ be denoted by ‘4’ and a negative value by ‘7’. Therefore ‘1’, ‘2’, ‘4’, ‘7’ are the encoded gaze shifts representing saccadic strokes ‘left’, ‘right’, ‘up’ and ‘down’. Similar saccadic encodings can be found in the literature (Drewes, Luca and Schmidt, 2007), but with the employment of a combination of digits and letters. This however complicates mathematical operations of the saccadic codes, i.e. it is not intuitive to add a digit with a letter. In the proposed method, the saccadic codes are selected to be digits only. They are chosen such that the addition of any pair of them will produce a code that is different to the existing ones. The two gaze shift vectors $\bar{\boldsymbol{v}}_{\boldsymbol{x}}$ and $\bar{\boldsymbol{v}}_{\boldsymbol{y}}$ are further summed and they produce $\bar{\boldsymbol{v}}_{\boldsymbol{s}}$, the integrated gaze shift vector. As a result, voluntary gaze shifts are recorded as a combination and repetition of the four digits, while a ‘0’ represents an unchanged eye position or an involuntary saccade. Finally, 7 types of gaze gestures are designed for HCI, shown in Table 4.2. Other saccadic codes (following mathematical operations ‘5’ = ‘1’+‘4’, ‘6’ = ‘2’+‘4’, ‘8’ = ‘1’+‘7’, ‘9’ = ‘2’+‘7’) denote diagonal saccadic strokes that are reserved for future works.

Table 4.2 Design of seven types of gaze gestures

Gesture No.	Gesture Sequence	Gesture Pattern	Gesture Name	HCI Event
1	1 → 2 → 4 → 7		Top-left gaze	Bring the *top-left* thumbnail advertisement to the screen centre
2/3/4	2 → 1 → 4 → 7/ 1 → 2 → 7 → 4/ 2 → 1 → 7 → 4	Similar to Gesture No. 1	Top-right gaze/ Bottom-left gaze/ Bottom-right gaze	Bring the *top-right / bottom-left / bottom-right* thumbnail advertisement to the screen centre
5	1 → 2 → 1 → 2		Reset gaze	*Reset* to default display
6	7 → 4 → 7 → 4		Zoom-in gaze	*Show details* of the selected advertisement (at the screen centre)
7	2 → 4 → 1 → 7 4 → 1 → 7 → 2 1 → 7 → 2 → 4 7 → 2 → 4 → 1		Change-content gaze	*Change all* the advertisement thumbnails

A '★' denotes the starting position of a gaze gesture and a '•' denotes the end of a gaze gesture; the circled numbers represent the encoded gaze shifts; the arrows denote saccadic strokes. Type 7 gaze gesture can start from any position. Type 6 gaze gesture will only be recognised as a subsequent gesture of type 1, 2, 3 or 4. It should be noted that these gaze gestures are designed to be intuitive in order to increase their usability. For example, the 'Zoom-in' gaze gesture is triggered by a user looking up and down twice. This is similar to the body language 'nodding', indicating that a user is keen to zoom in to view more of this advertising message. The 'Reset' gaze gesture, on the other hand, is similar to shaking one's head, indicating that the user is not interested of the displayed advertisement so as to request a reset of the content.

In the third stage, the gaze gesture patterns are recognised by searching for specific gesture sequences in $\overline{\boldsymbol{v}}_{\boldsymbol{s}}$, which will trigger pre-defined HCI events in the last stage.

It should be noted that the design of gaze gestures can be flexible such that it is suited for individual applications. The number of gaze gestures can be increased to provide higher functionality or can be decreased for better usability. Furthermore, eye saccades not only form gaze gestures for active HCI, they can be recorded by a system for a passive analysis in the form of an attentiveness energy map. By stacking all the midpoints of left and right eye centre coordinates, an attentiveness energy map can be constructed where denser clusters of high energy points can indicate the directions a user have gazed for longer time. Two case studies, a gaze gesture based map browser and a directed advertising billboard, are presented in Chapter 8 in order that the gaze gesture recognition algorithm can be evaluated from an applied point of view. By providing a means of contactless input to HCI systems, this approach should promise to serve HCI system users in generic scenarios, and especially to assist with the elderly and the disabled.

4.4 Discussion

4.4.1 The Benefits Brought by All the Modules in the Eye Centre Localisation Method

1. A Built-in Eye Detector

As stated previously, the first module computes an energy map for a whole face region (different from other isophote feature based methods that generate energy maps only for a pre-located eye region), and then reveals an eye centre candidate by calculating the energy moment. A rectangular region around the estimated centre location is cropped as the eye region that is fed to the second module for a regional analysis. As a result, the first module provides a built-in eye detector that can not only interact with the second module, but can be applied more generally to other eye detection tasks.

2. Tolerance to Head Poses

Also benefiting from the first module, the tolerance to head poses is embedded in the global analysis. When head rotation occurs, the relative eye position on a face changes dramatically and breaks the anthropometric relations set for a frontal face. Therefore, the extraction of a local facial region according to facial geometries can no longer ensure that it contains a complete eye region. This however has minimal influence on the proposed method since the true eye centre candidates always reside in the global energy map.

3. Tolerance to Shadows

The iris radius constraint drastically lowers the impacts (weights) of shadowed pixels even when they feature dark circular regions that resemble pupil/iris regions. This is facilitated by the prior knowledge that the size of a pupil/iris is a relative constant, compared to shadows of random radii. Therefore, when the radius of a dark circular region deviates from the pre-defined constant, the weights of the votes from this region should be diminished. However, the proposed method can be undermined by images where the pupil/iris regions are almost invisible (e.g. completely shadowed). This is because the proposed method is unsupervised and is intended to explore geometric features from the periocular regions. This limitation can be compensated by employing systems similar to that in Figure 3.1 to provide active illumination. Another possible solution is to utilise a classifier trained with poorly illuminated face/eye images to explore anthropometric relations or face appearance features.

4. Tolerance to Interfering Facial Edges

The SOG filter is what brings more robustness to the proposed algorithm in that it is able to measure the gradient similarity of any regular or irregular shape. The idea is based on the fact that the pupil/iris edges should be largely curved. Other gradients, regardless of their magnitudes, are more likely to be produced by shape edges from eyebrows, shadows and face accessories. The SOG filter results in a reduction of edges that have similar gradient directions, determined by means of statistical analysis.

5. Superior Efficiency

The simplicity and efficiency of the proposed eye centre localisation algorithm are further demonstrated by a comparison to the algorithm by Timm and Barth (2011), which claims to have achieved excellent real-time performance as one of its key features. Take the image containing a 41×47 eye region, i.e. 1927 pixels, as an example (Figure 4.6), the algorithm in comparison performs per-pixel estimation of the eye centre, assuming that every pixel is an eye centre candidate. Therefore 1927 iterations are needed before the optimal candidate is selected. The proposed method, on the other hand, resolves the problem by utilising the prior knowledge drawn from the first module which, through an initial estimation, avoids the per-pixel candidate assumption. The removal of the low-energy pixels in the first module largely reduced the number of candidates, i.e. number of iterations in the second module. In the 41×47 eye region, the iterations are decreased to only 67 from 1927, making the proposed algorithm 29 times faster.

The proposed algorithm was further tested with Microsoft Visual Studio 2012 and the

OpenCV library on a computer with an Inter(R) Core(TM) i5-4570 CPU and 12GB memory. The average execution time is calculated for every frame of a video (150 frames in total) recorded by a webcam at 640×480 resolution and at 30 frames per second. These results can be found in Table 3. Note that for face detection, the raw images were resized to a uniform size of 320×240 pixels, given the fact that the Viola-Jones detector was trained on small face images. However, the detected face regions are cropped from the raw images and were resized to 7 different sizes for evaluating their respective computational cost. Evaluation experiments on all the databases employed face regions resized to 128×128 and obtained superior results. The frame rate of 37 frames per second in this setting ensures that the proposed algorithm can run in real time. The algorithm performs eye analysis for every image frame instead of tracking face and eye movements across frames. This is because erroneous tracking results from a particular frame can cause larger errors in subsequent frames and eventually cause the algorithm to go beyond recovery. In comparison, per-frame detection/localisation does not generate localisation results based on preceding frames so as to avoid error propagation. Therefore, it can be deemed more applicable to real-world scenarios where human behaviours and environmental variations are dynamic and complex.

Table 4.3 The efficiency of the proposed eye centre localisation and gaze gesture recognition algorithm

Face size (pixels)	Execution time (milliseconds)				Frame rate (frame per second)	
	for face detection	for eye centre	for gaze gesture	Total	exclude face detection	include face detection
64 × 64	21.3 (raw images resized)	1.2	0.1	22.6	>700	44
96 × 96		2.9		24.3	>300	41
128 × 128		5.6		27	176	37
160 × 160		9.4		30.8	105	32
192 × 192		18.8		40.2	53	25
224 × 224		28.8		50.2	35	20
256 × 256		29.4		50.8	34	20

In Table 4.3, those underlined are the settings for the evaluation experiments on the publicly available datasets and the self-collected dataset introduced in the preceding section. With a webcam running at 30 frames per second, the proposed algorithm is capable of localising eye centres in detected faces accurately in real time. This would allow for its implementation, in various forms, for control of assistive technologies and other types of HCI systems.

4.4.2 Gaze Gesture vs. Eye Fixation

It has been illustrated previously that gaze gestures reflect relative eye movement while gaze fixation points reveal absolute gaze direction (Hyrskykari, Istance and Vickers, 2012). In Table 4.4, a comparison of the two types of gaze analysis is listed with regard to 7 criteria.

Table 4.4 A Comparison of Gaze Gesture and Eye Fixation

Criterion	Gaze gesture	Eye fixation
user attention measurement	relative	absolute
prerequisite	instructions	device-specific
calibration	not needed	often required
influence of user-camera distance	small	large
influence of head pose	small	large
hardware dependency	low	high
interaction level	high	low

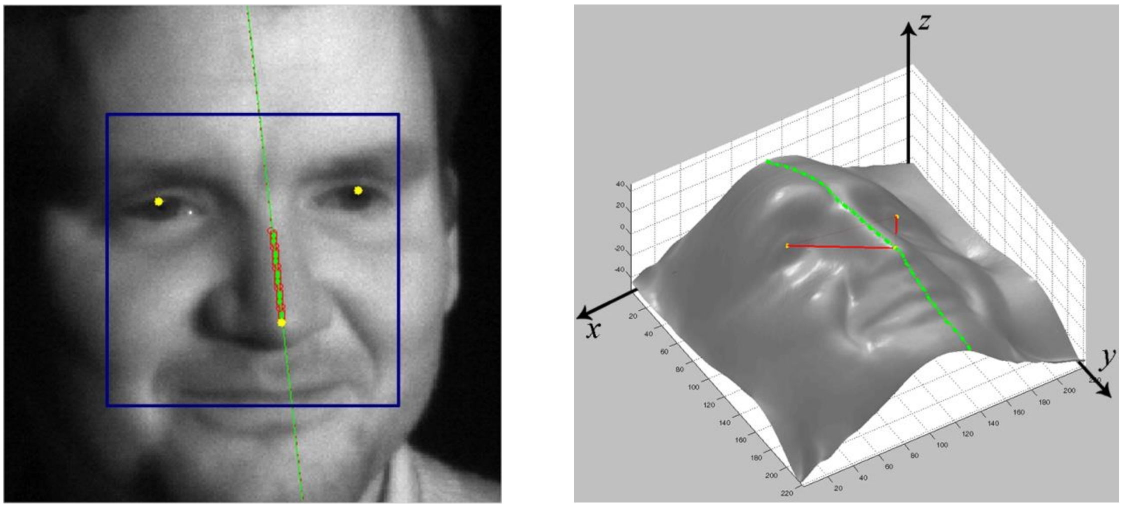
In the gaze gesture experiments, algorithms for the recognition of diagonal saccadic strokes were also implemented. However, it appeared difficult for inexperienced users to issue diagonal saccadic strokes accurately. As a result, only horizontal and vertical saccadic strokes are adopted eventually. Nevertheless, different combinations of horizontal and vertical strokes should give sufficient flexibility for HCI.

4.4.3 Eye fixation from 2D and 3D facial landmarks

As reflected by Table 4.4, gaze gesture analysis and eye fixation analysis have their respective advantages by estimating relative gaze or absolute gaze. Although this research places focus on analysis of relative gaze due to it being calibration-free, robust and interactive, preliminary works on eye fixation analysis have been carried out in order that the effectiveness and applicability of the proposed eye centre localisation algorithm (see Section 4.1) and the two-source photometric stereo (PS) algorithm (see Section 6.2) can be further demonstrated. The method and results delivered by these works are preliminary but are nevertheless significant to future works on facial landmark detection and head pose estimation, as well as eye fixation analysis.

Most facial landmark detection approaches are based on colour or greyscale images. While the centre of an eye can be reliably located by exploring its unique geometry and pattern in 2D, nasal tip is a type of facial landmark that is skin-coloured with relatively homogeneous reflectance. In addition, the protruding nasal profile commonly results in unpredictable changes (e.g. shadows or specularities) in colour or greyscale images due to non-uniform illumination or varied viewpoint. However, it is the unique nasal topology that allows the nasal area to be distinctive in a 3D face image. Therefore, to resort to 3D facial topology is intuitively an effective and robust way for nasal tip detection.

Following Section 4.1 where 2D based eye centre localisation was introduced, this subsection introduces a 3D based method for nasal tip localisation, as preliminary works for head pose estimation and eye fixation analysis. The two-source PS method utilised for 3D face reconstruction will be introduced in Section 6.2.



(a)

(b)

Figure 4.17 An example of 2D and 3D based facial landmark detection, displayed in (a) a greyscale face image and (b) a 3D depth face image.

Given a 3D face image (see Figure 4.17(b)) reconstructed by the proposed two-source PS algorithm, this method firstly selects a region of interest by defining a bounding box to exclude image margins that are 20% of the image size (see Figure 4.17(a) as a visualisation in 2D). This is performed such that redundant information (e.g. background or protruding cheekbones) is removed. It then locates the maximum value along the z axis and records its coordinate as the nasal tip (marked in yellow in Figure 4.17). This is due to the fact that the nasal tip is commonly the highest point along the z axis, considering the 3D plane defined in Figure 4.17(b). Searches for the maximum z value in each row of the image are then repeated above the nasal tip, for 20% of the total rows in the face image. These points are assumed to be samples from the raised nasal bridge. Linear regression is then applied to these samples which are fitted by a linear polynomial, displayed in red in Figure 4.17(a). Samples that cause large regression errors (larger than the average error from all samples) are then removed, which can be caused by image aliasing, image noise or 3D reconstruction error. For a second time, linear regression is applied to the remaining samples, giving a linear polynomial that should pass the nasal bridge, displayed in green in Figure 4.17.

The face image in Figure 4.17(a) is one of the PS images recorded during the data capture experiments introduced in Chapter 3. That in Figure 4.17(b) is a reconstructed 3D face image from a pair of PS images recorded during the same experiments. Therefore, the facial landmarks localised on both images are spatially consistent. The eye centres estimated from the 2D image and the nasal tip from the 3D image form a triangle that deforms when the head pose changes. According to facial anthropometrics, a perfectly frontal face corresponds to an isosceles triangle in the x-y 2D plane. In-plane head roll rotations (around the z axis) will only cause the isosceles triangle to rotate in the 2D plane, but not to change its shape; Head pitch rotations (around the x axis) will cause variations to the sides of the isosceles triangle; Head yaw rotations (around the y axis) will result in the triangle being scalene. As a result, head poses can be estimated from the deformation of this triangle, in a similar way to the study by Kaminski, Knaan and Shavit (2009). When the triangle is further calibrated in HCI applications, higher accuracy can be expected. With known head pose and localised eye centres, estimation of eye fixation is then achievable. This part of research is further outlined in the ‘Future Work’ section (Section 9.3).

4.5 Summary

This chapter introduces a novel method for eye centre localisation and gaze gesture recognition which is intended to use in gathering human behavioural data for enabling a better HCI experience. The main contributions of the proposed methods include the following algorithm designs: 1) a modular eye centre localisation method consisting of two modules and a SOG filter, and 2) communication through initiation of gaze gestures and employment of a gaze gesture recognition algorithm.

Tested on the BioID database, the proposed eye centre localisation approach has the highest accuracy, outperforming ten other methods. Tested on the GI4E database where different head poses are present, the proposed method outperformed all the six other methods in comparison. Tested on a subset of the extended Yale Database b, the proposed method exhibited excellent robustness under challenging illumination conditions. Moreover, results from three groups of self-collected data have again validated the high accuracy and robustness of the proposed eye centre localisation method. Based on the high reliability of the proposed eye centre localisation approach, seven types of gaze gestures have been designed which, upon recognition, trigger pre-defined HCI events in real time and therefore allow for user-centred HCI in an active and contactless manner. The practicability of gaze gestures is demonstrated by two types of case study HCI systems in Chapter 8.

Overall, the two algorithms introduced in this chapter contribute theoretically to research in the area of eye/gaze analysis; as well as demonstrating their potential for providing a richer HCI experience. The modular eye centre localisation approach only employs 2D facial features, while the proposed nose tip localisation algorithm is based on 3D face topography. As discussed in Section 4.4.3, the utilisation of 2D and 3D facial data can enable head pose estimation and eye fixation estimation, as well as gaze gesture analysis. This has the potential to bring increased functionality and usability to HCI implementations.

Chapter 5 Fisher Vectors for 2D Gender Recognition

Understanding a user is essential for creation of user-centred HCI environments. In natural face-to-face communications, gender is the very first piece of demographic information perceived by human beings. This explains why numerous works have been carried out to automatically recognise gender labels of human beings. As stressed in Section 1.2, gender recognition is a more favourable route chosen by this research over face recognition, in that it avoids directly linking face images to personal identities so that identity-neutral HCI strategies can be put forward to serve human beings routinely and extensively. It is also considered more effective than ethnicity recognition since variations due to ethnicity in daily lives are decreasing with countries and societies becoming increasingly multi-ethnic. The physical and mental distinctions due to gender, however, are intrinsic to some level and thus lead to the respective needs and preferences of male and female groups.

As seen from the literature, most methods regarding gender recognition lack robustness and suffer from various limitations in real-world scenarios. To bridge these gaps, novel methods are explored in this chapter, which can boost the discriminative power of facial features while overcoming challenging environmental variations. In this chapter, a Fisher Vector (FV) encoding method is proposed to reliably predict gender labels from greyscale facial images. Subsequent chapters further extend the FV encoding method such that 3D facial features can be employed and age classification can also be achieved. Section 8.2 further demonstrates how gender and age recognition can be beneficial to real-world HCI scenarios, applicable but not restricted to the case study – directed advertising.

5.1 Fisher Vector principle

A Fisher Vector (FV) is an encoded vector that applies Fisher kernels on visual vocabularies where the visual words are represented by means of a Gaussian Mixture Model (GMM). The Fisher kernel function is derived from a generative probability model, and provides a generic mechanism that combines the advantages of generative and discriminative approaches, meaning that it has the potential to construct a powerful face descriptor as well as benefiting a particular classification task.

As a core component of a FV, a GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities as given by equation (5.1) (Reynolds, 2009)

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} \beta_i \, g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\delta_i}) \tag{5.1}$$

where $\boldsymbol{x}$ is a D-dimensional data vector, $\lambda = \{\beta_i, \boldsymbol{\mu_i}, \boldsymbol{\delta_i}, i = 1,2, \ldots, N\}$ is the collective representation of the GMM parameters – β_i the mixture weights, $\boldsymbol{\mu_i}$ the mean vector and $\boldsymbol{\delta_i}$ the covariance matrix. N is the number of Gaussians. The component $g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\delta_i})$ is further described in equation (5.2), while the mixture weights are subject to the constraint in equation (5.3) (Perronnin and Dance, 2007).

$$g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\delta_i}) = \frac{e^{\{-\frac{1}{2}(x-\boldsymbol{\mu_i})' \boldsymbol{\delta_i}^{-1}(x-\boldsymbol{\mu_i})\}}}{(2\pi)^{D/2}|\boldsymbol{\delta_i}|^{1/2}} \tag{5.2}$$

$$\sum_{i=1}^{N} \beta_i = 1 \tag{5.3}$$

The covariance matrices are assumed to be diagonal since any distribution can be decomposed into a number of weighted Gaussians with diagonal covariances. Let $\boldsymbol{X} = \{\boldsymbol{x_t}, t = 1,2, \ldots, T\}$ be the set of low-level features vectors extracted from an image, and it is assumed that all the vectors are independent. Equation (5.4) can be found (Perronnin and Dance, 2007):

$$log\, p(\boldsymbol{X}|\lambda) = \sum_{t=1}^{T} log\, p(\boldsymbol{x_t}|\lambda) \tag{5.4}$$

As it can be assumed that the generation of $\boldsymbol{X}$ samples is modelled by the probability density function $p(\boldsymbol{X}|\lambda)$, $\boldsymbol{X}$ can be characterised by the gradient vector:

$$\psi_\lambda^{\boldsymbol{X}} = \frac{\nabla_\lambda\, logp(\boldsymbol{X}|\lambda)}{T} \tag{5.5}$$

The parameters of the GMM are tuned to best fit $\boldsymbol{X}$ samples according to direction defined by this gradient of the log-likelihood, which is responsible for transforming a variable length $\boldsymbol{X}$ into a fixed length vector that is only dependent on GMM parameters. A natural kernel (Perronnin and Dance, 2007) on these gradients is:

$$\varkappa(\boldsymbol{X}, \boldsymbol{Y}) = \psi_\lambda^{\boldsymbol{X}'}\, \mathcal{F}_\lambda^{-1}\, \psi_\lambda^{\boldsymbol{Y}} \tag{5.6}$$

$$\mathcal{F}_\lambda = E_X\{\nabla_\lambda\, logp(\boldsymbol{X}|\lambda)\, \nabla_\lambda\, logp(\boldsymbol{X}|\lambda)'\} \tag{5.7}$$

where $\mathcal{F}_\lambda$ is the Fisher information matrix (Jaakkola and Haussler, 1999), which can be decomposed as $\mathcal{L}_\lambda{}'\mathcal{L}_\lambda$; $\Psi_\lambda^{\boldsymbol{X}} = \mathcal{L}_\lambda\, \psi_\lambda^{\boldsymbol{X}}$ is then referred to as the Fisher Vector of $\boldsymbol{X}$.

Let $\gamma_t(i)$ denote the soft assignment of vector $\boldsymbol{x_t}$ to the Gaussian component i:

$$\gamma_t(i) = p(i|\boldsymbol{x_t}, \lambda) = \frac{\beta_i g(\boldsymbol{x_t}|\boldsymbol{\mu_i}, \boldsymbol{\delta_i})}{\sum_{j=1}^{N} \beta_j \, g\left(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\delta_j}\right)} \tag{5.8}$$

The gradients of Gaussian component i with respect to the mean $\boldsymbol{\mu_i}$ and the covariance $\boldsymbol{\sigma_i}$ respectively are (Simonyan et al., 2013):

$$\Psi_{\mu,i}^{\boldsymbol{X}} = \frac{1}{T\sqrt{\beta_i}} \sum_{i=1}^{T} \gamma_t(i) \left(\frac{\boldsymbol{x_t} - \boldsymbol{\mu_i}}{\boldsymbol{\delta_i}}\right) \tag{5.9}$$

$$\Psi_{\delta,i}^{\boldsymbol{X}} = \frac{1}{T\sqrt{2\beta_i}} \sum_{i=1}^{T} \gamma_t(i) \left[\frac{(\boldsymbol{x_t} - \boldsymbol{\mu_i})^2}{\boldsymbol{\delta_i}^2} - 1\right] \tag{5.10}$$

Finally a FV is represented as:

$$\boldsymbol{\Phi} = \left\{\Psi_{\mu,1}^{\boldsymbol{X}}, \Psi_{\delta,1}^{\boldsymbol{X}}, \dots, \Psi_{\mu,N}^{\boldsymbol{X}}, \Psi_{\delta,N}^{\boldsymbol{X}}\right\} \tag{5.11}$$

A FV derived from the GMM parameters describes the variation of the Gaussian distribution of one image from that of the entire training database. As a result, a FV as a feature vector is provided with contextual definition and results in enhanced saliency for classification.

5.2 Fisher Vector encoding for gender recognition

FVs have been used for face recognition and have proved to be superior to a number of other low-level feature types and encoded feature types (Simonyan et al., 2013). The FV encoding approach consists of five main stages: 1) face pre-processing, 2) low-level feature and face descriptor computation, 3) dimensionality reduction, 4) FV encoding and 5) classifier training and feature selection. The five primary stages adopted by the proposed method are illustrated in Figure 5.1.

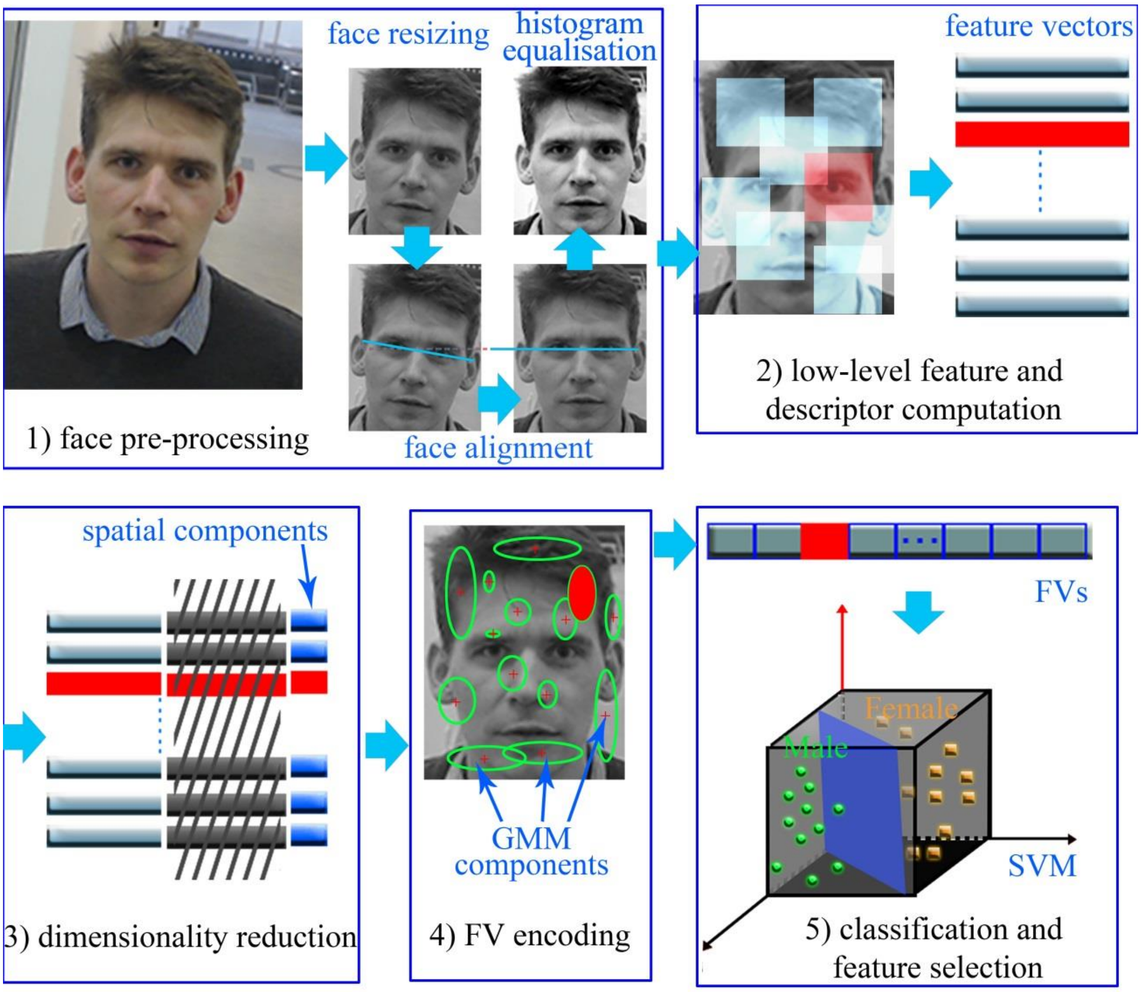


Figure 5.1 An illustration of the five primary stages in a FV encoding process. The correlation between a facial region and a FV segment is indicated by the red regions. Histogram equalisation and facial alignment are experimented in this study, but are not compulsory in this approach.

In more detail, these stages are illustrated as follows:

1) Face pre-processing

The techniques experimented at this stage include face detection, image resizing, histogram equalisation and face alignment. In the experiment, the Viola-Jones face detector is used to obtain the face region in the first place. The facial region obtained by the face detector is then reshaped so that it incorporates the hair region and the chin. All the reshaped facial regions are further resized to the same size in the next step. Face alignment and histogram equalisation have also been experimented with so as to evaluate their impact on the recognition accuracy.

2) Low-level feature and face descriptor computation

Dense descriptors at every pixel location are extracted. Firstly, a face image is divided into a number of overlapping patches of the same size. Specifically, these patches are obtained by sliding a $ps \times ps$ window across an image horizontally and vertically by a predefined sampling step ss ($ss \in \mathbb{Z}$). One descriptor per patch rather than one descriptor per image is constructed. The geometry of the patch-based descriptors is shown in Figure 5.2.

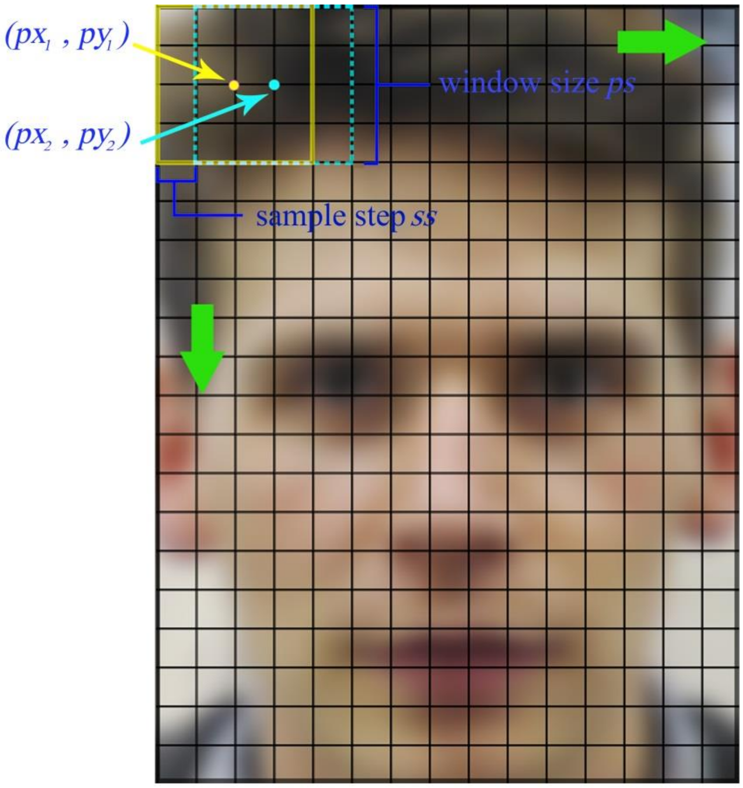


Figure 5.2 Geometry of patch-based dense feature descriptors

For example, vector (px_c, py_c) records the centre position of the c^{th} $(c \in \mathbb{N}, c \leq pn)$ patch in the image. The centre position of the first patch is therefore $(ps/2, ps/2)$. For an $m \times n$ image, the total number of patches is:

$$pn = \frac{m - ps + 1}{ss} \times \frac{n - ps + 1}{ss}, \quad \begin{matrix} ps < \min(m, n) \\ 1 \leq ss \leq \min(m, n) \end{matrix} \tag{5.12}$$

For every patch, features are sampled densely on all pixels, collectively forming a descriptor.

Although a number of pixels at the margins of an image cannot be the centres of a sliding window, they are incorporated by at least one image patch and are therefore involved in the formation of image descriptors.

Low-level dense features can be visualised by mapping the top three principal components to the RGB colour space. A similar process can be found in a study by Liu, Yuen and Torralba (2011), which mapped dense SIFT features to the principal RGB colour space. As an example, dense SIFT features extracted with different parameters can be seen in Figure 5.3.

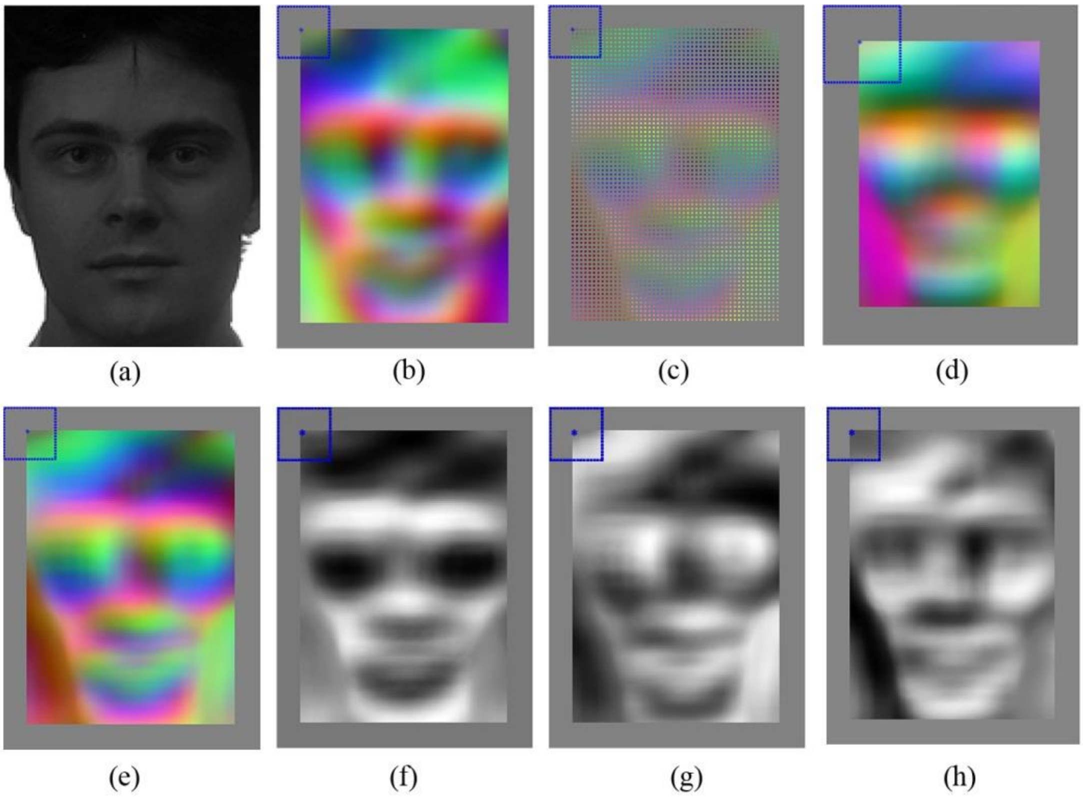


Figure 5.3 Dense SIFT feature visualisation in the RGB colour space, with different parameters. The square on the top left corner denotes the size and centre of the first descriptor patch. (a) The original facial image. (b) Visualisation for ps = 24 and ss = 1, with histogram equalisation and root SIFT applied. (c) Visualisation for ps = 24 and ss = 2. (d) Visualisation for ps = 36 and ss = 1. (e) Visualisation for ps = 24 and ss = 1. (f) Visualisation of the first principal component for ps = 24 and ss = 1. (g) Visualisation of the second principal component for ps = 24 and ss = 1. (h) Visualisation of the third principal component for ps = 24 and ss = 1.

The visualisation implies that the regions of similar structure are captured by densely sampled local features, except for the grey margins where no sampling windows can be applied to for feature extraction. It should be noted that only 3 of the 128 principal components are used for visualisation and therefore the structural cues in Figure 5.3 only manifest a general and coarse representation. In practice, 64 principal components (after dimensionality reduction) are used so that fine edges and facial structures can be reflected in more detail by these dense local features.

The majority of feature types, regardless of them being extracted from a 2D face or a 3D face, can be embedded into this framework by the FV encoding method. These low level features can be visualised in similar ways. This provides the algorithm with excellent generalisation property in that FVs can be derived from almost any type of feature.

3) Dimensionality reduction

As one image produces multiple descriptors, linearly increasing the total number of observations as opposed to conventional methods, dimension reduction is essential in that it cuts down memory consumption and that it potentially compresses raw features into more discriminative representations.

Principal Component Analysis (PCA) is employed to reduce the dimensionality to 64 features per descriptor. Being an unsupervised technique, PCA decorrelates the features while avoiding the removal of class specific information. It is assumed that every patch is large enough to produce a descriptor whose length is larger than 64 so that the dimensionality reduction is valid. For example, when intensity values are directly used as features, each patch should be larger than 8×8 pixels.

For aid with visualisation, the centre position of a patch is appended to the end of the corresponding compressed descriptor in the form of a 2D vector (Simonyan et al., 2013), increasing the dimensionality from 64 to 66. This vector is rescaled to [-0.5, 0.5] so that it has minimal effect compared to the other 64 features. As a result, although spatial information is embedded into the feature vectors, the overall features are still relatively independent of global facial geometry and are therefore robust to variation caused by head rotation. However local geometry is still represented by individual patches and the size of a patch determines the extent of geometric information to be preserved.

4) FV encoding

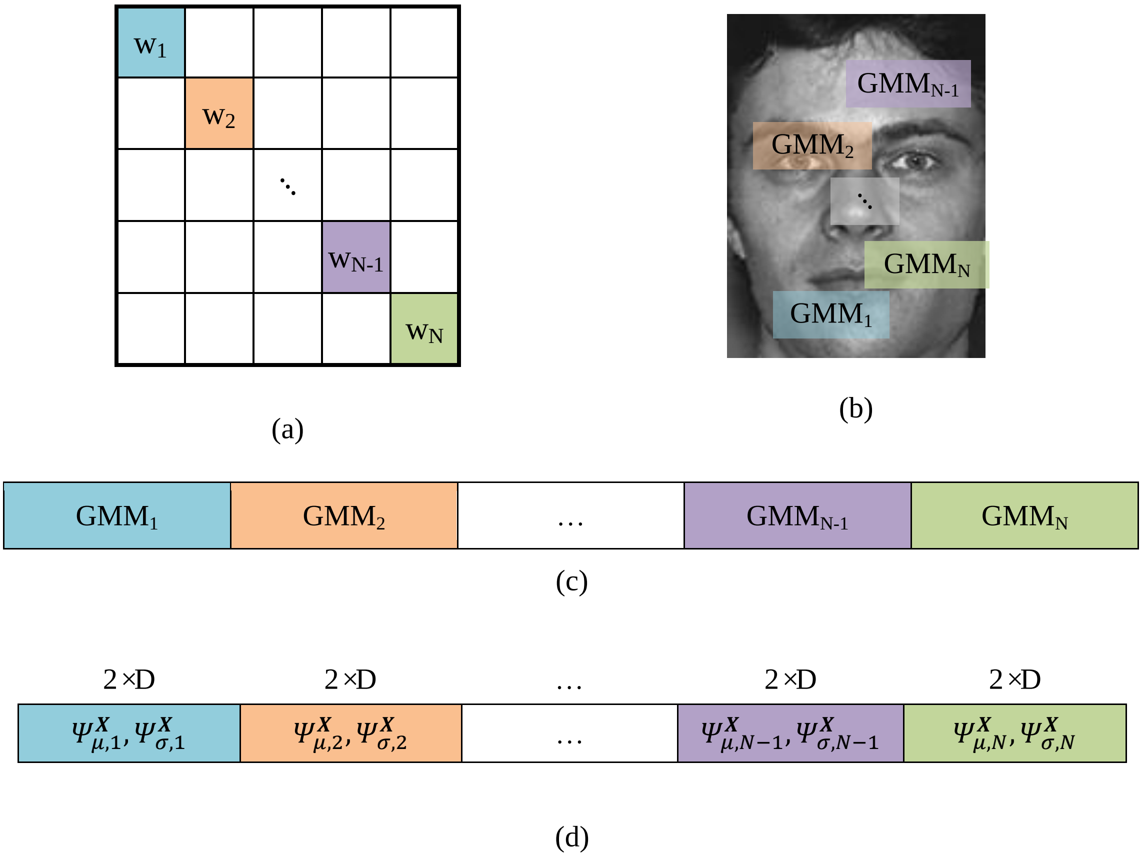
The aggregate of all descriptors from the training images trains a GMM and yields the model parameters which, according to equation (5.9) – equation (5.11), lead to FVs as encoded features. For example, when the trained GMM has 512 components, it can be calculated from equation (5.11) that the dimension of a FV is $2 \times N \times D = 2 \times 512 \times 66 = 67584$, where N is the number of Gaussian components and D is the dimensionality of a patch-based feature descriptor.

In this stage, the decomposed patches are reunited to characterise a complete image, in the form of the derivatives of all Gaussian components. In the computation of FVs, ℓ_2 normalisation and power normalisation are applied to the vectors since they are reported to bring improved classification performance (Perronnin, Sánchez and Mensink, 2010). In the experiments, a publicly available toolbox (Vedaldi and Fulkerson, 2010) was utilised for GMM training and SIFT feature extraction.

5) Classifier training

The algorithm learns a SVM classifier from all the training FVs. The employment of a SVM classifier has a twofold purpose. Firstly, it naturally performs classification tasks as an excellent classifier, with or without a kernel. The output from it gives the predicted labels (e.g. age or gender) for all testing images. Secondly, when acting as a linear classifier, it reveals the discriminative power of each variable/feature in the FVs. From the perspective of a SVM classifier, a SVM learns a hyper-plane that separates two classes with a maximum margin. As the hyper-plane is defined by a decision hyper-plane normal vector $\vec{\boldsymbol{w}}$ which is perpendicular to the hyper-plane and an intercept term b, the absolute values of the elements in $\vec{\boldsymbol{w}}$ imply the significance of the corresponding elements in the FVs. From the perspective of metric learning, the diagonal linear transformation matrix $\boldsymbol{W}$ to be learnt has its diagonal values as in $\vec{\boldsymbol{w}}$. It can be considered as projecting the original data so that they are located on each side of a fixed hyper-plane, different from the former perspective where the data are fixed and the hyper-plane is unknown (Do et al., 2012). Both methods state that $\vec{\boldsymbol{w}}$, also known as the weight vector, reflects the discriminative power of individual features.

From the manner in which a FV is constructed (equation (5.11)), a mapping can be obtained between the features and the Gaussian components. As well as the mapping between the discriminative power and an individual feature, the relationship between the discriminative power and each Gaussian component can be established. When visualised with regard to its spatial location, a Gaussian component can be used to signify the discriminability of its corresponding facial region. The establishment of this mapping is illustrated in Figure 5.4.



w_1
w_2
$\ddots$
w_{N-1}
w_N
(a)
GMM_{N-1}
GMM_2
$\ddots$
GMM_N
GMM_1
(b)
GMM_1
GMM_2
…
GMM_{N-1}
GMM_N
(c)
2×D
2×D
…
2×D
2×D
$\Psi^X_{\mu,1}, \Psi^X_{\sigma,1}$
$\Psi^X_{\mu,2}, \Psi^X_{\sigma,2}$
…
$\Psi^X_{\mu,N-1}, \Psi^X_{\sigma,N-1}$
$\Psi^X_{\mu,N}, \Psi^X_{\sigma,N}$
(d)

Figure 5.4 Establishment of the mapping between the discriminative power and a facial region represented by the spatial component of a Gaussian. The mapping is colour coded in this example where those with the same colour are correlated. (a) A transformation matrix $\boldsymbol{W}$ with $\boldsymbol{w_1}$ to $\boldsymbol{w_N}$ being its diagonal elements obtained by a SVM learnt from a set of FVs. (b) A cropped facial image from the Grey FERET Database. GMM components are mapped to the image according to their spatial components. (c) An illustration of N GMM components. (d) The concatenated features that form a FV where $2\times D$ features are generated from one GMM component.

A linear SVM is trained to seek the most discriminative facial regions as well as predicting class labels of all testing images. A SVM with Radial Basis Function (RBF) kernel is also trained solely for the purpose of classification, in comparison to the linear SVM. This training process and the classification results can be found in the subsequent section.

Figure 5.1 and Figure 5.4 demonstrate the detection of the most discriminative facial regions. To sum up, the linear SVM detects the most discriminative features in the FVs which correspond to certain Gaussian components in the GMM. Therefore the most discriminative Gaussian components can be identified. As low-level features bear spatial information when training the GMM, the locations of these Gaussian components can be restored on the facial images and therefore reveal the discriminative facial regions with regard to gender recognition.

5.3 Gender Recognition Experiments and Results on Publicly Available Databases

In order to evaluate the performance of the proposed gender recognition algorithm under a controlled environment and real-world condition respectively, the Grey FERET database (Phillips et al., 1988), the Labelled Face in the Wild (LFW) database (Huang et al., 2007) and the FRGCv2 database (Phillips et al., 2005) were employed.

The Grey FERET database (referred to as the FERET database in the rest of this thesis) consists of 14051 greyscale images of frontal and profile face images. The fa partition of 1152 male patterns and 610 female patterns is used. Although captured under controlled environment, the FERET database still poses great challenges as it accommodates different ethnicities, facial expressions, facial accessories, facial makeup and illumination conditions. Some sample images from the FERET database partition fa are shown in Figure 5.5.

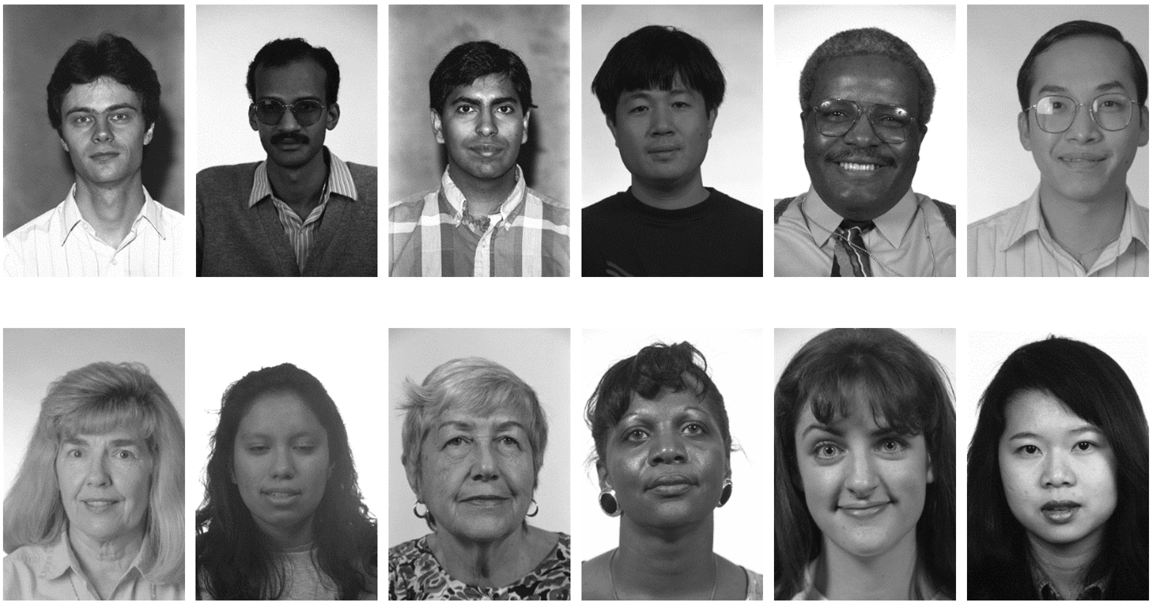


Figure 5.5 Sample images from the FERET database partition fa

The LFW database is considered one of the most challenging databases and has become the evaluation benchmark for face recognition under unconstrained environments. The 13233 colour facial images of 5749 subjects collected from the web include almost all types of variation and interferences (illumination, head poses, occlusion, image blur, chromatic distortion, etc.) and come in inconsistent image quality. Only one image per subject (the first image) is used in the experiment so that the same subject cannot appear in both the training set and the testing set. Some sample images from the LFW database are shown in Figure 5.6.

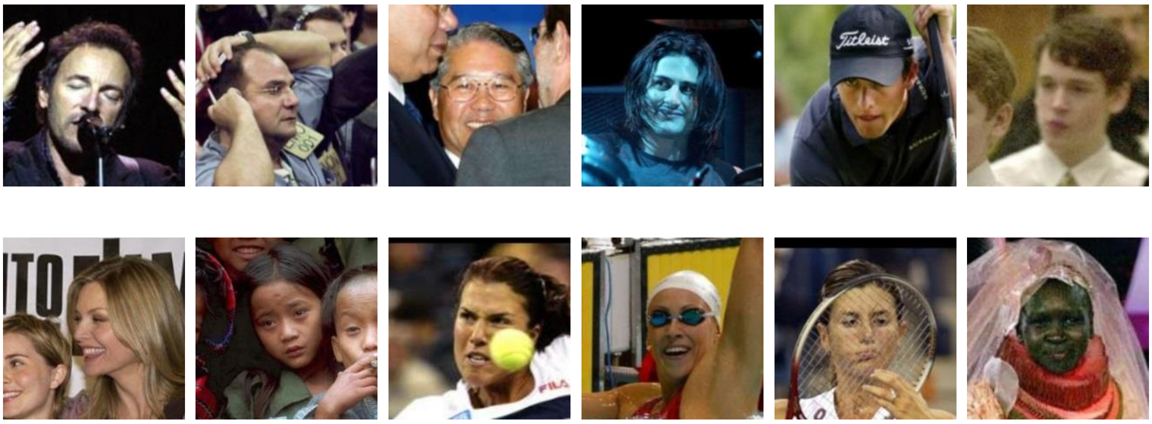


Figure 5.6 Sample images from the LFW database

The FRGCv2 database includes 4007 depth images belonging to 466 subjects. These data also include different ethnicities and age groups. Some sample images from this database containing male or female images are displayed in Figure 5.7.

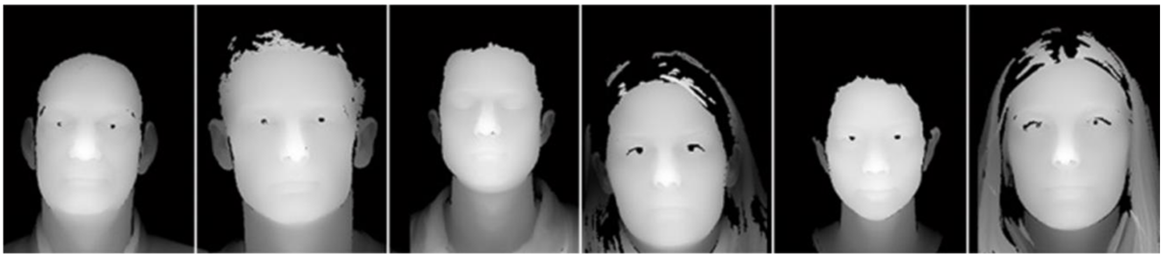


Figure 5.7 Sample images (depth images) from the FRGCv2 database

For all the databases used in the experiments, it holds that one particular subject only appears in either the training set or the testing set. The five-fold cross validation technique is used for evaluation. More specifically, the database is partitioned into 5 splits of similar size, 4 of which are used for training in each repetition while the remaining 1 split for testing. After 5 repetitions, the average classification rate is calculated as the final result.

Following the stages previously stated, different types of features were evaluated including the greyscales, LBP features (extracted using the circularly symmetric neighbour sets (Ojala, Pietikäinen and Mäenpää, 2000) with 8 neighbouring pixels and radius of 1), LBP histogram with uniform pattern, and SIFT features. Their respective performances on the FERET database are summarised in Figure 5.8.

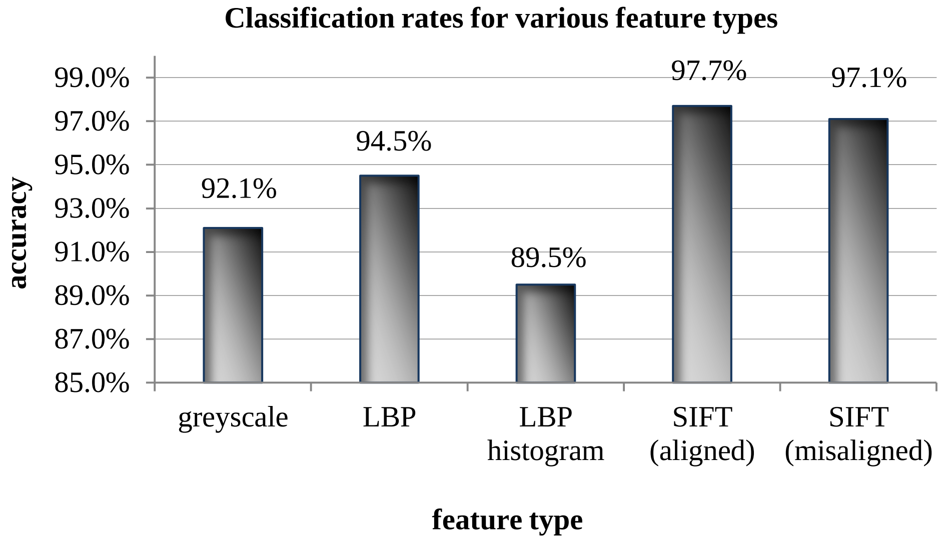


Figure 5.8 Gender classification rates for various feature types, tested on the FERET database

In the evaluation experiments, various parametric settings were investigated, among which the parameters that gave the highest accuracy were identified. The parameters inspected concern the size of the image, the size of the window for local patches, the sampling step, the number of components in the GMM, and the number of principal components calculated by PCA. One varying parameter is inspected at a time while the others remain fixed as the same specifications as in the last group of experiments (if not otherwise specified). However misaligned and non-normalised face images are used in this group of experiments.

1) Image size

The maximum size of the selected face regions from the FERET database is 160×120. With the aspect ratio fixed, it is scaled down by 10% of the maximum size to a minimum of 30% of the maximum size. Figure 5.9 reflects the declined performance as the size decreases.

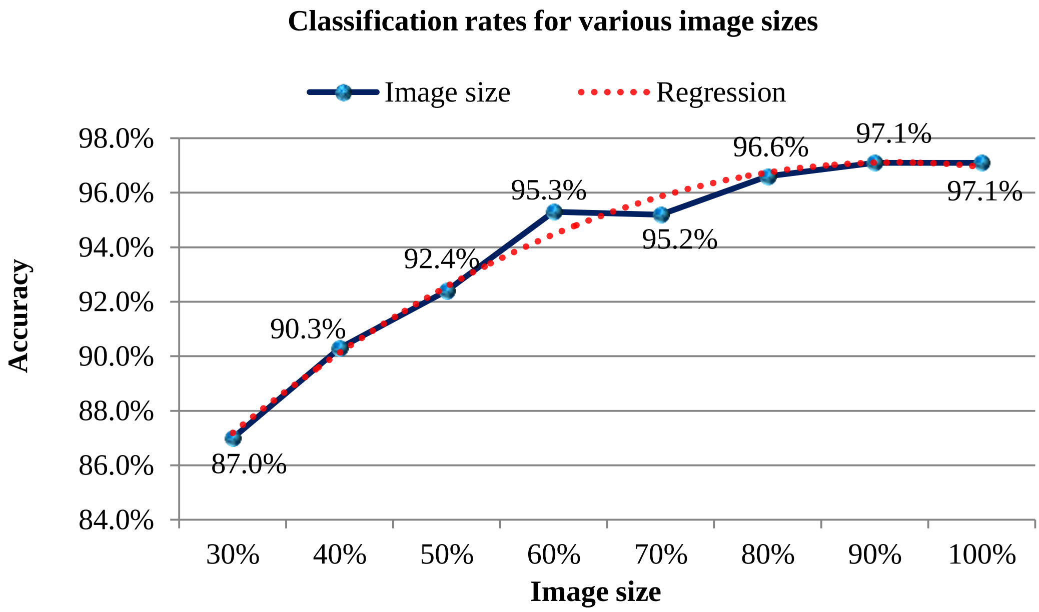


Figure 5.9: Gender classification rates for various image sizes, tested on the FERET database. The fluctuation trend is described by a quadratic polynomial regression fit, represented by the red dotted line.

The results suggest that larger images tend to produce higher classification rates. This coincides with the intuition that more details reside in larger images which generate more features for classification.

2) Window size of local patches

The algorithm introduced is an appearance-based method that performs at local level. It has the advantage of being insensitive to variation caused by illumination, viewpoint and facial expression. At the stage of low-level feature extraction, densely sampled features are drawn from the local image patches via a sliding window (see Figure 5.2). Broadly speaking, a larger window implicitly maintains more geometric information, defined by the spatial relationship amongst all pixels within the window. An extreme case concerns a window as large as the entire image which removes the advantage of using local features. Hence there is a critical need for the window to be defined with an appropriate size that allows sufficient tolerance to the variations.

The window size ranges from 12×12 to 40×40 in the experiments, with an interval of 4 pixels. The respective classification rate is summarised in Figure 5.10.

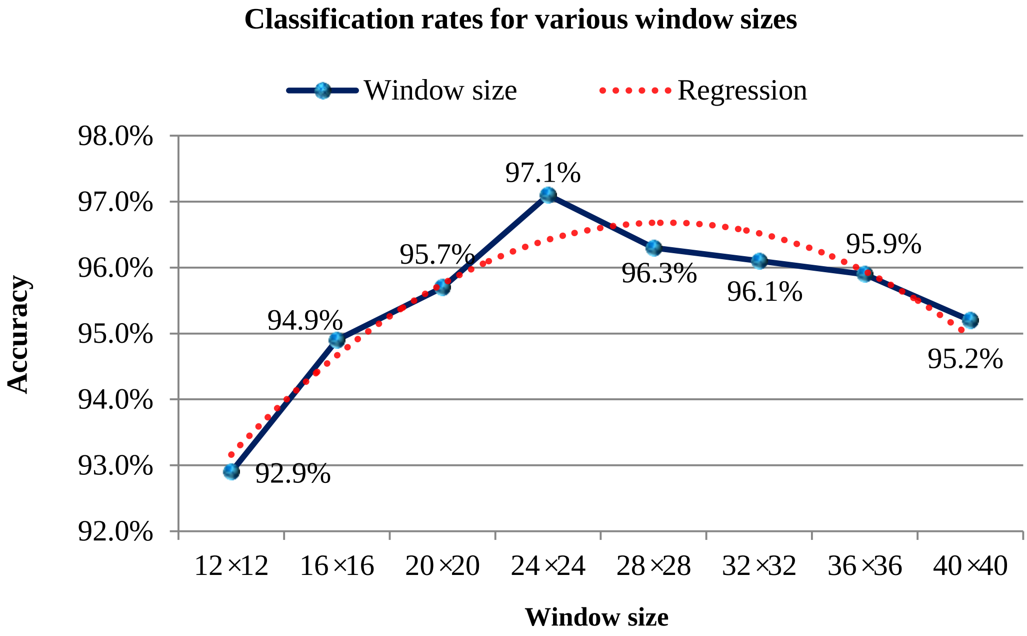


Figure 5.10: Gender classification rates for various window sizes, tested on the FERET database. The fluctuation trend is described by a quadratic polynomial regression fit, represented by the red dotted line.

It can be concluded from Figure 5.10 that a window 15% to 20% the size of the entire image is optimal, although its impact is seemingly less than the size of the images where the features are extracted.

3) Sampling step of local patches

One drawback of this method is that the appearance-based method at local level produces a great number of descriptors. The storage for these dense samples is thus more demanding than that for a method at global/holistic level. The sampling step sets a density restriction that controls the number of descriptors to be generated per image. Increasing sampling step in extracting local descriptors is comparable to decreasing the sampling rate in sub-sampling an image. Figure 5.11 summarises the impact of varying the sampling step. To avoid high memory usage, the facial images are further resized to

96×72 from 160×120. Sampling steps from 2 to 7 were experimented with in this setting. Descriptors with the highest density (the sampling step being 1 pixel) were not experimented with, because of the associated high memory usage for data storage. To compensate for this, the first 500 male and female images (1000 in total) were then employed and further resized to 80×60, in order that experiments with sampling step 1 could be conducted.

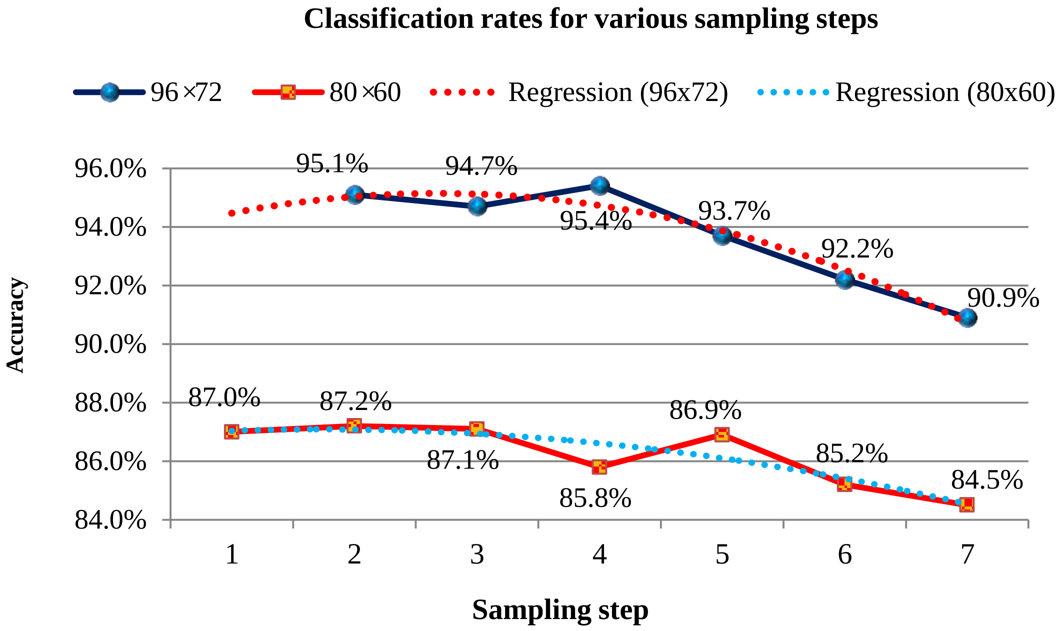


Figure 5.11 Gender classification rates for various sampling steps, tested on the FERET database. The fluctuation trends are described by quadratic polynomial regression fits, represented by dotted lines.

It can be seen from Figure 5.11 that although a smaller sampling step tends to produces a higher classification rate in general, it will only have a dramatic influence when it goes beyond 4 pixels. Therefore, a sampling step of 3 or 4 can be deemed an appropriate value without incurring high memory usage at the training stage.

4) GMM component number

Fitting a generative model, the GMM, to the features is a key step in this method. Figure 5.12 shows that the classification rate fluctuates as the number of Gaussians changes. The size of the facial images used is 128 × 96 (80% of the maximum size used in the experiments).

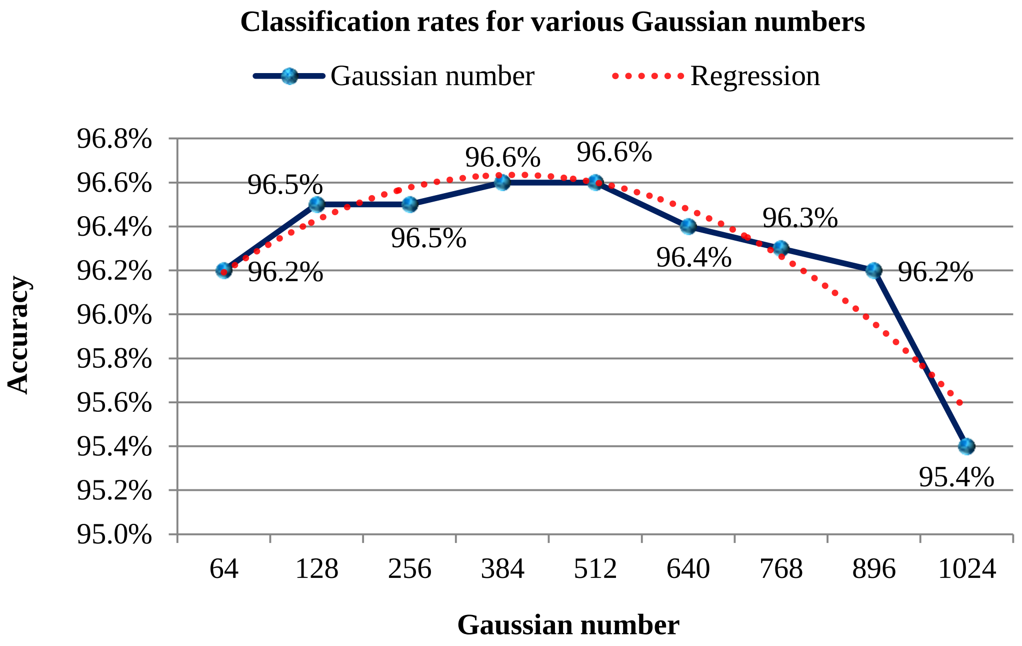


Figure 5.12: Classification rates for various Gaussian numbers, tested on the FERET database. The fluctuation trend is described by a quadratic polynomial regression fit, represented by the red dotted line.

The result agrees with the statement in S ánchez, Perronnin and Campos (2012) that an appropriate number of Gaussian components is needed since too many components result in very few ‘per Gaussian’ statistics that are pooled together, i.e. a sparse Gaussian representation. Too few Gaussians, on the other hand, are not sufficient enough to capture the uniqueness and reflect the separability of local descriptors.

5) Principal component number

As detailed in Section 5.2, PCA was implemented to reduce the dimensionality to 64 features per descriptor in the FV encoding experiments. By reducing the dimensionality of facial descriptors, highly correlated information that does not contributed to discriminability could be removed. The optimal number of principal components preserved is an empirical value as different component numbers were experimented with.

Their respective performance with regard to gender classification rates on the FERET database is shown in Figure 5.13.

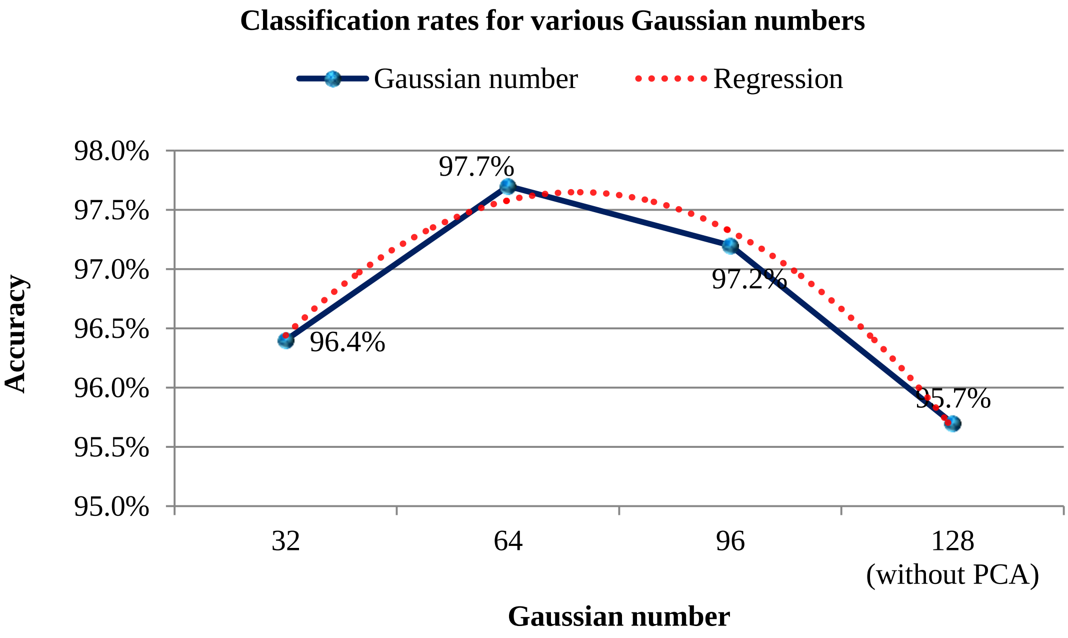


Figure 5.13 Classification rates for various principal component numbers, tested on the FERET database. The fluctuation trend is described by a quadratic polynomial regression fit, represented by the red dotted line.

As shown by Figure 5.13 that when the dimensionality of the original facial descriptors were reduced to 32, 64 and 96, the classification rates generally improved with 64 principal components yielding the highest rate.

6) Alignment, histogram equalisation and root SIFT

Face alignment has been reported in the literature to have a substantial impact on the classification rate. Without alignment, one may experience a drop in classification rate as much as 6% (Lee, Huang and Huang, 2010). However, applying histogram equalisation is a common pre-processing procedure that may increase the classification rate. When the SIFT features are specifically involved, a variation of SIFT computation, the root SIFT is recommended by Arandjelovic and Zisserman (2012), which uses the Hellinger kernel instead of the standard Euclidean distance to measure the similarity between SIFT descriptors. The impacts of the three factors are explored in this experiment and summarised in Figure 5.14.

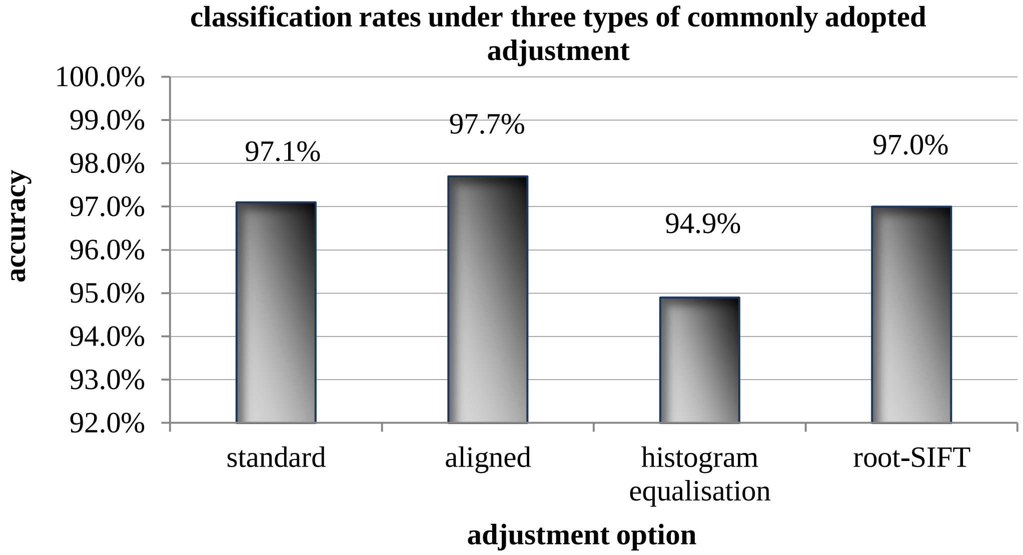


Figure 5.14: Gender classification rates with or without face alignment, histogram equalisation and root SIFT implementation.

Note that each time only one type of adjustment is made so that their impacts can be evaluated independently from other factors. As indicated by Figure 5.14, only face alignment plays a positive role in the classification results while the two other types of adjustment decrease the classification rate. It can be claimed that the algorithm is relatively robust against misalignment since misaligned face images only caused a 0.6% drop in the classification rate in the experiment.

To sum up, the optimal parametric setting is found to be 1) image size: 160×120 pixels (the largest image size experimented with for the FERET database), 2) sampling window size: 24×24 pixels, 3) sampling steps: 4 pixels, and 4) Gaussian component number: 512. Note that the images have been resized to keep an aspect ratio of 3:4 for this group of experiments and also other experiments in this study for consistency.

As SIFT features yielded the highest classification accuracy, it was further tested on the LFW database so that the FV gender classification algorithm could be evaluated on real-world scenes. The optimal parametric setting was adopted except that these facial images were resized to 112×84 due to the relatively small size of the original images, compared to the FERET database. Note that the SIFT features in this experiment were only extracted at one scale since improved accuracy was not observed when multiple scales are used. In addition, only a linear SVM was employed since the RBF kernel in the experiments did not contribute to higher classification rate.

The classification rates for aligned and misaligned images in the LFW database are 92.5% and 92.3%, respectively. The classification rate for misaligned images in the FRGCv2 database is 96.7%. Note that, for the FRGCv2 database, the SIFT features were extracted from the depth values instead of greyscale values. These results are displayed in Figure 5.15.

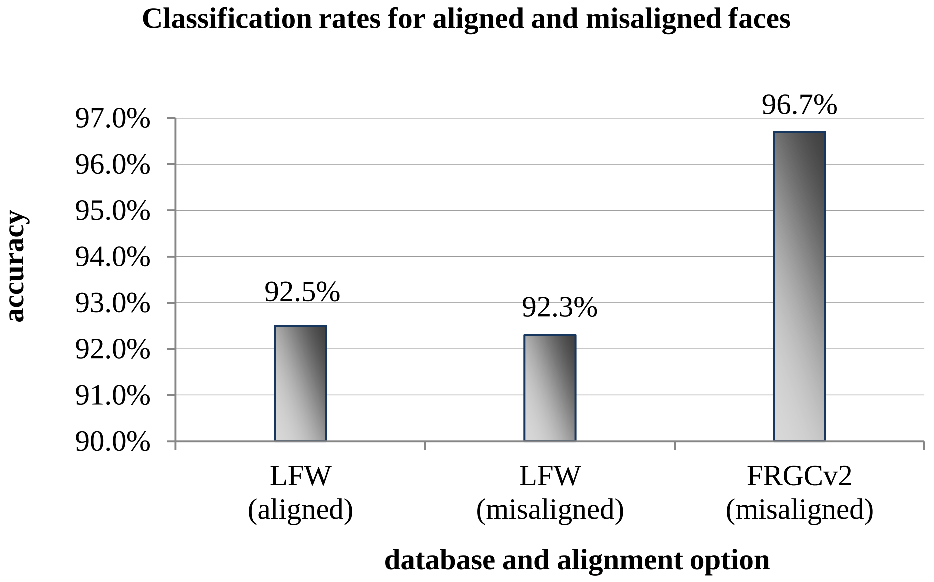


Figure 5.15 Gender classification rates for aligned and misaligned faces, tested on the LFW database and the FRGCv2 database

The endeavour to align the faces of the LFW database, in general cases, incurs a large amount of computation and sees very limited benefit; therefore it is not worthwhile. The reason for the reduced improvement (i.e. 0.2%) brought by face alignment on this database, compared to the 0.6% increase on FERET, is due to the large extent of out-of-plane head rotation which cannot be compensated for by conventional alignment algorithms. Another possible reason is that the proposed algorithm extracts features at local level and has tolerance to geometric variation. The proposed algorithm is insensitive to face alignment and therefore is robust against head rotation. This exempts the algorithm from the need for a face alignment algorithm, lowers the complexity of the algorithm, and increases its adaptability.

It has been stated in the preceding section that during the FV encoding process, spatial information of every image patch is implicitly embedded in a GMM. Therefore it is possible to restore the spatial coordinates from the Gaussian means (where the last two variables stand for the x and y coordinates). Similarly, the spatial variances can be restored from the Gaussian covariances, which indicate how well the spatial coordinates can represent individual patch locations. By visualising the Gaussians that correspond to the most discriminative image patches, it is possible to localise the facial regions that are most powerful in distinguishing male and female faces. Visualisation of facial discriminability is shown in Figure 5.16. Note that the face image displayed is only a generic representation of facial geometry and therefore is not gender-specific.

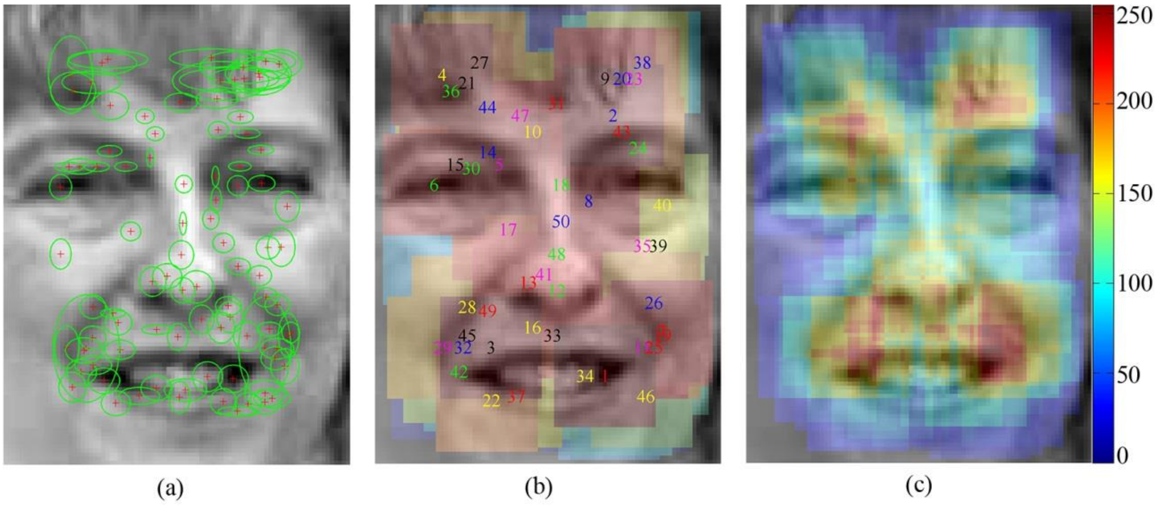


Figure 5.16 Visualisation of facial discriminability with regard to gender classification. (a) The top 128 Gaussians. (b) Facial patches at top 50 Gaussian locations. (c) The energy map for facial discriminability.

The first visualisation format (Figure 5.16(a)) shows that the most discriminative Gaussians agree with intuitive feature points (e.g. edges and corners), and therefore are deemed a suitable representation of face appearance. The Gaussians are further visualised in the form of dense image patches whose rankings in the discriminative power are indicated by the numbers centred at each patch (Figure 5.16(b)). It can be noticed that the patches overlap significantly, making it difficult to distinguish the most discriminative ones. One simple solution is to construct an energy map for all pixel locations, stacking up all the patches in the previous format. The construction of the energy map agrees with the following two rules: 1) image patches with higher rankings hold more energy; and 2) an overlapped pixel location draws energy from all the patches that cause this overlap. It can be seen that the mouth region (where male adults may have beard and moustache), the nasolabial furrows and the forehead region (where females are more likely to have fringes) are the most discriminative. This result provides insight into region-based facial discriminability so that future research on gender recognition can better target on facial regions with high discriminative power.

To further validate the proposed method, its accuracy is compared to ten other state-of-the-art methods with an evaluation of their respective limitations, detailed in Table 5.1.

Table 5.1 A comparison of the proposed method with ten other state-of-the-art gender recognition methods

Method	Description	Database	Validation method	Accuracy	Limitation
the proposed method	*FV encoding*	*FERET fa 1762*	*5-CV*	*97.7%*	*slow at training stage*
		FERET fa 1762	*50%/50%*	*96.9%*	
		FERET fa 1762	*50%/50%**	*97.9%*	
		FERET fa+fb 900 (u)	*5-CV*	*96.1%*	
		FERET fa+fb 2400	*50%/50%*	*98.3%*	
		FERET fa+fb 2400	*50%/50%**	*99.5%*	
		LFW all (u)	*5-CV*	*92.5%*	
		FRGCv2 depth all 466 subjects (u)	*5-CV*	*96.7%*	
(Rai and Khanna, 2014)	2DPCA Gabor space	LFW all	2-CV*	89.1%	*
(Xia, Amor and Daoudi, 2014)	Random Forest votes	FRGCv2 depth all 466 subjects	leave-one-out CV	97.2%	/
(Wang and Kambhamettu, 2013)	LBP, shape index	FRGCv2 depth all 466 subjects	5-CV*	93.7%	*
(Shan, 2012)	refined LBP histogram	LFW 7443 selected	5-CV	94.8%	manual data selection
(Ullah et al., 2012)	LBP & wavelet transform	FERET fa+fb 2400	50%/50%*	99.3%	manual data selection
(Lee, Huang and Huang, 2010)	facial strips + SVM	FERET fa 1763	not specified	98.8%	slow & need alignment
(M äkinen and Raisamo, 2008)	classifier fusion	FERET fa+fb 900 (u)	5-CV	92.9%	6 classifiers needed
(Phung and Bouzerdoum, 2007)	CNNs	FERET fa 1762	5-CV*	96.4%	*
(Tivive and Bouzerdoum, 2006)	CNNs	FERET fa 1762	5-CV*	97.2%	*
(Moghaddam and Yang, 2000)	RBF-SVM	FERET thumbnails 1855	5-CV*	96.6%	*

The '*' notation indicates that training data and testing data may contain different images of the same subject(s); '(u)' represents unique subjects; '5-CV' and '2-CV' represent five-fold and two-fold cross validation, respectively; '50%/50%' represents half the data for training and the other half for testing. With a number of evaluation methods implemented on different databases and database partitions, it can be seen from Table 5.1

that the proposed gender recognition algorithm outperforms most studies in comparison under the same experimental settings. It is recommended that, in the evaluation process, the five-fold cross validation technique should be adopted for strict and efficient evaluations. In addition, evaluations with different images of the same subjects in both training set and testing set should be avoided. The increase of accuracy (normally 1% or more) caused by this is most likely due to classification of gender-specific features that are already learnt by a classification model. Other validation methods listed in Table 5.1 are solely employed for a grain-to-grain comparison with other gender recognition studies.

Results show that the proposed gender recognition method outperforms all the other methods in comparison on the FERET database, expect for one study (Lee, Huang and Huang, 2010) which does not specify the evaluation method. Evaluated on a realistic database, the classification accuracy of the proposed method is only slightly lower than the study by Shan (2012), who manually removed challenging data from the LFW database. Excellent accuracy is also manifested on the FRGCv2 database containing depth facial data. The superiority of 3D facial data and an effective approach to 3D face reconstruction are presented in the subsequent chapters.

5.4 Summary

Fisher Vectors bring various benefits to object classification tasks. As well as adding high discriminability to facial features and resulting in superior classification accuracy, they offer the following advantages:

1) Fisher vectors are of uniform length (i.e. dimensionality). This allows different types of low level dense features to be encoded into feature vectors with the same dimensionality. As a result, various feature types, regardless of the source they are extracted from, can be easily manipulated, e.g. feature fusion.

2) This method is versatile in that it can encode almost any type of features. In Chapter 7, over ten different feature types are further encoded and compared.

3) At the classification stage, only a linear SVM is needed for the best accuracy. This makes the approach more efficient. The reason that SVMs with non-linear kernels do not offer additional benefits in the experiments is that “*learning a kernel classifier using the kernel is equivalent to learning a linear classifier on the Fisher vectors*” (Perronnin, Sánchez and Mensink, 2010).

4) FVs are robust to head pose and therefore a face alignment stage can be eliminated. This approach samples and encodes dense facial patches. Although features within a patch are bounded by local geometry, they are globally independent from facial geometry since individual patches are treated equally.

The FVs are of high dimensionality. Therefore, at the training stage, it requires the computer to have a large memory for data storage. This also causes a prolonged offline training time. However, computers nowadays can be easily equipped with large memories, and as soon as the classifier offline training is completed, the classification can be achieved online in real time.

The proposed method, for the first time, applied Fisher Vectors on gender recognition and achieved excellent classification accuracy on both controlled and realistic databases. Furthermore, it systematically explored different feature types and exhaustively investigated algorithmic parameters in pursuit of the maximised performance.

Chapter 6 3D Surface Reconstruction

3D facial features reveal facial topology by providing geodesic distances and surface curvatures. They have thus shown promise for bringing higher accuracy to face recognition, as well as improved robustness to practical applications where scenes are complex and dynamic. However, the exploitation of 3D vision is not currently sufficient enough to enable a wide array of 3D vision based applications. This is caused by the fact that many studies apply 3D surface reconstruction techniques to data from existing stereo imaging systems. However, most stereo imaging systems employ complex structures that limit their practical applications due to high manufacturing cost, difficulties of deployment and limited real-time performance. At the same time, most 3D surface reconstruction methods are evaluated on public databases where facial data are gathered in laboratory environments and are available in high resolution and consistency. Consequently, many algorithms struggle to find their way into real-world scenarios and others lack reconstruction accuracy.

In this chapter, a variation of the photometric stereo (PS) method is introduced and, for the first time, is applied to various types of realistic data. The aim of this endeavour is to provide a 3D reconstruction algorithm, together with a compatible stereo imaging system, suitable but not restricted to visual recognition and classification tasks. The reason for choosing the PS method over binocular vision and the Microsoft Kinect sensor for 3D surface reconstruction is based on a consideration of multiple factors including reconstruction accuracy, resolution, cost and applicability:

As reviewed in Section 2.4, the PS method reconstructs one surface normal vector per pixel, and therefore it is capable of recovering surface normals in high resolution. 3D reconstructions by PS are spatially consistent with PS images (greyscale) captured by a single camera. This eliminates the correspondence problem that perplexes binocular vision solutions, i.e. the problem of ascertaining how pixels in one image spatially correspond to those in the other image. Furthermore, the resolution of PS reconstructions can be flexible, which is solely determined by the camera employed. This allows a PS setting to cater to a specific device or application. In contrast, data obtained by the Kinect are normally of low spatial and depth resolution, which severely degrade as the sensor-object distance increases. In addition, PS reconstructions boast detailed high-frequency 3D texture information. 3D depth information can be derived from surface normals when necessary. In contrast, binocular stereo is more prone to noise and artefacts, since it directly recovers depth of surface (i.e. image centred) rather than surface orientations (i.e. object centred). Although being highly accurate and of high resolution, PS devices can be constructed at a similar or lower cost to the Kinect, with the potential flexibility of being portable or long-range (Hales et al., 2015), and is thus a more powerful solution to 3D imaging.

As reviewed in Section 2.4, the theories of PS have been increasingly sophisticated. Furthermore, studies in the recent decade have justified the superiority of surface normals over other 3D features (e.g. cloud points and depth maps) as face representations. This is one of the motivations for the PS method being utilised by this study where 3D face representations play a significant role in the proposed HCI strategy. Its contribution to facial landmark detection has been exemplified in subsection 4.4.3, which could lead to robust head pose estimation and eye fixation analysis.

While this chapter introduces the theories of PS and its two-source variations, as well as evaluating them on a publicly available database, the subsequent chapter further applies the proposed two-source PS method to a self-collected dataset and thus validates its contribution to 3D gender and age recognition by evaluating up to ten types of 2D and 3D features.

6.1 Photometric Stereo Principles

As reviewed in Chapter 2, PS allows estimation of surface normals from reflectance maps obtained from images of the same object captured under different illumination directions. It was first introduced by Woodham (1980) whose work illustrates that three views are sufficient to uniquely determine the surface normals as well as albedos at each image point, provided that the directions of incident illumination are not collinear in azimuth. Other works employ four views for improved reconstruction performance.

Let $I_1(x, y), I_2(x, y)$ and $I_3(x, y)$ be the three images captured under varied illumination directions. By varying the illumination direction, the reflectance map is changed accordingly, giving equation (6.1).

$$\begin{cases} I_1(x, y) = R_1(p, q) \\ I_2(x, y) = R_2(p, q) \\ I_3(x, y) = R_3(p, q) \end{cases} \tag{6.1}$$

where $R_1(p, q)$, $R_2(p, q)$ and $R_3(p, q)$ are the reflectance maps under different illumination directions, while p and q are gradients of the surface in the x and y directions, respectively. A general reflectance map in the gradient representation of the surface orientation and illumination direction is expressed in equation (6.2) (Woodham, 1980).

$$R(p, q) = \frac{\varrho(1 + pp_s + qq_s)}{\sqrt{1 + p^2 + q^2}\sqrt{1 + {p_s}^2 + {q_s}^2}} \tag{6.2}$$

where ϱ is the albedo, $\vec{\boldsymbol{N}} = [-p, -q, 1]$ defines the surface normal vector, and $\vec{\boldsymbol{L}} = [-p_s, -q_s, 1]$ defines the illumination direction. Let the surface be $z = f(x, y)$, the gradients in x and y directions become:

$$\begin{cases} p = -\dfrac{\partial f(x,y)}{\partial x} \\ q = -\dfrac{\partial f(x,y)}{\partial y} \end{cases} \tag{6.3}$$

These equations are derived under the assumptions that 1) the object size is small relative to the viewing distance. 2) The surface is Lambertian. 3) The surface is exempt from cast-shadows or self-shadows.

To simplify the expression, the light vector is further normalised as a unit vector $\overrightarrow{\boldsymbol{L_n}} = [a_x, a_y, a_z]$. The relationship between the intensity image and the reflectance map can also be written as:

$$I(x,y) = \frac{\varrho \cdot < \overrightarrow{\boldsymbol{N}}, \overrightarrow{\boldsymbol{L_n}} >}{|\overrightarrow{\boldsymbol{N}}|} = \varrho \cdot \frac{-pa_x - qa_y + a_z}{\sqrt{1 + p^2 + q^2}} \tag{6.4}$$

From equation (6.1) – (6.4), it is known that with three greyscale images $I_1(x,y)$, $I_2(x,y)$ and $I_3(x,y)$, along with three known light vectors $\overrightarrow{\boldsymbol{L_{n1}}}$, $\overrightarrow{\boldsymbol{L_{n2}}}$ and $\overrightarrow{\boldsymbol{L_{n3}}}$ pointing in the directions of their respective light source, the surface normal and albedo at each image point can be uniquely determined.

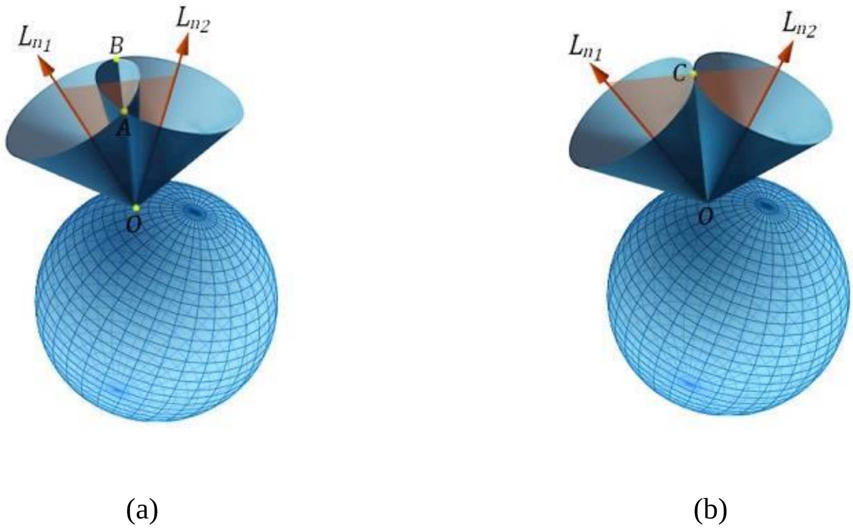
6.2 A Two-source PS Method for Face Reconstruction in Unconstrained Environments

Although the standard four-source PS method is highly accurate and relatively efficient in recovering 3D surface normals, it is however not ideal for facilitating real-world applications. Generally, its implementation is prohibited by the need for capturing at least 3 (or more commonly 4) images at high frame rate for every reconstruction. Another limitation is posed by the complex structure of the data capture system where a large set of light sources need to be deployed, which are likely to lead to a bulky and hazardous system in some cases.

Mecca and Durou (2011) combined a differential and a non-differential approach to predict the number of surface normal solutions with only two PS images. They achieved this by characterising the zones on a surface where unique surface normal solutions could be admitted. However, their work was only tested on simulated data. The authors also concluded that “more tests have to be performed in order to more clearly show the accuracy of our approach, since this work was rather theoretical”.

To conduct theoretical studies of two-source PS from an applied point of view, this section proposes to simplify both the data capture as well as the hardware design by employing a two-source PS variation where only two light sources are required and therefore only two images need to be captured. The proposed two-source PS algorithm and its compatible hardware system are intended for use in real-world environments.

In general, the equations in (6.1) are nonlinear with respect to p and q, and therefore any two of them admit more than one solution. This ambiguity problem is interpreted by Figure 6.1.



(a)

(b)

Figure 6.1: Illustration of ambiguous solutions from two images using PS. (a) two-source PS with ambiguous solutions at an image point. (b) two-source PS with a unique solution at an image point.

When only one light source is concerned, it can be seen from equation (6.4) that the surface normal vectors which produce a specific intensity value at point O form a cone with apex at this point and axis in the direction of illumination $\overrightarrow{\boldsymbol{L_{n_1}}}$. In the case of two illuminators, the surface normals should belong to two such cones and therefore exist at the intersections of the two cones. Two cones with the same apex either have two intersections or one intersection (the case of no intersection does not occur for PS images). These two scenarios corresponding to ambiguous solutions and a unique solution respectively are shown in Figure 6.1 (a) and Figure 6.1 (b) where $\overrightarrow{OA}$ and $\overrightarrow{OB}$ represent the ambiguous solutions and $\overrightarrow{OC}$ represent the unique solution.

This can be mathematically explained by deriving a pair of equations as in the general form of equation (6.4). If constant albedo is assumed for simplicity, and let $\overrightarrow{\boldsymbol{L_{n_1}}} = [a_x, a_y, a_z]$, $\overrightarrow{\boldsymbol{L_{n_2}}} = [b_x, b_y, b_z]$, two PS images yield

$$\begin{cases} I_1 = \dfrac{-pa_x - qa_y + a_z}{\sqrt{1+p^2+q^2}} \\ I_2 = \dfrac{-pb_x - qb_y + b_z}{\sqrt{1+p^2+q^2}} \end{cases} \tag{6.5}$$

The solutions for p and q are produced in a similar way to Onn and Bruckstein (1990). Let

$$T \triangleq \sqrt{1+p^2+q^2} \tag{6.6}$$

Rearranging equation (6.5) yields

$$\begin{cases} pa_x + qa_y = a_z - I_1 T \\ pb_x + qb_y = b_z - I_2 T \end{cases} \tag{6.7}$$

Solving equation (6.7) for p and q in terms of T produces equations of the form (Onn and Bruckstein, 1990)

$$\begin{cases} p = \varepsilon_1 T + \varepsilon_2 \\ q = \varepsilon_3 T + \varepsilon_4 \end{cases} \tag{6.8}$$

where ε_1, ε_2, ε_3, ε_4 are functions of known values I_1, I_2, a_x, a_y, a_z, b_x, b_y and b_z. Combining equation (6.6) and (6.8) provides a quadratic equation for T of the form

$$\lambda_2 T^2 + \lambda_1 T + \lambda_0 = 0 \tag{6.9}$$

where

$$\begin{cases} \lambda_2 = \dfrac{(I_2 a_y - I_1 b_y)^2 + (I_1 b_x - I_2 a_x)^2 - (a_x b_y - b_x a_y)^2}{(a_x b_y - b_x a_y)^2} \\ \lambda_1 = \dfrac{2\big(a_z b_y - b_z a_y\big)\big(I_2 a_y - I_1 b_y\big) + 2(a_x b_z - a_z b_x)(I_1 b_x - I_2 a_x)}{(a_x b_y - b_x a_y)^2} \\ \lambda_0 = \dfrac{(a_z b_y - b_z a_y)^2 + (a_x b_z - a_z b_x)^2 - (a_x b_y - b_x a_y)^2}{(a_x b_y - b_x a_y)^2} \end{cases} \tag{6.10}$$

Solving this quadratic equation ($\lambda_2 \neq 0$ in this case) gives

$$T_{1,2} = \frac{-\lambda_1 \pm \sqrt{{\lambda_1}^2 - 4\lambda_2\lambda_0}}{2\lambda_2} \tag{6.11}$$

where $\lambda_2 \neq 0$ and where $\sqrt{{\lambda_1}^2 - 4\lambda_2\lambda_0}$ is the discriminant of the quadratic equation. The two pairs of derivatives then become

$$\begin{cases} p_{1,2} = \dfrac{a_z b_y - b_z a_y + I_2 T_{1,2} a_y - I_1 T_{1,2} b_y}{a_x b_y - b_x a_y} \\ q_{1,2} = \dfrac{a_x b_z - a_z b_x + I_1 T_{1,2} b_x - I_2 T_{1,2} a_x}{a_x b_y - b_x a_y} \end{cases} \tag{6.12}$$

It should be noted that $a_x b_y - b_x a_y \neq 0$, in order to produce meaningful solutions. This practically sets a constraint to the deployment of the two illuminators that they cannot be placed diagonally with respect to the object to be illuminated. An example of diagonal lighting is illustrated in Figure 6.2 where the two LEDs are placed diagonally on the $x-y$ plane, with respect to the object position.

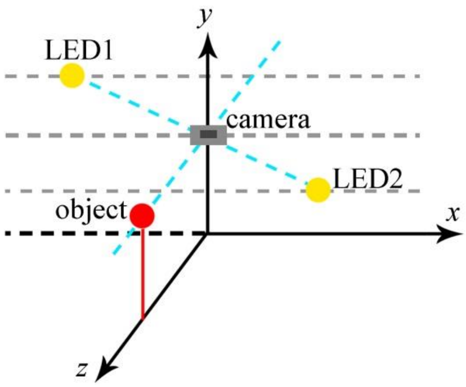
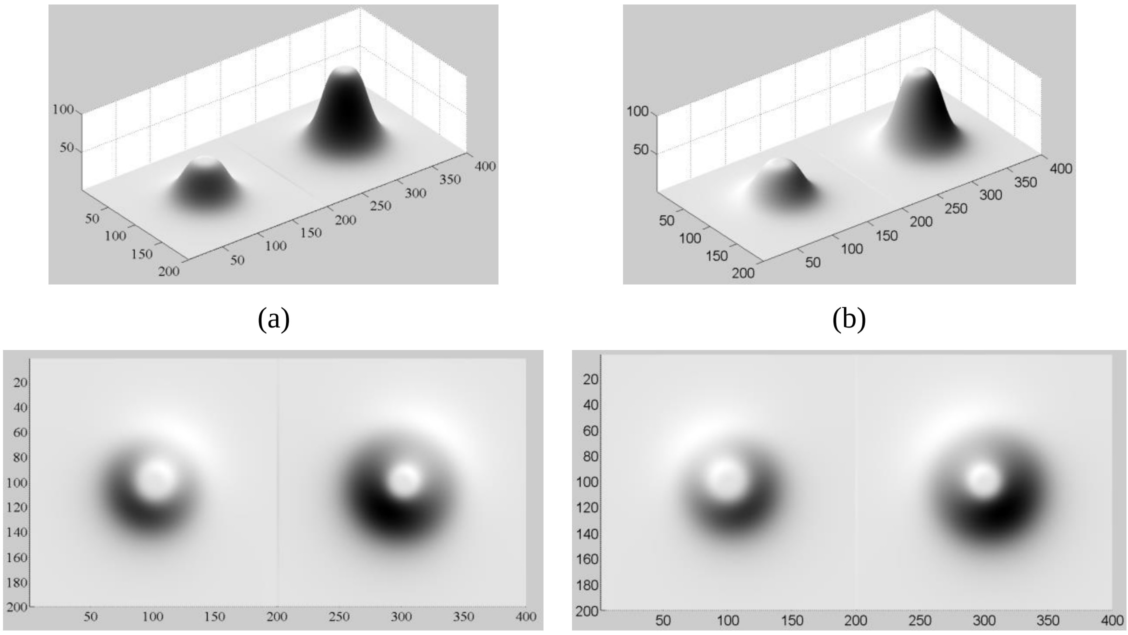


Figure 6.2 An example of diagonal lighting

To remove the ambiguity, Onn and Bruckstein (1990) enforced integrability and continuity properties of the surface of the scene, assuming that the surface normals are continuous and that the surface height is twice differentiable. According to the continuity property of the surface, this study suggests that an arbitrary surface can be divided into connected regions R_c $(c \in \mathbb{Z})$, where there exists either a unique solution for T or a pair of solutions. On the other hand, integrability provides equation (6.13)

$$\int_{(x,y)\in R_c} \left(\frac{\partial p_{1,2}}{\partial y} - \frac{\partial q_{1,2}}{\partial x}\right)^2 = 0 \tag{6.13}$$

The pairs of p and q that agree with this equation are the true gradients and therefore correspond to the true surface normals. It can be seen from Figure 6.1 that when there exists a unique solution, the surface normal lies on the plane defined by the two lighting vectors (illumination directions). In the case of two ambiguous solutions, the surface normals lie on both sides of the plane. Equation (6.13) can be therefore used to discard the false solution. An example on synthetic images is presented to demonstrate the process of the two-source PS algorithm. Figure 6.3 shows two 3D Gaussian surfaces illuminated from two different directions with unit lighting vectors $\overrightarrow{\boldsymbol{L_{n_1}}} = [0.2592, -0.4319, 0.8639]$ and $\overrightarrow{\boldsymbol{L_{n_2}}} = [-0.2592, -0.4319, 0.8639]$. The two vectors represent symmetric illuminations with regard to the y axis, which can be observed from Figure 6.3.



(a)

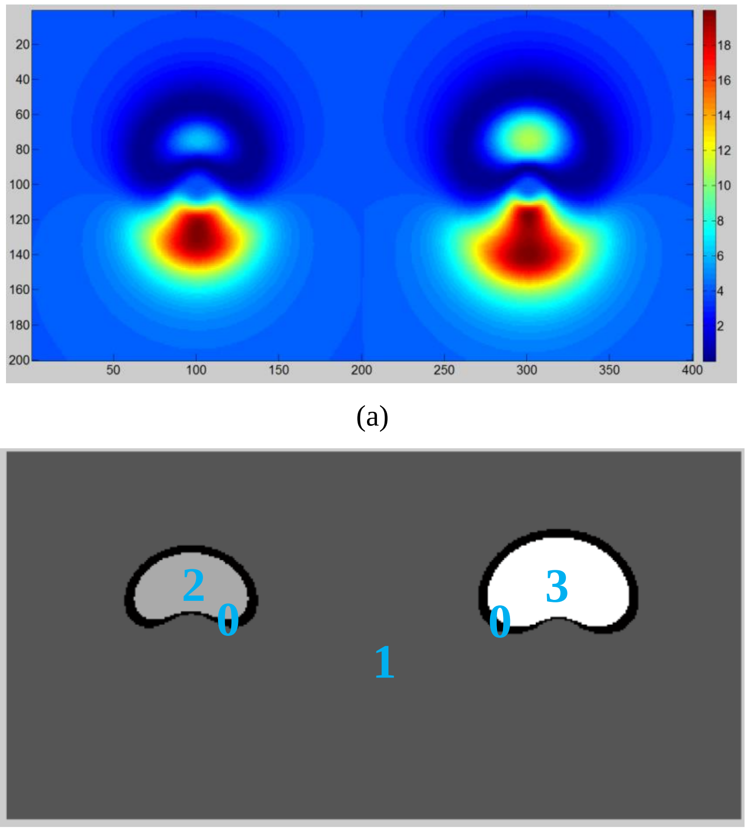
(b)

(c)

(d)

Figure 6.3: Two Gaussian surfaces illuminated from two different directions. (a) and (b) are illuminated surfaces in 3D view with the lighting vector L_{n_1} and L_{n_2}, respectively. (c) and (d) are illuminated surfaces in 2D view with the lighting vector L_{n_1} and L_{n_2}, respectively.

The regions on the whole surface where there exists a unique solution can be found by correlating them with $\sqrt{{\lambda_1}^2 - 4\lambda_2\lambda_0} = 0$ (this is when $T_1 = T_2$ in equation (6.11)). Naturally these regions define the boundaries that segment out the other regions with ambiguous solutions. The resulting image where pixel values are calculated by $\sqrt{{\lambda_1}^2 - 4\lambda_2\lambda_0}$ is referred to as the discriminant image shown in Figure 6.4 (a).



(a)

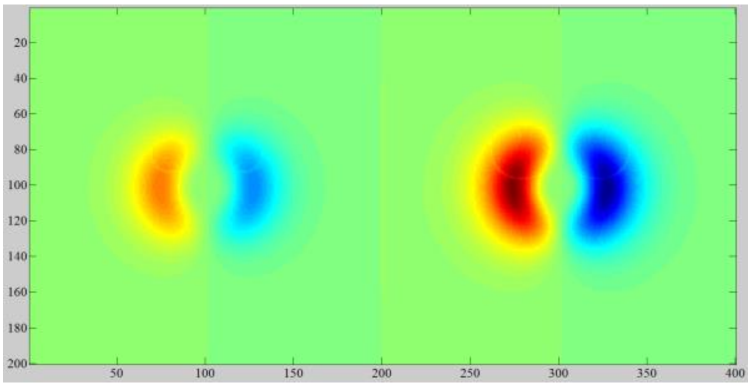
(b)

Figure 6.4: The discriminant image and the label image of the synthetic surface. (a) The discriminant image where the dark blue regions are where unique solutions exist for equation (6.11). (b) The corresponding label image that segments the surface into a number of connected regions, the number being 4 in this example, labelled by '0', '1', '2' and '3', respectively.

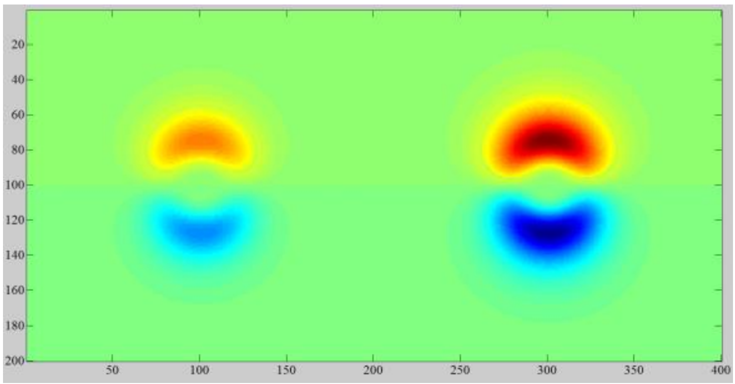
As the colour map shows, the dark blue regions are where the discriminant is close to zero. A threshold of 0.1 is used in practise to replace the threshold of 0 and accounts for calculation errors so that the boundary positions generated from the discriminant aggregate a connected region rather than discrete points. All the values below the threshold in the discriminant image are set to zero and otherwise to one as a binarization process. A label image is then calculated to uniquely number the segmented regions (where all the regions with zero values are labelled by '0'), as shown in Figure 6.4 (b).

When one or two solutions for gradients in the x and y directions are found by equation

(6.12) for each region, the restriction set by equation (6.13) effectively removes the ambiguity for those with two solutions. Finally the gradients are uniquely determined that yield surface normal vectors. These gradients are shown in Figure 6.5. The surface depth is obtained according to the algorithm introduced by Frankot and Chellappa (1988). The surface depth image is displayed in Figure 6.6 (a), with its ground truth displayed in Figure 6.6 (b) as a comparison.

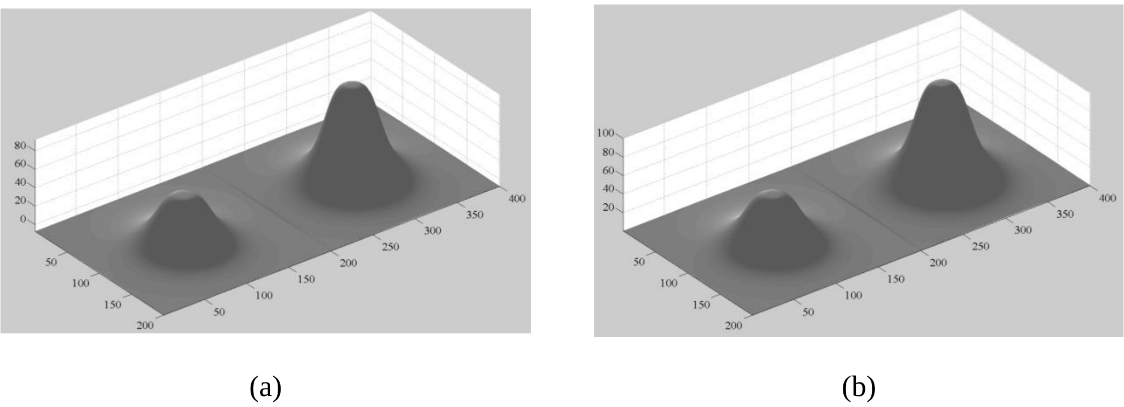


(a)



(b)

Figure 6.5: Gradients in the x and y directions with ambiguity removed. (a) Gradients in the x direction. (b) Gradients in the y direction.



(a)

(b)

Figure 6.6: The reconstructed surface and its ground truth. (a) The reconstructed surface. (b) The ground truth of the same surface.

It can be seen from the reconstruction results that the surface shape is preserved although the scaling is slightly different along the depth axis. This is a common problem caused by the integral error in the process of depth recovery, but this can be compensated by multiplication with a scalar which can be found heuristically. While the qualitative evaluation reflects accurate reconstruction results, quantitative evaluations further reveal the error between the reconstruction results and the ground truth. Specifically, the ground truth values are subtracted from the reconstructed surface normals (shown in Figure 6.7), as well as from the depth image (shown in Figure 6.8).

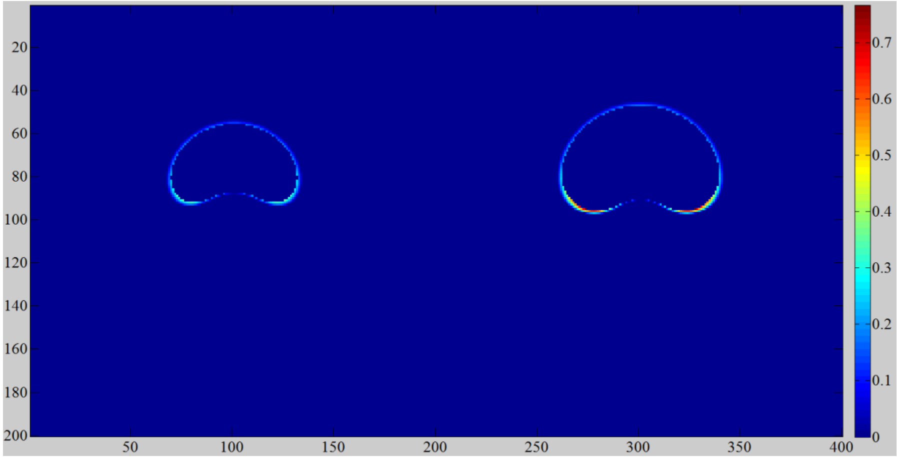
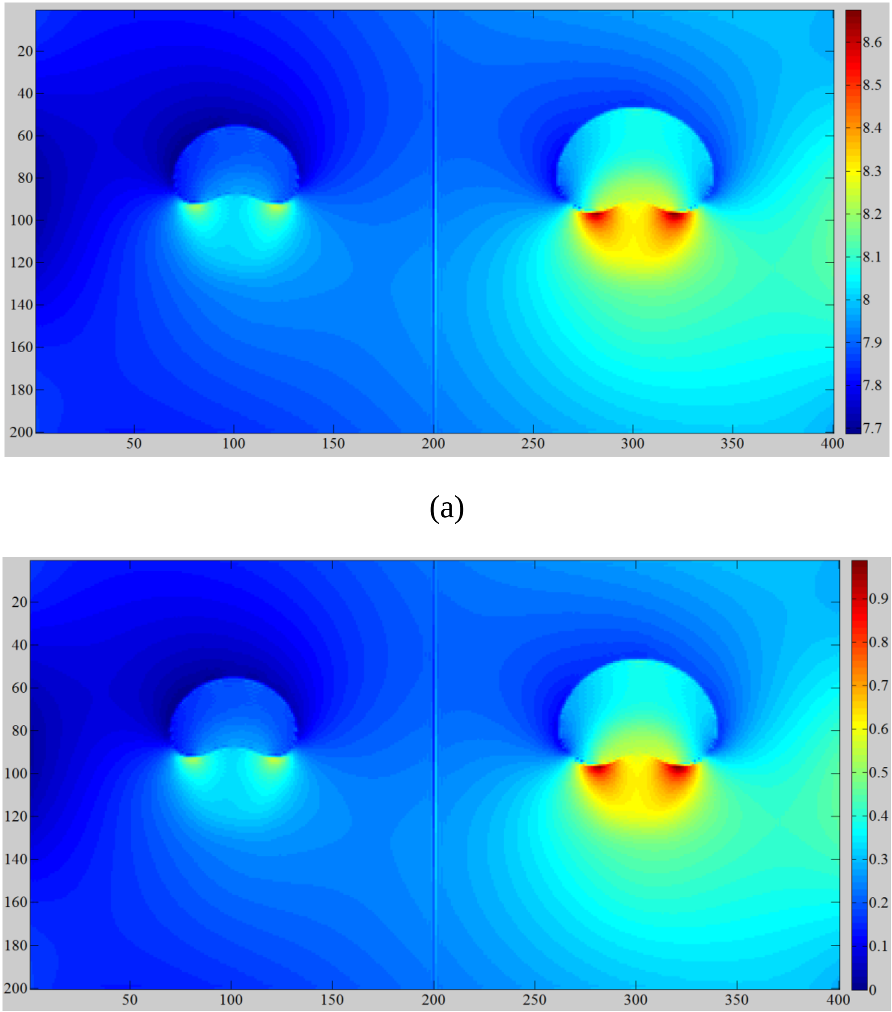


Figure 6.7: A visualisation of the computed error between the surface normals and the ground truth.



(a)

(b)

Figure 6.8: A visualisation of the computed error between the surface depth and the ground truth. (a) The original error image. (b) The error image with the minimum subtracted.

Let $\boldsymbol{g_r}(i,j) = \left[g_{rx}(i,j), g_{ry}(i,j)\right]$ be the recovered surface gradients at pixel location (i,j), and $\boldsymbol{g_t}(i,j) = \left[g_{tx}(i,j), g_{ty}(i,j)\right]$ be its ground truth, the average error for the surface normals is calculated following equation (6.14).

$$e_{\vec{N}} = \frac{1}{m \cdot n} \sum_{j=1}^{n} \sum_{i=1}^{m} ||\boldsymbol{g_r}(i,j) - \boldsymbol{g_t}(i,j)|| \tag{6.14}$$

where m and n are the image width and height. As a result, the average error for the surface normals is 0.0015, which is considerably smaller than the error for the surface depth. The former reflects the accurate reconstruction performance while the latter is caused by the scaling of the surface height in the depth recovery algorithm. With a subtraction of the minimum error, the range of error decreases while the pixel distribution remains unchanged.

The reason for employing synthetic data is that the ground truth can be easily computed for comparison. Section 6.4 further evaluates the two-source PS method using real data so that the impacts of complex shapes, image noise, shadows, specularities and experimental errors (lighting vectors, for example) can be assessed.

Although the standard PS method is capable of performing surface normal reconstructions in real time, reduced image capture time as well as decreased reconstruction time can further contribute to the efficiency and applicability of 3D imaging systems. As far as the proposed method is concerned, in the image acquisition stage, only two images are needed, as opposed to a minimum of three, or in most cases, 4. It effectively halves the time for exposure and storage during this stage. On the other hand, processing only two images involves much less calculation which further reduces the time consumed by the reconstruction algorithm. As for hardware configuration, only two illuminators are required, which lowers the complexity of the overall system and boosts the flexibility and applicability in installing and deploying systems for practical needs.

6.3 An alternative two-source PS method using diagonal lighting

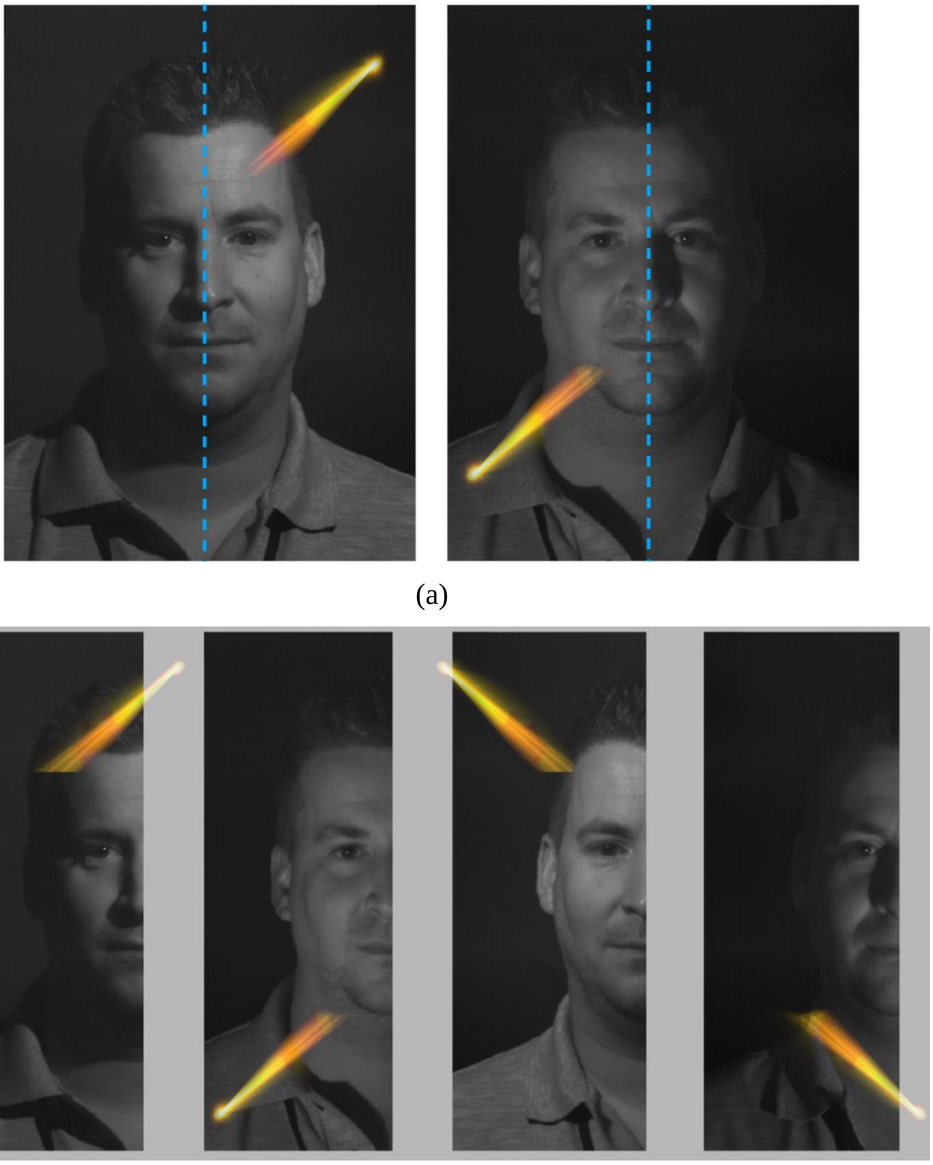
Reducing the number of illumination sources can give rise to PS systems with simplified structures. This is an essential requirement in practice since the deployment of a bulky system is often unmanageable due to limited space and safety issues. Higher flexibility of a system structure can enhance the applicability of a PS system. One limitation of the two-source PS method introduced in the preceding subsection is that the two illumination directions cannot be diagonal. Therefore, to increase the flexibility of a two-source PS system, as well as for the completeness of the two-source PS study, another two-source PS method is proposed to fulfil the reconstruction task in an alternative way.

A frontal face can be commonly considered symmetric. The symmetry property provides information that can facilitate the realisation of an alternative two-source PS method. Different from Zhao and Chellappa (2000) which employs the symmetry to reduce the unknowns in the SFS algorithm, the proposed method brings symmetry to the PS context and converts a two-source scenario to a standard four-source problem.

To calibrate facial symmetry, face alignment is required as the first step. The eye centre localisation algorithm introduced in Chapter 4 plays a key role in the alignment process. After two eye centre positions are detected in a facial image, the pupillary vector is determined to rotate the face such that the symmetry line lies vertically across the midpoint of the pupillary segment. As a result, when aligned to a frontal position, a face can be divided into the left half and the right half.

In the case of diagonal lighting with lighting vectors $\boldsymbol{L}_{\boldsymbol{n}_1} = (a, b, c)$ and $\boldsymbol{L}_{\boldsymbol{n}_2} = (-a, -b, c)$**,** two PS images I_1 and I_2 are obtained and aligned in the first place, corresponding to the two lighting vectors respectively. The left halves of the two images remain unchanged, preserving the lighting vectors $\boldsymbol{L}_{\boldsymbol{n}_1}$ and $\boldsymbol{L}_{\boldsymbol{n}_2}$. The right halves of the two images are flipped around the symmetry line, resulting in two mirrored halves. This causes the original two lighting vectors to be flipped left to right, becoming $\boldsymbol{L}_{\boldsymbol{n}_3} = (-a, b, c)$ and $\boldsymbol{L}_{\boldsymbol{n}_4} = (a, -b, c)$**,** corresponding to the mirrored image halves. Thus four image halves with four different lighting vectors are obtained, sufficient enough for the standard PS method.

This transformation from a two-source problem to a four-source problem is illustrated in Figure 6.9.

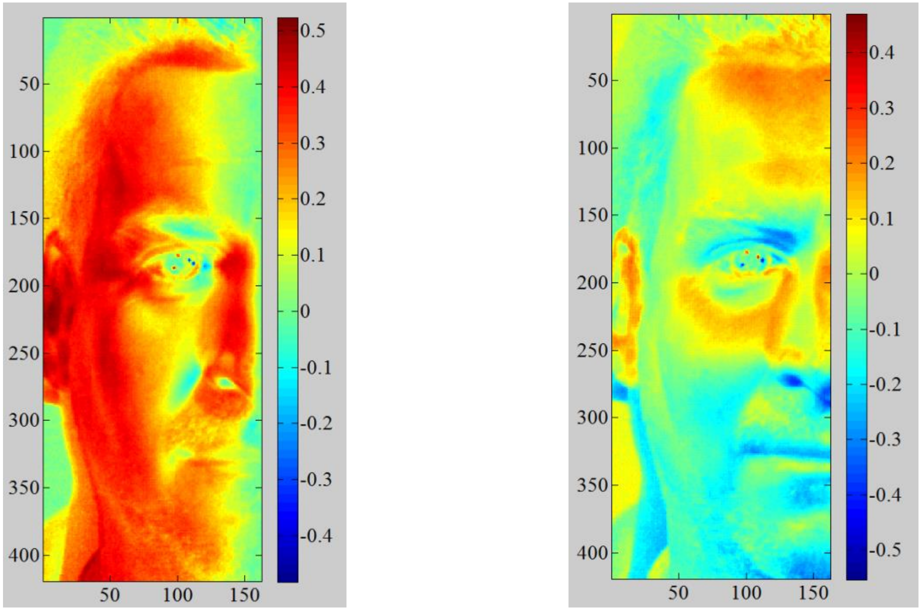


(a)

(b)

Figure 6.9: The transformation from a two-source PS problem to a four-source PS problem. The light directions are indicated by yellow light beams. (a) Two raw PS images under diagonal lighting. The symmetry of the face is indicated by blue dashed lines. (b) The four facial halves obtained by flipping right to left the two right halves. The first two columns are the original halves of the face and the last two columns are the flipped halves of the right face.

The standard four-source PS can then be employed to reconstruct half of a face using the four halves of the face. The reconstructed gradient images and the depth image are shown in Figure 6.10 and Figure 6.11, respectively.



(a)

(b)

Figure 6.10: The gradient images reconstructed by the two-source PS method using facial symmetry. (a) Gradients in the *x* direction. (a) Gradients in the *y* direction.

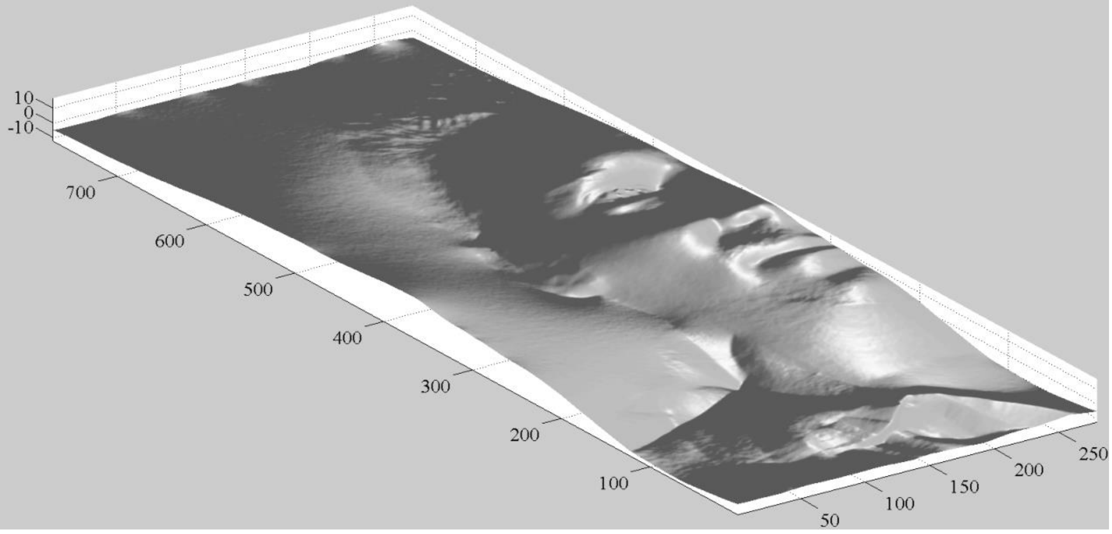


Figure 6.11: The depth image for half of the face.

It can be seen from the reconstruction result that the half of the face is as detailed as that from the standard PS method. Although only half of a face is available, it is intuitive that it contains almost as much information as a full face under the symmetry assumption.

6.4 3D Face Reconstruction Results

Under non-diagonal lighting conditions, the two-source PS method introduced in Section 6.2 has been evaluated on synthetic images. This subsection further evaluates this method by applying it to a series of image sets from the Photoface database (Zafeiriou et al., 2013). The reconstruction accuracy is then compared to that of the standard (four-source) PS method by calculating the ℓ_2-norm for reconstructed surface normals and the root-mean-square (RMS) error for surface depth.

The Photoface database is one of the few databases containing images captured under PS settings. It is deemed a suitable representation for realistic data as the data capture device was placed at the entrance to a workplace to ensure casual usage. The variations in this database include gender, age, expression, image blur, head pose and facial accessories.

An example of 3D face reconstruction is shown by applying the two-source PS method to the image set of the first subject (Figure 6.12) in the Photoface database.

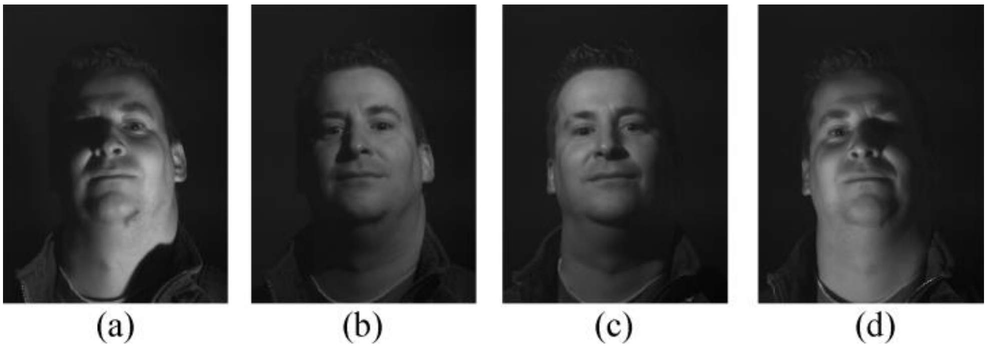


Figure 6.12 A sample PS image set from the Photoface database

Note that the proposed two-source PS method only employs the two images illuminated from the top (Figure 6.12 (b) and (c)) while the standard PS method needs all the four images for reconstruction. It has been experimentally observed that illuminating from the top-left and the top-right directions creates relatively less self-cast shadows. The recovered x gradient image and the depth image are compared with those from the standard PS in Figure 6.13. The recovery of the depth images in this thesis is based on the algorithm introduced by Frankot and Chellappa (1988).

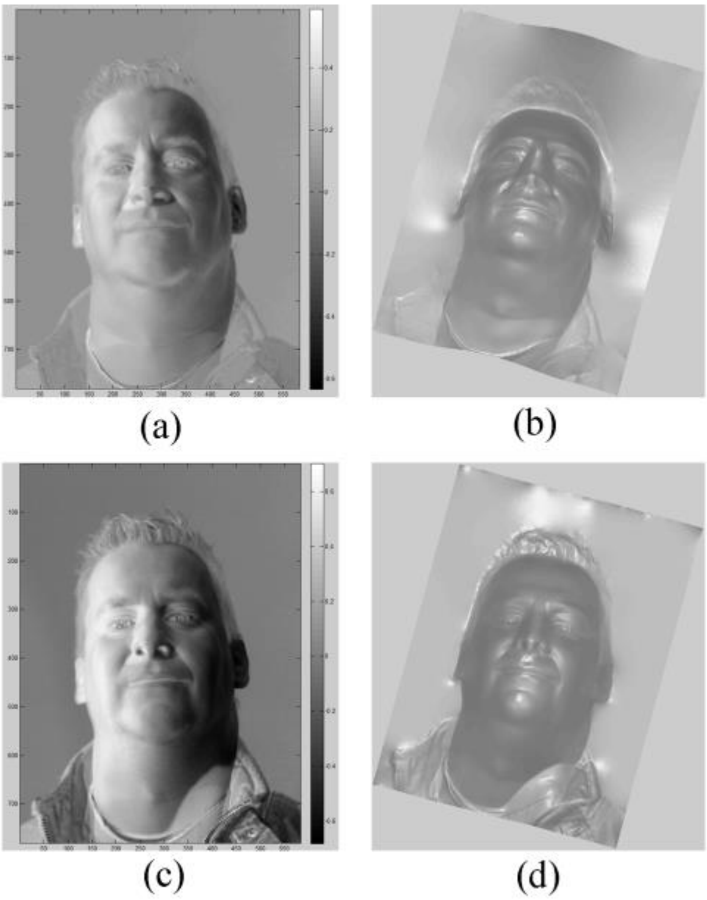


Figure 6.13 A comparison of 3D face reconstruction between the two-source PS and standard four-source PS. (a) and (b) are the x gradient image and the depth image from the two-source PS while (c) and (d) are from the four-source PS.

The difference between a pair of reconstructions by the two-source and four-source PS method is more likely to be seen in facial regions that are less continuous (e.g. eye regions). Possible causes include sharp changes in face depth, non-Lambertian reflectance and shadows. Overall, comparable reconstruction results can be reflected by this example in this particular visualisation form. Figure 6.14 further provides a comparison of the reconstructed depth images for 4 other subjects (subject 1002 to 1005) from the Photoface database.

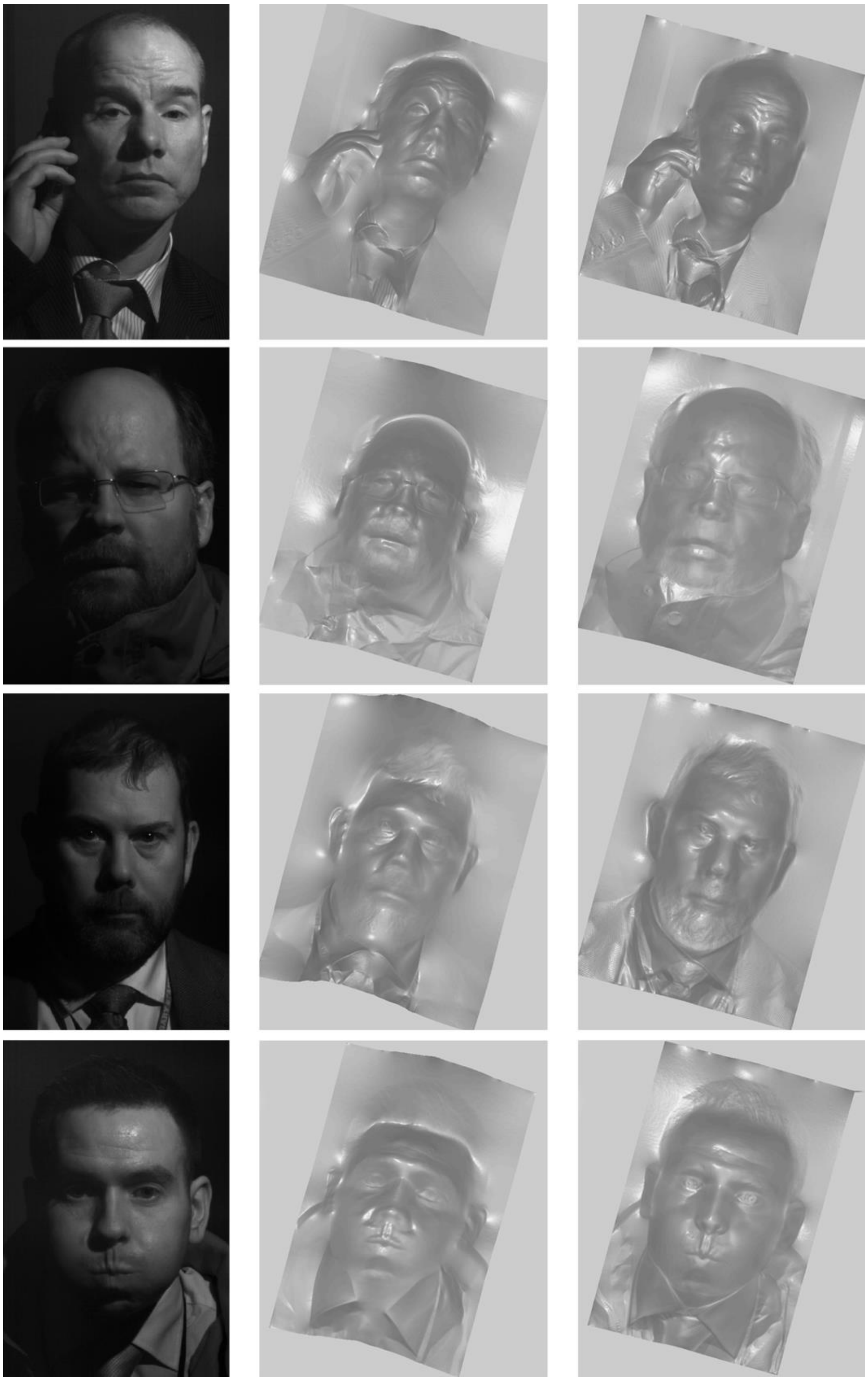


Figure 6.14 3D face reconstructions for four subjects from the Photoface database using the two-source PS and the standard (four-source) PS methods. Left to right: one of the PS images, reconstructions by the two-source PS and reconstructions by the standard PS

While the x gradient images and the depth images offer an intuitive visual comparison, a quantitative and comprehensive statistical analysis follows, which evaluates the surface normals and the depth images by measuring the ℓ_2-norm and the RMS errors for the first 100 subjects in the Photoface database. As Hansen et al. (2010) have already calculated the errors between the standard PS and the ground truth obtained by a 3dMD projected pattern range finder (3dMD, n.d.), to set the standard PS method as a reference then becomes a viable solution to evaluating the proposed two-source PS variation. Similarly to their evaluation method, all reconstructions are cropped to 160×200 facial regions centred on the nose tip in order that the evaluations are consistent.

While Hansen et al. (2010) measured the ℓ_2-norm and the RMS errors for 8 subjects from the Photoface database, this validation experiment calculated the errors for the first 100 subjects for a more objective evaluation. These results can be seen in Figure 6.15.

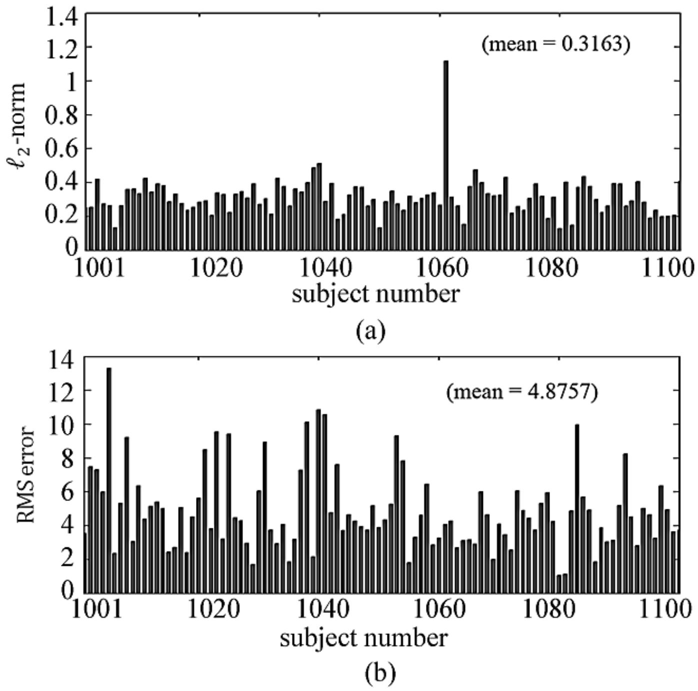


Figure 6.15 Estimation of 3D reconstruction errors. (a) The ℓ_2-norm errors for surface normals and (b) the RMS errors for surface depth (in pixels), for the first 100 subjects in the Photoface database, when the two-source PS method and the four-source PS method are used.

Note that the ℓ_2-norm for subject No. 1061 is relatively large due to the extreme head pose and the large background area.

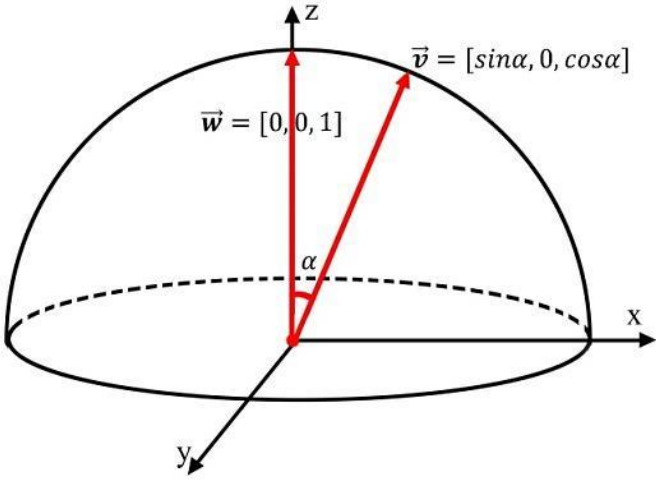


Figure 6.16 An illustration of surface normal deviation

The average ℓ_2-norm and the RMS errors for the 100 subjects are 0.3163 and 4.8757, respectively. Consider a Cartesian coordinate system with x, y and z axes (see Figure 6.16), a ℓ_2-norm error is caused by a unit vector deviating in the x and/or y directions. This is similarly to angular deviations in a spherical coordinate system represented by radial distance, azimuthal angle, and polar angle. ℓ_2-norm error of 0.3163 corresponds to an error of only 5.74 degrees when a unit surface normal vector $[x, y, z]$ deviates in either the x direction or the y direction. The correspondence between angular deviation and ℓ_2-norm error in a 2D Cartesian coordinate system can be found in Figure 6.17.

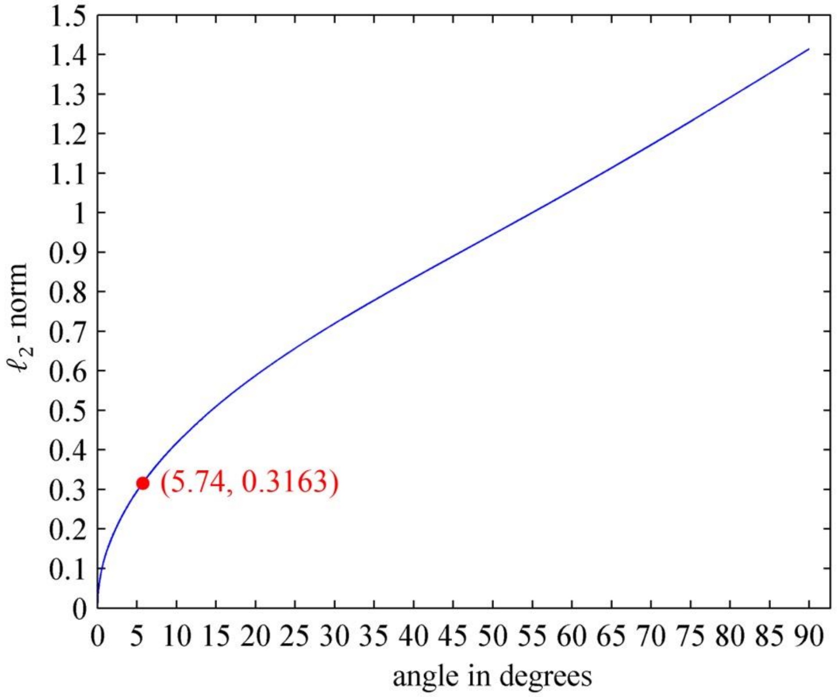


Figure 6.17 Correspondence between angular deviation and ℓ_2-norm error in a 2D Cartesian coordinate system.

On the other hand, RMS error of 4.8757 pixels corresponds to only 1.12% of the average length of the 100 face images (i.e. 420 pixels). Therefore, both the 3D face visualisation and the statistical study can validate that, when realistic data are concerned, the two-source PS method has achieved comparable results to those from the standard PS method, meaning that the accurate 3D reconstructions it generates should promise to facilitate different classification tasks similarly to the standard PS method.

The times spent for the first 100 reconstructions were also recorded in the 3D face reconstruction experiments. The computational times for the two-source PS and the four-source PS are plotted in Figure 6.18 for a comparison.

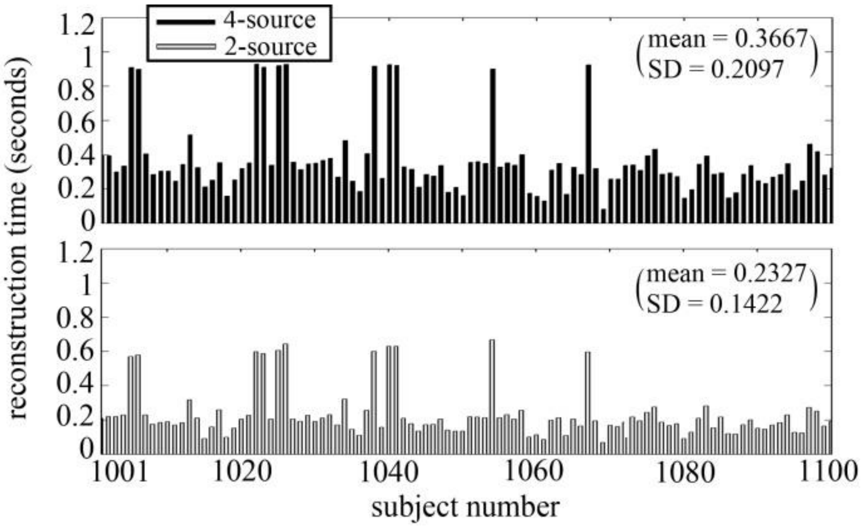


Figure 6.18 An evaluation of 3D face reconstruction efficiency. Note that the times include both surface normal recovery and depth estimation. The mean and the standard deviation (SD) of reconstruction times are calculated for both the two-source and the four-source PS method.

This test was performed with Matlab R2014a and on a computer with an Inter(R) Core(TM) i5-4570 CPU and 12GB memory. It can be seen from Figure 6.18 that, on average, the two-source PS method only consumes less than two thirds of computational time required by the standard PS in the 3D reconstruction stage. Note that the reconstruction times for the 100 subjects vary mainly due to their different image sizes (average size: 733×624 pixels).

Moreover, in the image capture stage, the two-source PS method halves the time consumption by capturing only two images per reconstruction instead of the commonly required 4 images. The standard PS has been able to spawn real-time 3D imaging systems (Malzbender et al., 2006), and therefore the two-source PS variation should promise to boost the efficiency of 3D imaging systems to a higher level.

Overall, the reduced image capture time and the shortened 3D reconstruction time are the two main factors that can result in high efficiency of the two-source PS method and therefore can increase its application potential.

6.5 Summary

Two types of two-source PS method are introduced in this chapter. The former one employs the integrability and continuity properties to remove ambiguous surface gradients and realises surface normal reconstruction. The latter one mirrors the illumination directions under the symmetry assumption of faces, and then recovers half of a 3D face. The two methods are complementary in terms of the deployment of the illuminators (non-diagonal or diagonal). They both have the merits that 1) only two images need to be captured and stored, which allows the PS technique to be more efficient and that 2) desirable reconstruction results can be obtained for both synthetic data and real data.

In addition, the two PS variations can give rise to simplified PS capture systems that have enhanced applicability to real-world scenarios and can be used to gather 3D data for more robust gender/age classification.

Combined with pupillary distance measures available from the proposed eye centre localisation algorithm, the PS method can be tuned to correspond to any position of a user by adjusting the lighting vectors. Therefore, the movement of a user will no longer be restrained as in conventional PS settings where lighting vectors are normally fixed and known. The utilisation of 3D data is thus more unconstrained and realistic.

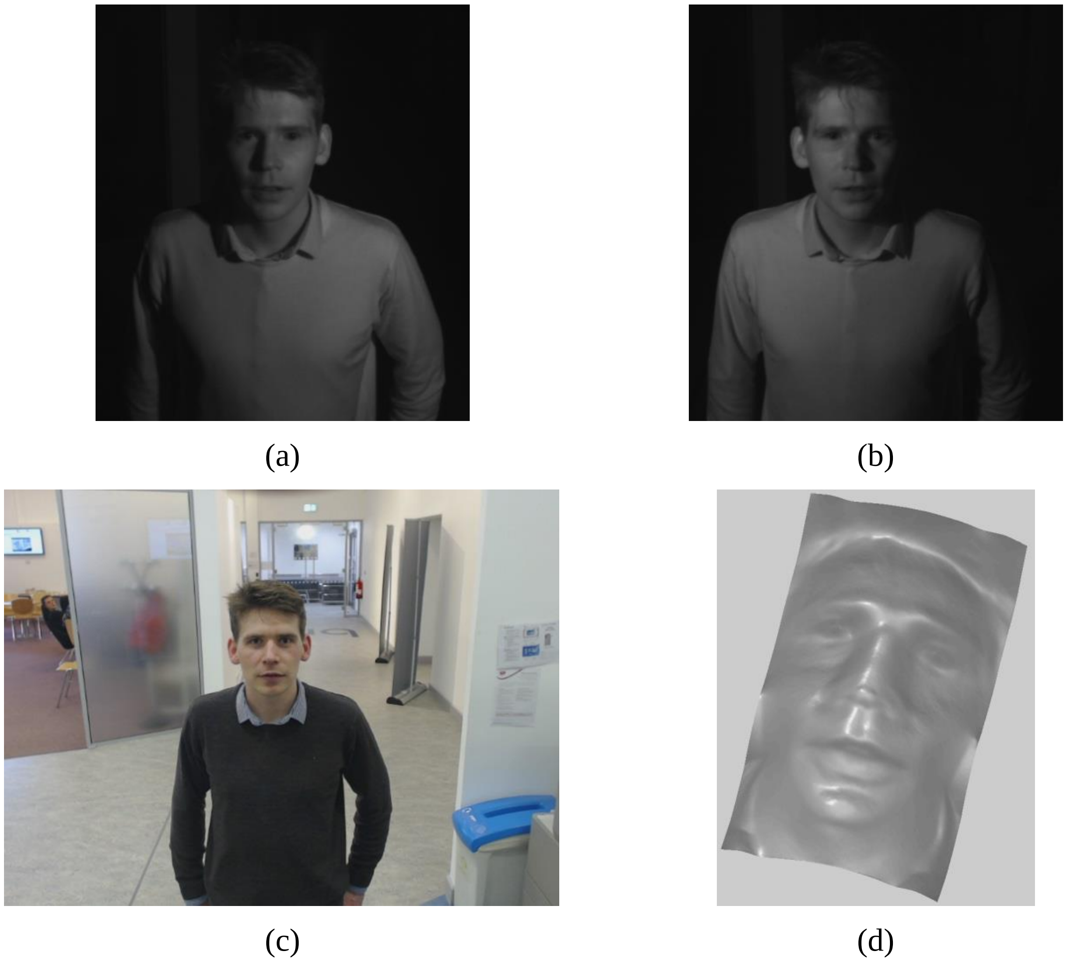
Chapter 7 Gender and Age Recognition: 2D + 3D

In Chapter 5, it has been demonstrated that even with greyscale images, the proposed FV encoding method is able to achieve state-of-the-art gender recognition accuracy. As much as this gives promising reliability, the inherent limitation of utilising greyscale or colour images mainly results from changes in lighting conditions. This cannot be resolved by designing classifiers with higher discriminative power, but can be tackled by seeking light-independent features –3D facial features.

This chapter applies the FV encoding method to age classification and presents a unique 2D+3D HCI scheme. It demonstrates that, with the incorporation of 2D and 3D data, gender and age recognition can be realised with higher reliability, robustness and adaptability.

7.1 Gender Recognition Experiments and Results on 3D images

In this section, a comparison between utilising 2D facial features and 3D features is drawn for FV gender recognition. For 2D facial features, the SIFT features which had gained the highest classification rate in previous experiments, were evaluated as a reference. For 3D facial features, 3D reconstructions were performed at the first stage by employing the two-source PS method introduced in Section 6.2. Following 3D face reconstructions, many types of 3D facial features are extracted. They include: 1) x gradient features, 2) depth features, 3) SIFT features extracted from x gradient images (gradient-SIFT), 4) SIFT features extracted from depth images (depth-SIFT), 5) LBP features extracted from x gradient images (gradient-LBP) and 6) LBP features extracted from depth images (depth-LBP). x gradients have been utilised instead of y gradients since the two light sources are deployed along the x axis but they have the same y coordinate. This indicates that surface reconstructions are largely based on illumination changes along the x axis rather than the y axis.



(a)

(b)

(c)

(d)

Figure 7.1 A group of image data for one subject: the PS image set in (a) and (b), the colour image of the same subject in (c), and the reconstructed 3D face in (d).

Currently, no public databases have been discovered that contain both PS image sets and the corresponding greyscale/colour images which, at the same time, have a suitable male-female ratio. Therefore the data captured by the 2D+3D imaging system (introduced in Chapter 3) were employed for 2D and 3D gender recognition evaluations.

For all the 150 subjects, both the colour images (which are converted into greyscale images before FV encoding) and the PS image sets are employed for evaluations respectively. A representative group of image data for one subject is shown in Figure 7.1.

Various types of features were extracted from them and were encoded into FVs following the method introduced in Chapter 5. The optimal parameters identified in Section 5.3 were employed while the images were resized to 192×144 for a consistent aspect ratio.

The classification rates resulting from the x gradient features, the depth features, the gradient-SIFT features, the depth-SIFT features, the gradient-LBP features and the depth-LBP features are shown in Figure 7.2.

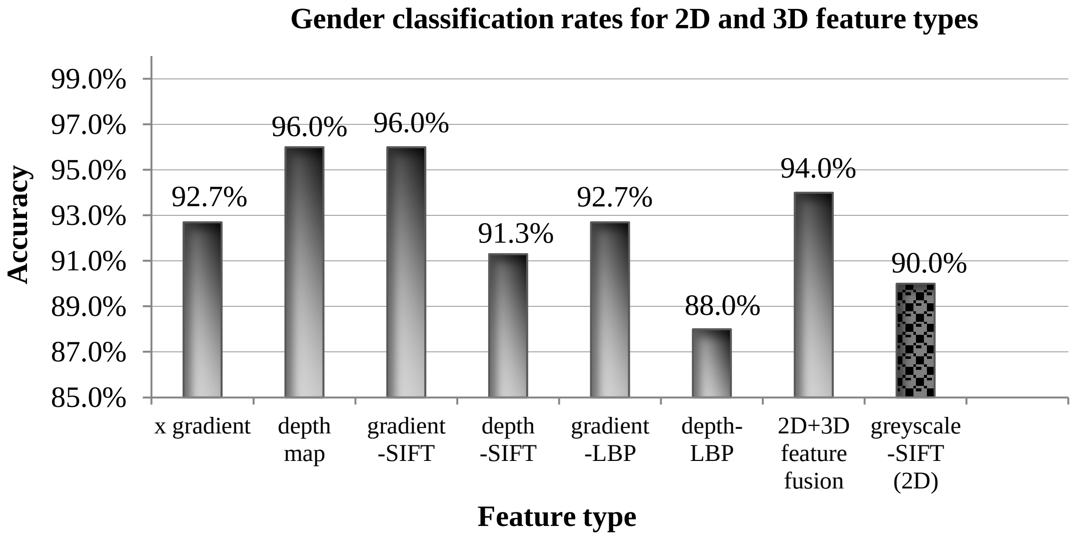


Figure 7.2 Evaluations of 2D and 3D features for FV encoding regarding gender recognition

It can be seen from Figure 7.2 that 3D features generally have resulted in superior classification accuracies over 2D features, except for the depth-LBP features which are most likely to have been interfered by image noise. The highest accuracy obtained in this experiment surpassed the 2D SIFT features by 6%. In this experiment, this increase in accuracy is achieved when the illumination conditions are acceptable. Therefore, it can be inferred that when illumination conditions worsen, 2D features will suffer more while 3D features are much less influenced, hence at such times greater superiority of 3D features can be observed. In addition, the fusion of greyscale-SIFT features (2D) and gradient-SIFT features (3D) was also investigated. This fusion process was achieved by extracting both types of features that correspond to the top 256 Gaussian components learnt by the GMM. The concatenation of these features then became the fused facial descriptors to be classified. However, the fused features were outperformed by 3D features alone. This is likely due to the redundancy/interference of 2D features that could cause the feature space to become less separable. In order to verify this intuitive hypothesis, an additional evaluation experiment was conducted, regarding gender recognition under different illumination conditions. This experiment was also enabled by the 2D+3D data capture experiment introduced in Section 3.2, which provided image data of the same subjects under various and controllable illumination conditions. Overall, three different illumination conditions were simulated, including NIR illumination from top-left direction, NIR illumination from top-right direction (i.e. the two-source PS setting), and relatively uniform illumination in an indoor environment. One set of the resulting NIR and greyscale face images, as well as the x gradient image from two-source PS reconstruction, is displayed in Figure 7.3.

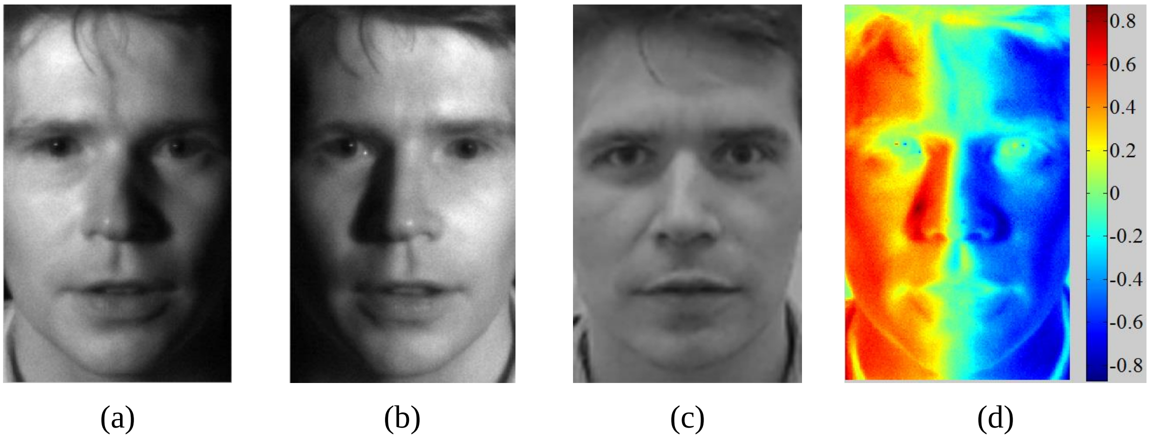


Figure 7.3 Images of the same subject illuminated differently: (a) and (b) under NIR illumination from opposite directions, and (c) by indoor visible lighting, with a comparison to their corresponding *x* gradient image in (d), which reflects facial topography and is independent of illumination.

SIFT features were employed in this experiment as they proved to be effective by previous results where SIFT features extracted from images (example in Figure 7.3(c), 2D) under uniform indoor illumination yielded 90% accuracy, while those from x gradient reconstructions (example in Figure 7.3(d), 3D) yielded 96% accuracy. As reliable as 2D images seemed to be by achieving a classification rate of 90%, they proved to be volatile by this experiment. When the training set contained face images illuminated from the left and the testing set contained those illuminated from the right, the gender classification rate plunged severely to only 50.7% in a five-fold cross validation test. This test complied with the convention that the training set and the testing set should not contain images of the subjects. More interestingly, when this convention was violated in a 2-fold cross validation test, i.e. images illuminated from the top-left direction were used for training and images of *the same subjects* illuminated from the opposite direction were used for testing, and vice versa, the classification rate did not receive any improvement. These results are summarised in Figure 7.4.

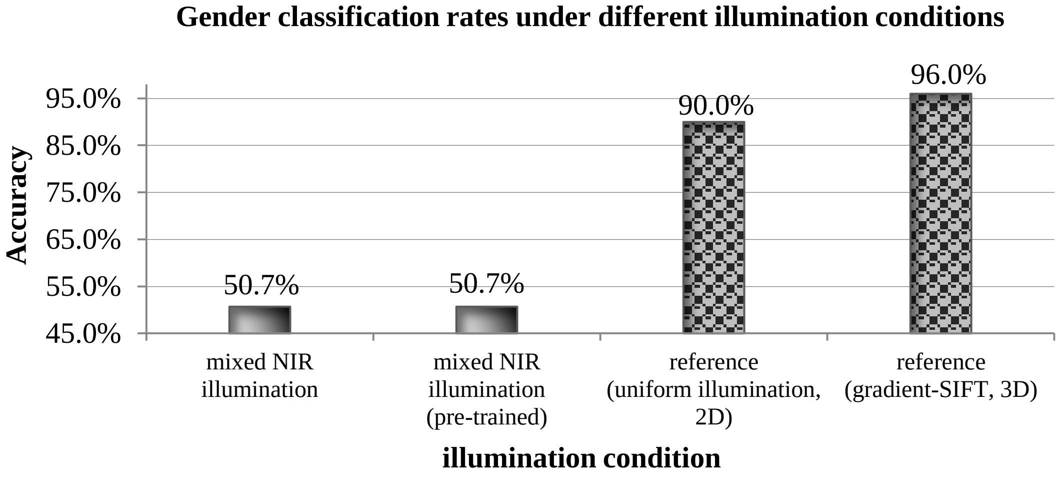


Figure 7.4 Gender classification results under different illumination conditions.

This result showed that even with a classifier pre-trained by the same subjects used for testing, the impact of illumination variation interfered to such an extent that the classifier would not be able to re-identify the gender of nearly 50% of the subjects. One natural reflection accounting for this is that 2D-based methods extract illumination-dependent features which are volatile and thus unreliable compared to 3D features. Being illumination independent, 3D features can tolerate variations in dynamic scenes which commonly render 2D features unstable and incapable of discriminating gender groups or other groups that need to be classified.

7.2 Age Classification Experiments and Results on 3D images

It has been illustrated in Chapter 5 that FV encoding is a generic method that can encode almost any type of low level dense features into more discriminative representations. For gender recognition, over ten types of 2D and 3D features have been encoded and then fed into a linear SVM for classification. It has been proved by various experiments on four databases (three public databases and one self-collected database) that FV encoding for gender recognition is an effective and robust method that can promise state-of-the-art accuracy in both controlled and uncontrolled environments. It has been also evidenced that 3D features generally produce superior classification rates over those by 2D features.

Therefore, it is intuitive to extend this method to the classification of age labels and to make use of 3D features for high accuracy and robustness.

Although no PS based databases have been identified that are suitable for the evaluation of 3D gender recognition, the Photoface database can be employed for the evaluation of 3D age classification since it contains 261 subjects with a wide age range. The only limitation is that this database does not contain an age label for every subject. However, it is possible to manually divide these facial data into two age groups – group ‘young’ and group ‘senior’ – with sufficient reliability. Trying to predict the exact age label for every subject in the Photoface database is not a realistic and feasible solution since the database lacks precise age labels, while constructing a new PS based database that contains a large number of subjects with a wide age range is extremely time-consuming. In addition, given that investigations on age classification in the literature have not achieved great success, it is more realistic and sensible to treat age classification as a two-class problem, rather than an ambitious attempt to reach for more classes. This can simplify the problem, and can at the same time objectively evaluate the proposed age classification algorithm.

More specifically, the subjects that appear to be below 40 years old are placed into group ‘young’, each with an age label denoted by a ‘0’; Those that appear to be of or above 40 years old are placed into group ‘senior’, each with age label denoted by a ‘1’. This manual labelling process was performed by one person initially and was then verified by two volunteers to ensure the attached age group labels were objective and reliable. In the experiment, the first 200 subjects in both groups were employed, from which only one facial image per subject was used such that the training set and the testing set would not contain different facial images of the same subject. These facial images were then processed following the five stages in the FV encoding method – face pre-processing, low-level feature and face descriptor computation, dimension reduction, FV encoding and feature selection (see Section 5.2 for more details). Every face region is resized to a maximum of 320×240 pixels with an aspect ratio of 3:4 such that it is consistent with previous gender recognition experiments. Similarly to 3D gender recognition experiments, the features extracted for age classification include: 1) x gradient features,

2) depth features, 3) SIFT features extracted from x gradient images (gradient-SIFT), 4) SIFT features extracted from the depth images (depth-SIFT), 5) LBP features extracted from the x gradient images (gradient-LBP), 6) LBP features extracted from the depth images (depth-LBP), and 7) SIFT features extracted from albedo images (albedo-SIFT). It should be noted that although albedo images seem to be 2D face representations, they are recovered from the PS method based on the reconstructions of surface normals. Therefore they are illumination-independent and should be differentiated from other 2D face representations such as greyscale images or colour images. The age classification rates for the seven types of 3D features are shown in Figure 7.5, with a comparison to 2D features extracted from images illuminated differently.

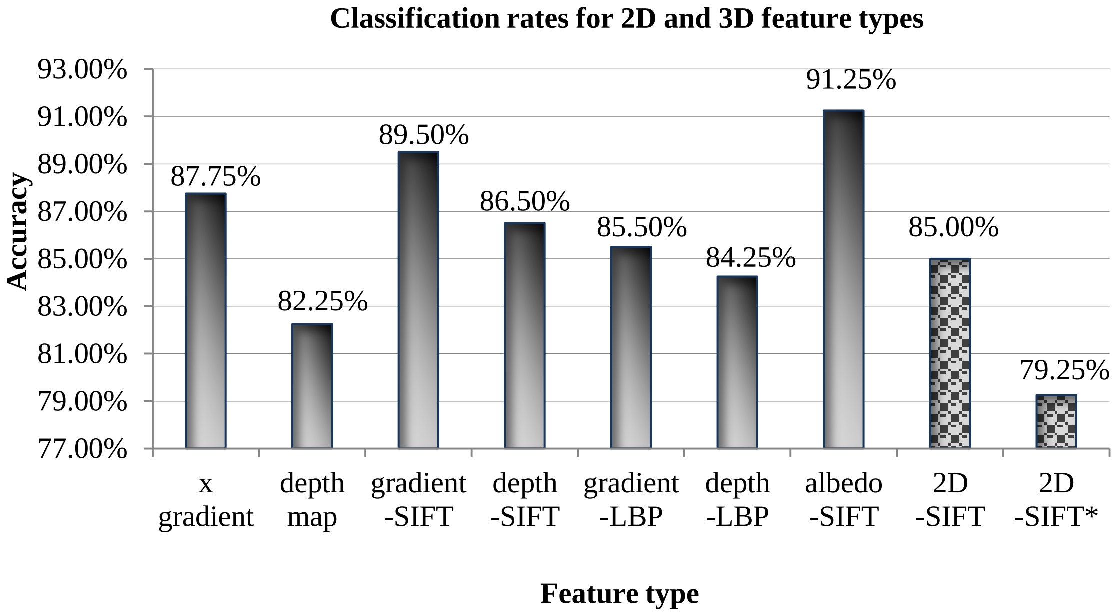


Figure 7.5 Evaluations of 3D features for FV encoding regarding age recognition

The first 7 bars in Figure 7.5 represent results obtained by 3D features. Similarly to other studies in the literature (Gökberk, İrfanoğlu and Akarun, 2006; Hansen, 2012), the results indicate that depth maps are not as effective as other 3D feature types. Encoded x gradient features and albedo features produced the highest classification rates in this experiment. This is likely due to the integration errors during depth calculation, meaning that surface gradients and albedo features from PS reconstructions should be more accurate than reconstructed depth maps. The last 2 bars represent results obtained by 2D features. Denoted by ‘2D-SIFT’ in Figure 7.5, features extracted from greyscale images of subjects illuminated from the top-left direction produced 85% accuracy; Denoted by ‘2D-SIFT*’, features by a mixture of top-left and top-right illuminations (the mixture rate being 50% to 50%) produced a worse classification rate of 79.25%. These results indicate that consistent illumination tends to generate more effective 2D features. The results also reinforces the previous conclusion to gender recognition that 3D features are generally more reliable than 2D features and are thus worthy of extensive exploitation.

7.3 Summary

This chapter demonstrates the effectiveness of employing 3D facial features for demographic recognition tasks, i.e. gender recognition and age recognition.

For gender recognition, the advantages of 3D features have been illustrated by a set of experiments. More specifically, it has been proved in Section 5.3 that encoding dense SIFT features into FVs outperforms most state-of-the-art gender recognition methods. Subsequently, it has been further proved in Section 7.1 that encoding 3D feature types, compared to 2D dense SIFT features, can further boost the performance of automatic gender recognition. Therefore, it can be concluded that the FV encoding method when employing 3D features exhibits superior excellence in gender recognition. By simulating different illumination conditions, it has been demonstrated that 2D features are volatile due to environmental variations which has negligible impact on 3D features. In the experiment, the employment of 3D features brought an increase of up to 45.3% to gender classification when the illumination was dynamic (see Figure 7.4).

For age classification, the FV encoding method continues to contribute to desirable classification rates for a two-class problem. In a comparison where seven types of 3D features battled against 2D features, it was found that the albedo features achieved the highest classification rate and that illumination variation was a persistent negative factor for both classification tasks. The superiority of 3D features is therefore sufficiently evidenced by all the experimental results demonstrated in this chapter.

The predicted gender and age labels of a user can serve as demographic statistics for personalised HCI sessions. The 3D recognition algorithms and the 2D+3D imaging systems have been designed in a way such that they are fully compatible and that they can be seamlessly combined and implemented as a HCI system which can function in real-world environments.

Chapter 8 Behavioural Analysis and Demographic Recognition for Human-Computer Interaction

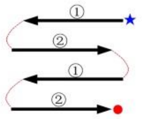
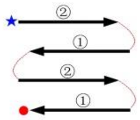
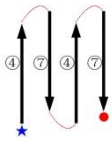
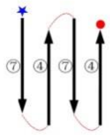
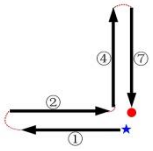
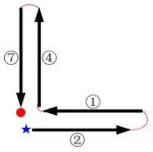
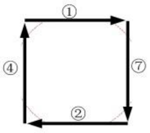
This chapter is dedicated to the presentation of two case studies, intended to validate the practicability of all the proposed algorithms and to elaborate the tremendous potential of the proposed HCI strategy that features a combination of demographic recognition and behavioural recognition. In addition, these two case studies are the embodiment of the theoretical illustration of the proposed HCI strategy. They shed light upon the broad applicable areas that benefit from understanding user gaze, gender and age, as well as indicating the potential of a boosted HCI strategy from reliable analysis of more abundant and diversified user data.

8.1 Case Study One – A Remote Map Browser by Gaze Gestures

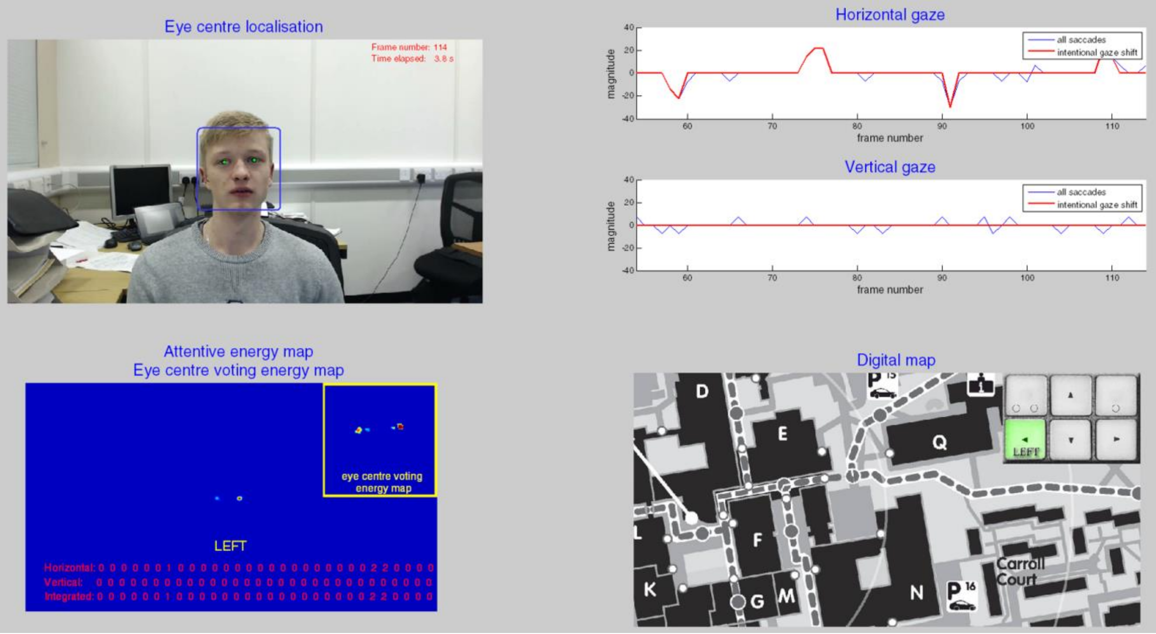
Books, brochures and other print materials are becoming widely available in digital forms. Access to them has turned pervasive and convenient, even in public venues. However, these messages are normally approached and browsed by contact with hands, fingers and styluses. While this manner fits reasonably well to individual-targeted and close-range settings, out-of-home HCI scenarios may suffer from it. For example, many digital billboards at transportation venues come in large sizes, e.g. as large as 56 inches, and are placed at a great distance from the ground for safety reasons. Interactive forms of such billboards have not been widely put into use due to the inconvenience that might be incurred: 1) they will be beyond reach of users of average height; and 2) user perception will be limited due to the field of view at close range during an interaction. Another example concerns those with motor disability who cannot easily use their hands to perform actions such as ‘swipe/scroll to navigate’ or ‘pinch to zoom’. In these circumstances, a contactless HCI method is in high demand. A solution to resolving such limitations is the proposed gaze gesture recognition algorithm that can be employed by a user to remotely browse digital messages on a screen with ease and convenience.

This section introduces a remote map browser that allows its users to manipulate a digital map. Being facilitated by the proposed gaze gesture recognition algorithm (introduced in Section 4.3), this map browser only requires a webcam to capture facial images and a screen for map display. Seven types of input commands have been designed to navigate map ‘up’, ‘down’, ‘left’ and ‘right’; to ‘zoom in’ and ‘zoom out’; or to issue a ‘print’ command, i.e. to request a printed copy of the map. The correspondence between these pre-defined gaze gestures and HCI events are listed in Table 8.1, in a similar way to Table 4.2. It should be noted that the complexity of performing a gaze gesture is mainly determined by the number of saccadic codes that form the gesture pattern. A large number of saccadic codes in a gaze gesture will increase the uniqueness of it, but they will raise its complexity (i.e. lower usability). Too few saccadic codes, on the other hand, would likely to cause a gaze gesture to be sensitive to unintentional eye movements. In the proposed design, every gaze gesture is comprised of four saccadic codes. It was found empirically that this would allow for intuitive user input while achieving robust performance.

Table 8.1 Gaze gestures defined for the remote map browser

Gesture No.	Gesture Sequence	Gesture Pattern	Gesture Name	HCI Event
1	1 → 2 → 1 → 2		Left gaze	navigate map *left*
2	2 → 1 → 2 → 1		Right gaze	navigate map *right*
3	4 → 7 → 4 → 7		Up gaze	navigate map *up*
4	7 → 4 → 7 → 4		Down gaze	navigate map *down*
5	1 → 2 → 4 → 7		Zoom-in gaze	*zoom in* on map
6	2 → 1 → 4 → 7		Zoom-out gaze	*zoom out* on map
7	2 → 4 → 1 → 7 4 → 1 → 7 → 2 1 → 7 → 2 → 4 7 → 2 → 4 → 1		Print gaze	Request a *print* copy of map

As the proposed gaze gesture recognition method is robust to changes in head pose, a user is enabled to issue any gaze gesture with eye movement, head movement, or a combination. A video (Gaze_gesture_map_browser_1.mp4) is made to demonstrate the functionality of the gaze gesture based remote map browser, where a first-time user browses a digital map with mainly eye movements. A representative frame from the video is shown in Figure 8.1.



```
Frame number: 114
Time elapsed:   3.8 s
```

```
Horizontal: 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0
Vertical:   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Integrated: 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0
```

Figure 8.1 A demonstration of navigating the map left. Other navigation gaze gestures result in similar map display.

As shown by Figure 8.1, the 'Left' gaze gesture has been issued and recognised, which results in the view of the map being shifted to the left. The top-left plot shows the raw face image. The detected face is within the blue bounding box, and the automatically localised eye centres are marked by green dots. The top-right plot shows the recorded horizontal gaze shifts and vertical gaze saccades that are essentially derivatives of the eye centre positions in the x and y directions. All the eye saccades are represented by blue pulses, while intentional eye saccades (those exceeding a threshold) are represented by red pulses. Note that the threshold is defined to be a value relative to the pupillary distance such that it can adapt to changes in user-camera distance (see Section 4.3). The bottom-left plot is the attentive energy map that accumulates the eye centre positions over time to reflect the relative user attentiveness. Energy values are mapped to the jet colourmap, meaning that red pixels denote high energy points while blue pixels denote low energy points. In addition, another energy map constructed for eye centre localisation is displayed at the top-right corner of this plot. Note that low energy points have been removed, as explained in subsection 4.1.1. The encoded horizontal and vertical eye saccades are displayed at the bottom of this plot as sequences of ‘0’, ‘1’, ‘2’, ‘4’ and ‘7’. The bottom-right plot is the view of a digital map controlled by gaze gestures.

While some users may prefer to use eye movements to input gaze gestures, others choose a combination of eye movements and head movements. To demonstrate the robustness of this method against changes in head pose, a video (Gaze_gesture_map_browser_2.mp4) is made where a user is issuing gaze gestures in a more natural manner. A representative frame from this video is shown in Figure 8.2.

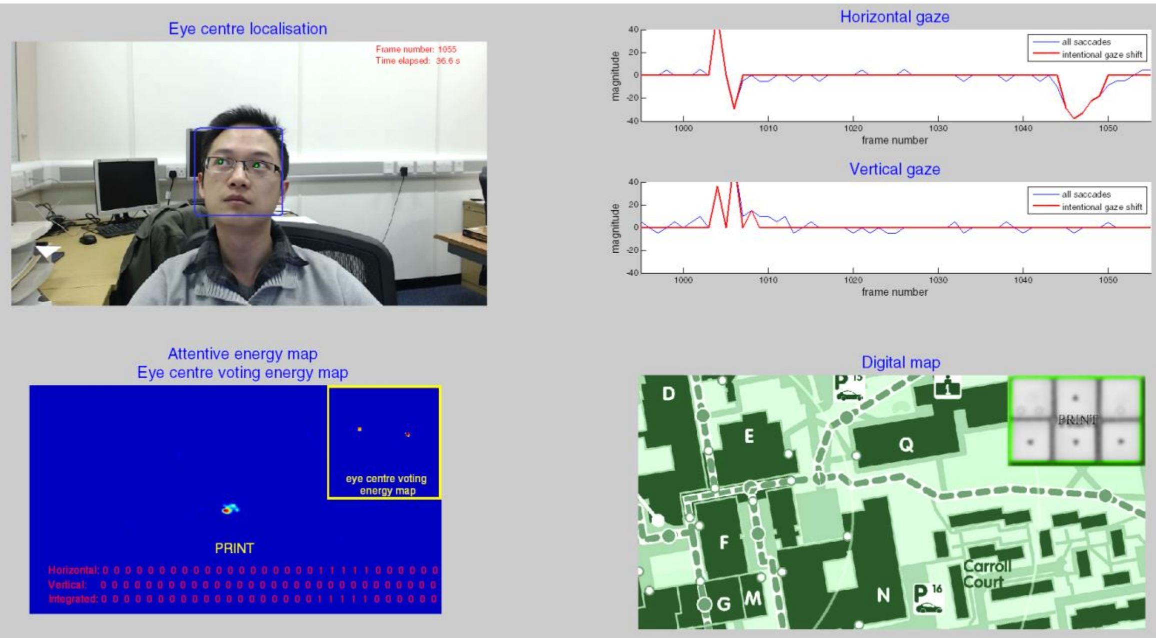

Eye centre localisation
Frame number: 1055
Time elapsed: 36.6 s
Horizontal gaze
Vertical gaze
magnitude
40
20
0
-20
-40
1000
1010
1020
1030
1040
1050
frame number
all saccades
intentional gaze shift
Attentive energy map
Eye centre voting energy map
eye centre voting energy map
PRINT
Horizontal: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0
Vertical: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Integrated: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0
Digital map
D
E
Q
F
G
M
N
P 15
P 16
Carroll Court
PRINT

Figure 8.2 A demonstration of issuing the ‘Print’ gaze gesture with natural eye/head movements.

It can be reflected by Figure 8.2 and the video that eye centres can be accurately localised on a face with different head poses and with a glass frame. The efficiency of the proposed eye centre localisation algorithm and the gaze gesture recognition algorithm has been evaluated and illustrated in subsection 4.4.1. Over 30 frames can be processed per second, meaning that the proposed method can promise to facilitate real-time implementations with satisfactory accuracy and robustness.

Other different gaze gestures can be designed to cater to HCI applications of similar functionalities. Users can then manipulate virtual objects on a screen (e.g. switch TV channels or turn up the speaker) or browse specific content (e.g. an interactive menu or an electronic brochure). It has great potential for assistance of the disabled and the elderly who cannot easily manipulate HCI systems at public venues such as airports, train stations and shopping malls.

8.2 Case Study Two – An Intelligent Directed Advertising Billboard

This section presents an intelligent directed advertising billboard that serves to deliver personalised advertising messages for different age groups, gender groups and dynamic user behaviours. Digital signage systems have been prevalent for years in this digital era. Their use for advertising has become ubiquitous and can be found at venues such as restaurants, shopping malls, airports and other public spaces. Often referred to as digital out-of-home (DOOH) advertising (Lasinger and Bauer, 2013; Stalder, 2011), this advertising format aims to extend the exposure and the effectiveness of marketing messages by engaging consumers to an increased extent, compared to conventional print based billboards (Pieters, Warlop and Wedel, 2002). Following the idea of switching from print media to digital media, it is only intuitive to reform a conventional DOOH advertising system toward a HCI system (Adams, 2004) for enhanced interactivity and adaptability.

To this end, an experimental interactive advertising system was designed as one of the case studies. It demonstrates how behavioural recognition and demographic recognition can be integrated to increase the interactivity and flexibility of a HCI system. The design of the system follows the same structure as the 2D+3D imaging system introduced in Chapter 3 (Figure 3.1), where the HD display acts as a billboard capable of displaying an advertisement. The algorithms employed to drive the interaction process include the eye centre localisation approach and the gaze gesture recognition method for eye/gaze analysis, the unconstrained two-source PS technique for 3D face reconstruction, and the FV encoding method for gender and age recognition.

Randomly selected advertisement thumbnails will be brought to circulation on the screen until a face is detected by the system. When an individual approaches the system, his/her facial images are captured by both cameras. The FV encoding method then outputs the predicted age and gender label for the facial image in every frame, using 3D data for higher accuracy and robustness (see Chapter 7) when the detected face region is larger than 100×100 pixels, or using 2D data otherwise. As a result, advertisement thumbnails that are highly relevant to the predicted age and gender group are displayed in replacement to previous advertising messages. The eye centre localisation algorithm then detects eye centres in every greyscale image frame and supplies the information to the gaze gesture recognition algorithm. When a gaze sequence matches one of the seven pre-defined gaze gesture patterns (see Table 4.2), the advertising messages displayed on the screen can be manipulated correspondingly. Although gender and age specific advertisements will be delivered by default, a user can issue the ‘reset’ gaze gesture to deactivate the use of gender and age labels for advertisement selection. One can also issue the ‘change-content’ gaze gesture for advertisement browsing.

A video demonstration (see Directed_advertising.mp4) is made to illustrate this HCI process where the display of advertisements responds to the gender, age and gaze of a user. A few representative frames are shown in Figure 8.3, Figure 8.4, Figure 8.5 and Figure 8.6.

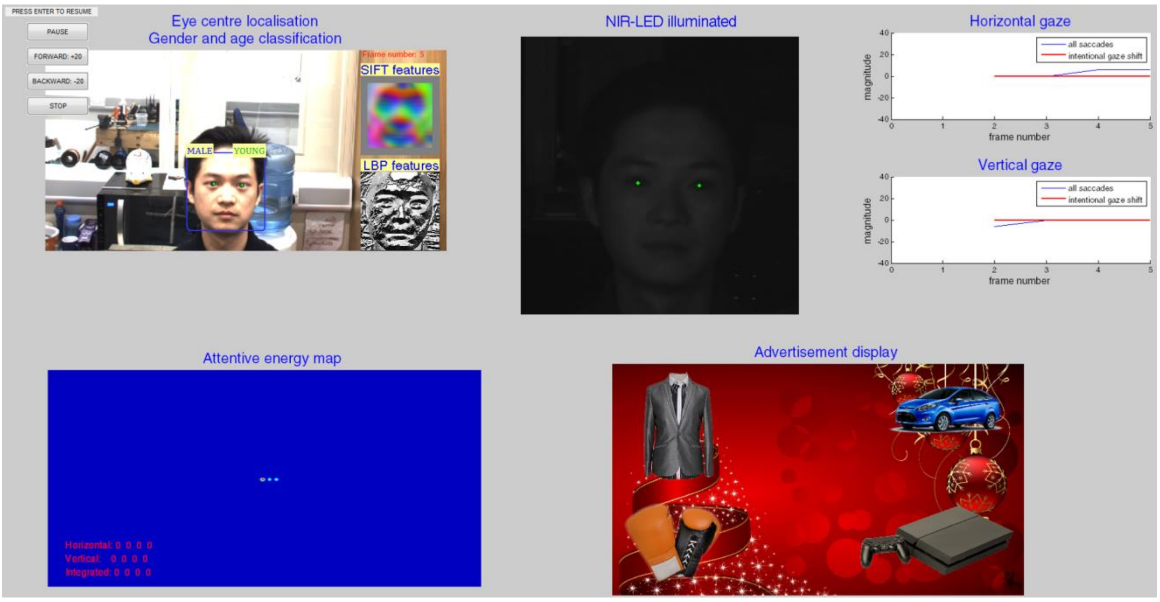
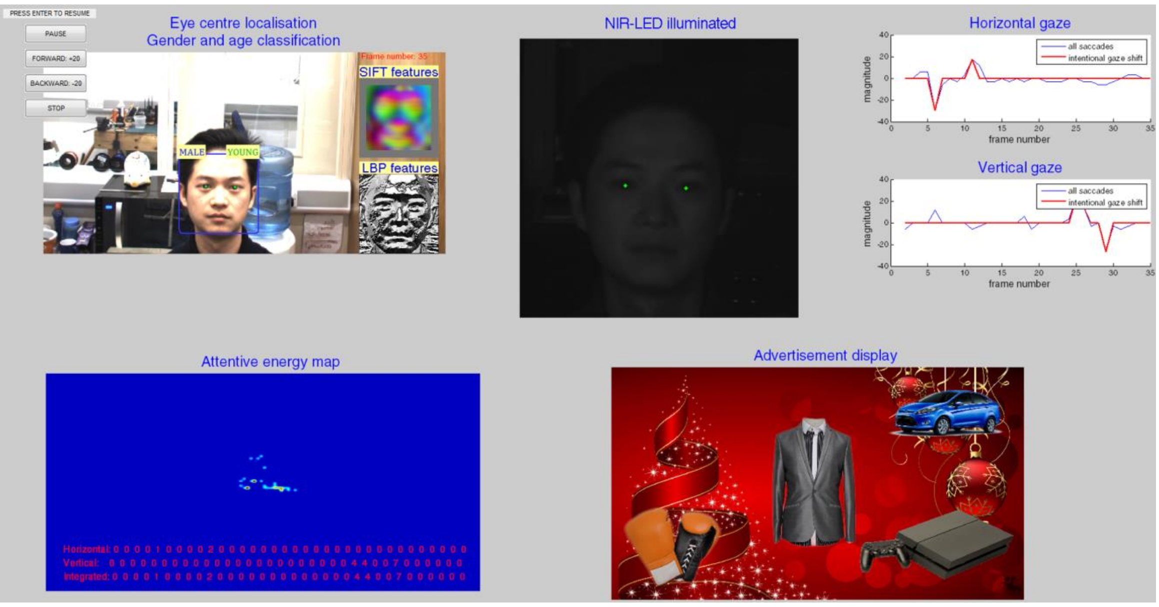


Figure 8.3 A frame from the video demonstration showing that the directed advertising system delivered advertisements based on the user's gender and age. Eye centres were accurately localised on both the colour image and the NIR image. An attentive energy map was constructed to reflect the relative attentiveness of the user over time.

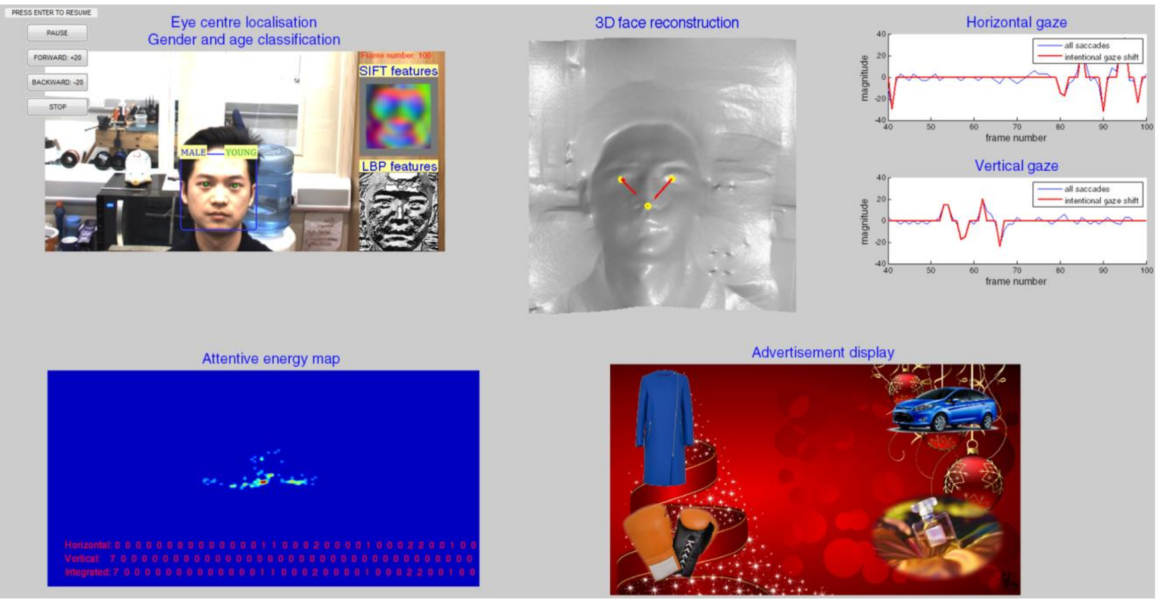
In Figure 8.3, the top-left image shows the detected face region with a blue bounding box and the localised eye centres represented by two green dots. It also displays the estimated gender and age labels, as well as visualising the SIFT features and the LBP features (see Chapter 5). The top-middle image displays the average image of the two NIR images captured under the two-source PS setting. Eye centres have been accurately localised on these poorly illuminated images by including a simple contrast enhancement step in the pre-processing stage. The top-right image displays horizontal and vertical eye saccades, as in the preceding case study. The bottom-left image shows the attentive energy map and the encoded eye saccades. The bottom-right image shows the advertisements being delivered.



```
Horizontal: 0 0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Vertical:   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 4 0 0 7 0 0 0 0 0 0
Integrated: 0 0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 4 4 0 0 7 0 0 0 0 0 0
```

Figure 8.4 A frame from the video demonstration showing that the top-left gaze gesture was issued consisting of four saccadic codes ①, ②, ④, ⑦. Upon the recognition of this pattern, the advertisement thumbnail on the top-left corner was placed at the centre of the screen. Refer to Table 4.1 for six other pre-defined gaze gestures for directed advertising.

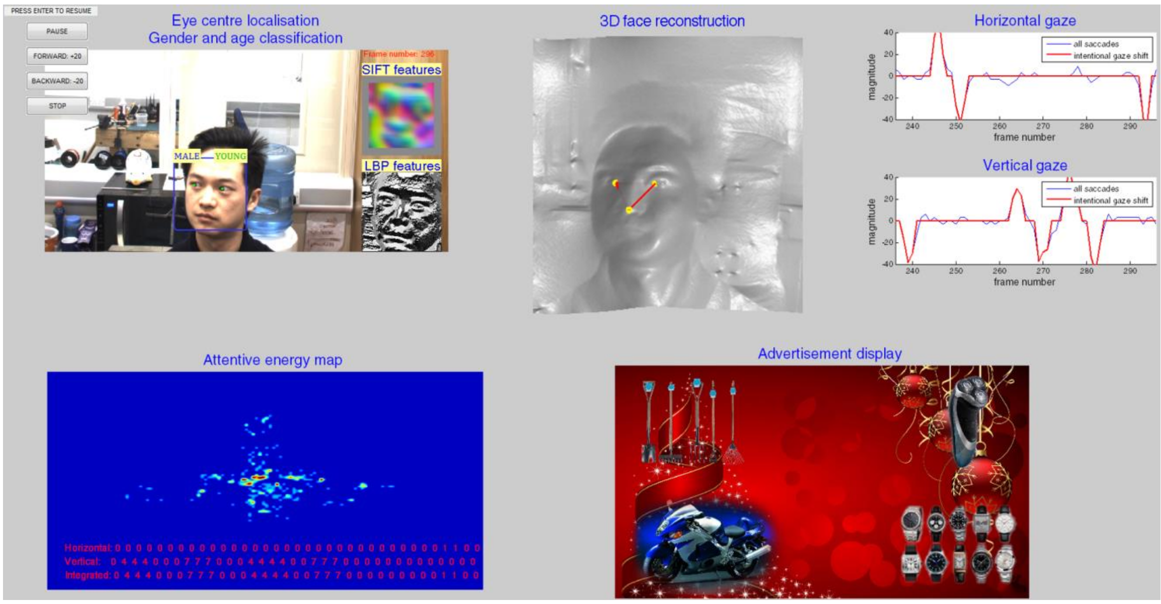
Figure 8.4 shows that when the top-left gaze gesture was detected, the top-right advertisement thumbnail was displayed at the screen centre. If the ‘zoom-in’ gaze gesture was detected as a subsequent gesture, an enlarged view of the centred thumbnail would become available.



```
Horizontal: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 2 0 0 0 0 1 0 0 0 2 2 0 0 1 0 0
Vertical:   7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Integrated: 7 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 2 0 0 0 0 1 0 0 0 2 2 0 0 1 0 0
```

Figure 8.5 A frame from the video demonstration showing that the user performed the 'reset' gaze gesture such that the advertisements delivered were no longer gender and age specific. In addition, the corresponding 3D face reconstruction is displayed where the eye centres (localised on the pair of NIR images) and the nose tip (localised on the depth image) form a triangle that can be used for head pose estimation and eye fixation estimation (see Section 4.4.3).

As shown by Figure 8.5, the 'reset' gaze gesture was issued and detected such that the advertisements delivered were no longer specific to the 'young' 'male' group. The top-middle image shows the 3D face of the user reconstructed by the proposed two-source PS method. The eye centres and the nose tip of the user were localised on the NIR images and the depth image, respectively. The triangle formed by the three facial landmarks will deform as the user makes eye/head movement. This offers a huge potential for head pose estimation and eye fixation estimation (see subsection 4.4.3).



```
Horizontal: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
Vertical:   0 4 4 4 0 0 0 7 7 7 0 0 0 4 4 4 4 0 0 7 7 7 0 0 0 0 0 0 0 0 0 0 0 0 0
Integrated: 0 4 4 4 0 0 0 7 7 7 0 0 0 4 4 4 4 0 0 7 7 7 0 0 0 0 0 0 0 0 0 1 1 0 0
```

Figure 8.6 A frame from the video demonstration showing that the user performed the 'change-content' gaze gesture using both eye movement and head movement. Results show that all the proposed algorithms are highly robust to moderate head rotations (which are typical scenarios for HCI).

Figure 8.6 shows that the user performed the 'change-content' gaze gesture by combining eye movement and head movement to browse advertisements. It demonstrates that the proposed eye centre localisation algorithm, the gaze gesture recognition algorithm and the two-source PS algorithm can cope with variations in head pose.

In the case where a group of individuals are recognised by the system, the distance between an individual and the cameras will be estimated. This is also facilitated by the proposed eye centre localisation algorithm that calculates the pupillary distance in the image domain. The one individual that is the closest to the system (the one with the largest pupillary distance in the image) obtains the highest priority in controlling the system. This is based on the fact that the pupillary distance in the real-world domain is relatively constant among individuals. The estimated user-camera distance can also contribute to an adaptive mechanism that switches between the 2D modality and the 3D modality. At a relatively large user-camera distance (e.g. three metres and above), 3D reconstructions will be hindered by low resolution data and a low signal-to-noise level.

Therefore the directed advertising system should automatically switch to the 2D modality by only employing 2D features for high efficiency. Otherwise, 2D features should be employed for eye centre localisation and gaze gesture recognition while 3D features should be incorporated for gender and age classification.

8.3 Summary

The two proofs of concept presented in this chapter are intended to exemplify the HCI strategy where demographic recognition and behavioural recognition are employed individually or collectively. The purpose of the two systems is to demonstrate the proposed algorithms in real-world scenarios. As stated in Section 1.3, validation of the effectiveness of the resulting HCI systems is beyond the scope of this research.

By gathering behavioural data, the remote map browser can recognise intentional gaze gestures issued by users who need to navigate a digital map on a screen. The simplicity of its design and implementation does not compromise its effectiveness and reliability, and therefore brings huge potential to the development of other HCI systems that can be remotely controlled, at close range or at a long distance.

The intelligent directed advertising billboard embodies the integration of demographic recognition and behaviour recognition for a user-centred HCI experience. Gender label and age label of a user are recognised by this HCI system, which can be used to determine the theme of advertising messages. As well as being able to receive personalised advertisements (which can be deactivated by the ‘reset’ gaze gesture) in a passive manner, users are enabled to actively interact with the system and selectively view the content. This HCI strategy is not confined to a particular case, but it can be applied to other scenarios such as care at hospitals and nursing homes, where systems can be designed to assist with patients from different gender and age groups, and in different physical conditions.

Chapter 9 Conclusions and Future Works

9.1 Conclusion

This thesis presents a HCI strategy that employs the visual modality to advance HCI system designs towards a more effective and natural state. This HCI strategy is enabled by a combination of demographic recognition and behavioural recognition in 2D and 3D. As it is not realistic to perform an exhaustive study of all types of user demographic data (e.g. gender, age, ethnicity, etc.) and behavioural data (e.g. gaze, facial expression, head pose, etc.), the benefits delivered by the proposed HCI strategy is firstly illustrated by theoretical discussions that:

1) Comprehension of user characteristics and user behaviours can serve to fulfil user expectations, to augment user satisfaction and to boost user experience in HCI sessions;

2) Demographic data can only determine the initial state of a HCI process, while behaviour data progressively and adaptively address the transient state of the HCI process, accounting for the dynamic nature of HCI;

3) Demographic analysis only serves to facilitate HCI suggestively rather than decisively, i.e. it should be user behaviours ('what they need'), instead of user demographics ('who they are'), that define the manners of HCI.

These theoretical illustrations have then been put into practice by capturing and analysing a selection of representative demographic and behavioural cues (gaze, gender and age). As a result, they exemplify the HCI strategy in an uncompromised manner by exhibiting high functionality, usability and reliability as well as revealing the potential to generalise the proposed HCI strategy, provided that richer and diversified user data are available.

More specifically, novel gender and age recognition algorithms are designed to predict user demographics during HCI sessions; an unsupervised modular eye centre localisation algorithm and a gaze gesture recognition algorithm are also proposed to gather user behavioural data such that user attention and intention can be estimated dynamically. Furthermore, in order that these algorithms are highly reliable in various environmental settings, a PS variation is introduced that can reconstruct light-independent 3D facial features for 3D recognition tasks. The functionalities and the merits of each algorithm can be summarised as follows:

1) The eye centre localisation algorithm

   This novel eye centre localisation algorithm proves to be highly accurate and robust. Tested on the BioID database, it has outperformed ten other methods of its kind. Evaluations on a self-collected database have further validated its robustness on faces with head pose/eye movement and faces illuminated under challenging lighting conditions.

   This algorithm is an unsupervised approach that does not involve any training stage and thus is not biased toward a certain type of training data. It is also efficient and can facilitate eye centre localisation tasks in real time.

   It further provides facial landmarks (eye centre coordinates) for face registration that can boost the accuracy of the proposed gender recognition and age recognition tasks.

2) The gaze gesture recognition algorithm

   Given a sequence of localised eye centres, this algorithm allows users to issue gaze gestures to actively control a HCI system in a remote manner. Its realisation only requires simple hardware configuration and it can particularly assist the elderly and those with motor disabilities to access HCI systems with ease and comfort.

3) The Fisher Vector encoding algorithm

This algorithm is intended for classification tasks and therefore becomes the core of the proposed age and gender recognition methods. Facial features, when encoded as Fisher Vectors, can better serve to accurately and robustly classify faces with different appearance due to gender and age. This is verified by experiments on both facial data gathered in laboratory environments (the FERET database) and those gathered in real-world environments (the LFW database). State-of-the-art recognition rates have been achieved for both databases where challenging illumination conditions and head pose variations are present. This algorithm builds up the foundation for demographic recognition as part of the HCI strategy.

4) The two-source PS algorithm

   This algorithm enables 3D recognition tasks in real-world environments and thus further improves the applicability of the proposed HCI strategy for practical use.

   This algorithm features accurate reconstruction results, desirable reconstruction time and less dependency on hardware configuration. This is critical for HCI systems to gain superior functionality and usability. In addition, implementations of the proposed two-source PS algorithm are low-cost such that their pervasive use in various forms can be expected in the near future.

5) The 3D recognition algorithm for age and gender recognition

   As well as inheriting the merits of Fisher Vectors, the proposed 3D age and gender recognition scheme is facilitated by robust 3D features and could resolve the unreliability of conventional age and gender recognition algorithms due to environmental variations. Experiments on the Photoface database and a self-collected database have proved that this scheme can provide more accurate recognition results compared to 2D recognition methods.

The realisation of these algorithms leads to a number of intelligent HCI systems that can deliver richer user experience by catering to the needs of different user groups. Aimed at creating user-centred HCI environments, two case studies have been explored to evaluate the proposed algorithms and the novel HCI strategy.

The remote map browser and the intelligent directed advertising billboard designed by this research demonstrate that, when reliable demographic recognition and behavioural recognition are employed individually or when combined, prior knowledge about a user can be collected to deliver personalised HCI system feedback. As a result, HCI sessions can be tailored to individuals or groups of individuals for escalated effectiveness and naturalness.

9.2 Limitations

This section discusses the limitations of the proposed algorithms and the 2D+3D imaging system. Possible solutions have also been proposed to motivate future works that can potentially resolve these limitations.

1) The proposed eye centre localisation algorithm is based on geometric eye models. Therefore it requires that a pupil/iris region is not fully occluded in order to perform accurate localisations, meaning that the localisation errors are likely to increase otherwise. This can be resolved by designing a separate module for detection of eye blinks which can result in fully closed eyes. By discarding these frames, less impact will be brought to the gaze gesture algorithm. In addition, this algorithm simply employs the Viola-Jones face detector at the pre-processing stage, as this research is not dedicated to face detection. Therefore, as much as the periocular features employed possess rotational invariance, a face with a large head pose (e.g. rotated by 40 degrees) cannot be detected. Therefore, a more versatile and efficient face detector would bring higher robustness, efficiency and applicability to this algorithm. 2) The FV encoding method for gender and age classification results in a large number of face descriptors, which incur prolonged computational time and a relatively high demand for data storage. However, this only affects the offline training stage but does not impair the efficiency of the online classification. Higher processing power of computers nowadays will continue to allow for increased data storage and faster classifier training. This can potentially facilitate the FV encoding process regarding the scale of training images and computational time.

3) The two-source PS technique functions the best when the two NIR LEDs are at a similar distance to the object being imaged. When the distances between the object to the LEDs are largely unequal, the reconstruction errors will likely increase. This can be resolved by modelling light attenuation with respect to illumination distance. While the inverse square law is a common assumption, the model can be more specifically tailored to a particular type of LED for higher effectiveness.

4) Age classification in this research is considered as a binary classification task. This is due to the lack of a PS database with accurate age labels. Construction of such a database would thus allow for future evaluations of the FV encoding method for multi-class age classification.

9.3 Future Works

This thesis delivers the concept that understanding user characteristics and behaviours supplies critical knowledge to the formation of a natural HCI strategy. As well as being able to enrich functionality and usability of HCI systems by gathering and analysing user information, this strategy serves to fulfil user expectations, to augment user satisfaction and to boost user experience. This HCI strategy is validated by 2D and 3D based computer vision analysis of gaze, gender and age as representative facial cues. It is only intuitive that inclusion of more abundant and diversified facial cues, such as ethnicity, head pose and facial expression, should promise to bring more intelligent and lifelike HCI systems and to enable more natural HCI sessions. It is however unrealistic to examine all these facial cues in this thesis, therefore it is suggested that the remainder of them be pursued in the future.

Overall, the future works described in this section are either an extension of the proposed algorithms and experiments or a continuation of preliminary methods and results presented by this thesis.

9.3.1 2D and 3D face database construction and multi-class age recognition

Chapter 3 is dedicated to introducing a 2D+3D data capture system based on the proposed two-source PS setting. This system provided hardware support for the data capture experiments which gathered facial data for algorithm evaluations in real-world environments. Facial data captured incorporate an array of variations, such as gender, age, head pose, eye/pupil position, camera-object distance and illumination condition.

As reviewed in Section 2.3, age recognition studies in the literature lack reliability and consistency. Therefore, in this thesis, automatic age recognition from facial images is treated as a two-class classification problem, which can be extended to deliver a multi-class solution by modifying the SVM utilised in the proposed algorithm. To evaluate a multi-class age recognition algorithm, capturing sufficient facial data with accurate age labels is critical and is a more arduous task than that for a two-class problem. The main difficulty lies in requiring every age group to have a similar number of subjects.

One convention in the literature is to construct a database with 7 age groups, 0-2, 3-7, 8-12, 13-19, 20-36, 37-65 and 66+ (years old). However, for HCI scenarios, a different group division, i.e. 3-10, 11-20, 21-30, 31-40, 41-50, 51-60, 70+ (years old), can be deemed more suitable since this ensures a more consistent age span for every group so that HCI systems can cater to a user’s needs and capabilities more specifically and appropriately. The number of subjects for every group should be no less than 50, leading to a total number of 350 subjects.

Capturing these data in both 2D and 3D forms has rarely been conducted in the literature. This has largely prohibited the exploitation of 3D based methods whose comparative evaluations against 2D based methods are also restrained.

9.3.2 Head pose and Eye fixation analysis

The preliminary works regarding 3D based head pose and eye fixation analysis have been introduced in subsection 4.4.3. The idea is based on facial landmark detection from greyscale images (face appearance) and depth images (face topography). Localised eye centres and nose tip form a triangle that deforms as head rotation occurs. As one triangle is uniquely associated with a particular head pose, trigonometric analysis can then reveal the head pose defined within a 3D coordinate system.

Given an image frame containing a face, the coordinate of the face, its head pose and eye centre positions can be united to calculate the eye fixation point. Absolute gaze, combined with relative gaze, i.e. gaze gestures, can serve to further improve the functionality and usability of HCI systems by providing interactive and immersive user experience.

9.3.3 Analysis of other facial attributes

While head pose can be an additional facial attribute for assisting with user attentiveness analysis, other facial attributes other than gender, age and gaze can further assist HCI systems in achieving higher intelligence. The idea of associating multiple demographic and behavioural cues is to avoid deriving biased knowledge from a user. Appropriate manners of a HCI system, i.e. HCI strategies, can only be assumed once sufficient knowledge of user characteristic is available. They can then be tuned by the recognition of dynamic user behaviours which can act as feedback to previously formed HCI strategies. For example, apart from gaze analysis that reveals user attentiveness, recognising facial expressions can infer user emotions triggered during a HCI session. This information can overwrite judgements made by a HCI system merely based on the user’s gender and age, and therefore they can lead to a more objective HCI strategy.

References

3dMD *3dMDface System*, [Online], Available: http://www.3dmd.com/3dmd-systems/3d-systems/3dmdface/ [05 April 2015].

Adams, R. (2004) 'Intelligent advertising', *AI & SOCIETY*, vol. 18, no. 1, pp. 68-81.

Aldebaran softbank group *Who is Pepper?*, [Online], Available: https://www.aldebaran.com/en/a-robots/who-is-pepper [23 August 2015].

Alexandre, L.A. (2010) 'Gender recognition: A multiscale decision fusion approach', *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1422-1427.

Alexiadis, D.S., Kelly, P., Daras, P., O'Connor, N.E., Boubekeur, T. and Moussa, M.B. (2011) 'Evaluating a dancer's performance using kinect-based skeleton tracking', Proceedings of the 19th ACM international conference on Multimedia, 659-662.

Amon, C., Fuhrmann, F. and Graf, F. (2014) 'Evaluation of the spatial resolution accuracy of the face tracking system for kinect for windows v1 and v2', Proceedings of the 6th Congress of the Alps Adria Acoustics Association.

Arandjelovic, R. and Zisserman, A. (2012) 'Three things everyone should know to improve object retrieval', IEEE Conference on Computer Vision and Pattern Recognition.

Asadifard, M. and Shanbezadeh, J. (2010) 'Automatic adaptive center of pupil detection using face detection and cdf analysis', Proceedings of International MultiConference of Engineers and Computer Scientists.

Baek, S.J., Choi, K.A., Ma, C., Kim, Y.H. and Ko, S.J. (2013) 'Eyeball model-based iris center localization for visible image-based eye-gaze tracking systems', *IEEE Transactions on Consumer Electronics*, vol. 59, no. 2, pp. 415-421.

Ballesteros, S. and Heller, M.A. (2006) *Conclusions: Touch and blindness*.

Barsky, S. and Petrou, M. (2003) 'The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. Pattern Analysis and Machine Intelligence', *IEEE Transactions on*, vol. 25, no. 10, pp. 1239-1252.

Bevan, N. (2001) 'International standards for HCI and usability', *International journal of human-computer studies*, vol. 55, no. 4, pp. 533-552.

BioID Technology Research *BioID Face Database*, [Online], Available: https://www.bioid.com/About/BioID-Face-Database [5 May 2013].

Blanz, V. and Vetter, T. (2003) 'Face recognition based on fitting a3D morphable model', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1063-1074.

Brewster, S. and Murray-Smith, R. (2000) ' Haptic Human-Computer Interaction', Proceedings of the first international workshop on haptic human-computer interaction, Glasgow.

Campadelli, P., Lanzarotti, R. and Lipori, G. (2006) 'Precise eye localization through a general-to-specific model definition', British Machine Vision Conference, 187-196.

Chang, C.-Y., Lange, B., Zhang, M., Koenig, S., Requejo, P., Somboon, N., Sawchuk, A. and Rizzo, A. (2012) 'Towards pervasive physical rehabilitation using Microsoft Kinect', 6th International Conference on Pervasive Computing Technologies for Healthcare, 159-162.

Chang, C.C. and Lin, C.J. (2011) 'LIBSVM: A library for support vector machines', *ASM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27.

Chen, W., Qian, W., Wu, G., Chen, W., Xian, B., Chen, X., Cao, Y., Green, C.D., Zhao, F., Tang, K. and Han, J.-D.J. (2015) 'Three-dimensional human facial morphologies as robust aging markers', *Cell Research*, vol. 25, pp. 574-587.

Coleman, E.N. and Jain, R. (1982) 'Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry', *Computer graphics and image* *processing*, vol. 18, no. 4, pp. 309-328.

Cristinacce, D., Cootes, T. and Scott, I. (2004) 'A Multi-Stage Approach to Facial Feature Detection', British Machine Vision Conference, 1-10.

Dalal, N. and Triggs, B. (2005) 'Histograms of oriented gradients for human detection', IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 886-893.

Daugman, J. (2004) 'How iris recognition works', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 21-30.

Dix, A., E.Finlay, J., Abowd, G.D. and Beale, R. (2004) *Human-Computer Interaction*, 3rd edition, Essex: PearsonEducation Limited.

Do, H., Kalousis, A., Wang, J. and Woznica, A. (2012) 'A metric learning perspective of SVM: on the relation of LMNN and SVM', International Conference on Artificial Intelligence and Statistics, 308-317.

Drewes, H., Luca, A.D. and Schmidt, A. (2007) 'Eye-gaze interaction for mobile phones', Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology , 364-371.

Drewes, H. and Schmidt, A. (2007) 'Interacting with the computer using gaze gestures', Human-Computer Interaction–INTERACT, 475-488.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012) *Pattern classification*, John Wiley & Sons.

Fagertun, J., Andersen, T., Hansen, T. and Paulsen, R.R. (2013) '3D gender recognition using cognitive modeling', 2013 International Workshop on Biometrics and Forensics, Lisbon, 1 - 4.

Fagertun, J., Andersen, T. and Paulsen, R.R. (2012) 'Gender recognition using cognitive modelling', Proceedings of Computer Vision–ECCV, 300-308.

Farahat, A. and Bailey, M.C. (2012) 'How effective is targeted advertising?', Proceedings of the 21st international conference on World Wide Web, 111-120.

Farooq, A.R., Smith, M.L., Smith, L.N. and Midha, S. (2005) 'Dynamic photometric stereo for on line quality control of ceramic tiles', *Computers in industry*, vol. 56, no. 8, pp. 918-934.

FDIs, I. (2009) *9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems (formerly known as 13407)*, Switzerland.

Frankot, R.T. and Chellappa, R. (1988) 'A method for enforcing integrability in shape from shading algorithms', vol. 10, no. 4, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 439-451.

Gallagher, A. and Chen, T. (2009) 'Understanding images of groups of people', IEEE Conference on Computer Vision and Pattern Recognition, 256-263.

GBTIMES (2015) *Bank replaces manager with cute singing robot* , 19 August, [Online], Available: http://gbtimes.com/china/bank-replaces-manager-cute-singing-robot [28 August 2015].

Georghiades, A.S., Belhumeur, P.N. and Kriegman, D.J. (2001) 'From few to many: Illumination cone models for face recognition under variable lighting and pose', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643-660.

Gökberk, B., İrfanoğlu, M.O. and Akarun, L. (2006) '3D shape-based face representation and feature extraction for face recognition',. *Image and Vision Computing*, vol. 24, no. 8, pp. 857-869.

Gupta, A. and O'Malley, M.K. (2006) 'Design of a haptic arm exoskeleton for training and rehabilitation', *IEEE/ASME Transactions on Mechatronics*, vol. 11, no. 3, pp. 280-289.

Hales, I.J., Williamson, D.R., Hansen, M.F., Broadbent, L. and Smith, M. (2015) 'Long-range concealed object detection through active covert illumination', SPIE Security+ Defence, 964806-964806.

Hamouz, M., Kittler, J., Kamarainen, J.K., Paalanen, P., Kalviainen, H. and J. Matas (2005) 'Feature-based affine-invariant localization of faces', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1490-1495.

Hansen, M.F. (2012) *3D face recognition using photometric stereo*, Bristol, UK: BMVA.

Hansen, M.F., Atkinson, G.A., Smith, L.N. and Smith, M.L. (2010) ' 3D face reconstructions from photometric stereo using near infrared and visible light', *Computer Vision and Image Understanding*, vol. 114, pp. 942-951.

Hassanpour, R., Wong, S. and Shahbahrami, A. (2008) 'Vision based hand gesture recognition for human computer interaction: A review', *International Conference Interfaces and Human Computer Interaction*, pp. 125-134.

Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D. (2012) 'RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments', *The International Journal of Robotics Research*, vol. 31, no. 5, pp. 647-663.

Huang, X., Acero, A., Alleva, F., Hwang, M.Y., Jiang, L. and Mahajan, M. (1993) 'Microsoft Windows highly intelligent speech recognizer: Whisper', Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing, 93-96.

Huang, G.B., Ramesh, M., Berg, T. and Learned-Miller, E. (2007) 'Labeled faces in the wild: A database for studying face recognition in unconstrained environments', *University of Massachusetts Technical Report 07-49*, pp. 1–11.

Hu, J., Hu, R., Wang, Z., Gong, Y. and Duan, M. (2013) 'Kinect depth map based enhancement for low light surveillance image', 20th IEEE International Conference on Image Processing, 1090-1094.

Hu, Y., Yan, J. and Shi, P. (2010) 'A fusion-based method for 3D facial gender classification', Proceedings of the 2nd International Conference on Computer and Automation Engineering, 369–372.

Huynh, T., Min, R. and Dugelay, J.L. (2012) ' An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data', Proceedings of Computer Vision–ECCV, 133-145.

Hyrskykari, A., Istance, H. and Vickers, S. (2012) 'Gaze gestures or dwell-based interaction?', Proceedings of the Symposium on Eye Tracking Research and Applications, 229-232.

I. Ullah, M.H.H.A.G.M.A.M.M.a.G.B. (2012) 'Gender recognition from face images with dyadic wavelet transform and local binary pattern', *Advances in Visual Computing*, pp. 409–419.

Iyer, G., Soberman, D. and Villas-Boas, J.M. (2005) 'The targeting of advertising', *Marketing Science*, vol. 24, no. 3, pp. 461-476.

Jaakkola, T. and Haussler, D. (1999) 'Exploiting generative models in discriminative classifiers', *Advances in neural information processing systems*, pp. 487-493.

Jesorsky, O., Kirchberg, K.J. and Frischholz, R.W. (2001) 'Robust face detection using the hausdorff distance', *Audio-and video-based biometric person authentication*, vol. 2091, pp. 90-95.

Kaminski, J.Y., Knaan, and Shavit, A. (2009) 'Kaminski, Jeremy Yrmeyahu, Dotan Knaan, and Adi Shavit. "Single image face orientation and gaze detection', *Machine Vision and Applications*, vol. 21, no. 1, pp. 85-98.

Kapoor, A., Burleson, W. and Picard, R.W. (2007) 'Automatic prediction of frustration', *International journal of human-computer studies*, vol. 65, no. 8, pp. 724–736.

Khoshelham, K. (2011) 'Accuracy analysis of kinect depth data', *Remote Sensing and*

Spatial Information Sciences, vol. 38, no. 5, pp. 133-138.

Khoshelham, K. and Elberink, S.O. (2012) ' Accuracy and resolution of kinect depth data for indoor mapping applications', *Sensors*, vol. 12, no. 2, pp. 1437-1454.

Koenderink, J.J. and Doorn, A.J.v. (1992) 'Surface shape and curvature scales', *Image and vision computing*, vol. 10, no. 8, pp. 557-564.

Kroon, B., Hanjalic, A. and Maas, S.M. (2008) ' Eye localization for face matching: is it always useful and under what conditions?', Proceedings of the 2008 international conference on Content-based image and video retrieval, 379-388.

Lai, K., Konrad, J. and Ishwar, P. (2012) 'A gesture-driven computer interface using Kinect', 2012 IEEE Southwest Symposium on Image Analysis and Interpretation , 185-188.

Lasinger, P. and Bauer, C. (2013) ' Situationalization, the New Road to Adaptive Digital-out-of-Home Advertising', Proceedings of IADIS International Conference e-Society, 162-169.

Lee, P.H., Huang, J.Y. and Huang, Y.P. (2010) 'Automatic gender recognition using fusion of facial strips', Proceedings of 20th IEEE International Conference on Pattern Recognition , 1140–1143.

Lee, J.R.J., Smith, M.L., Smith, L.N. and Midha, P.S. (2005) 'A mathematical morphology approach to image based 3D particle shape analysis', *Machine Vision and Applications*, vol. 16, no. 5, pp. 282-288.

Leo, M., Cazzato, D., Marco, T.D. and Distante, C. (2014) 'Unsupervised Eye Pupil Localization through Differential Geometry and Local Self-Similarity Matching', *PloS one*, vol. 9, no. 8.

Lichtenauer, J., Hendriks, E. and Reinders, M. (2005) 'Isophote properties as features for object detection', Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 649-654.

Lim, J., Ho, J., Yang, M.H. and Kriegman, D. (2005) 'Passive photometric stereo from motion', *IEEE International Conference on Computer Vision*, vol. 2, pp. 1635-1642.

Liu, C., Yuen, J. and Torralba, A. (2011) 'Sift flow: Dense correspondence across scenes and its applications', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978-994.

Mäkinen, E. and Raisamo, R. (2008) 'An experimental comparison of gender classification methods', *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1544-1556.

Mäkinen, E. and Raisamo, R. (2008) 'Evaluation of gender classification methods with automatically detected and aligned faces', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541-547.

Malzbender, T., Wilburn, B., Gelb, D. and Ambrisco, B. (2006) 'Surface Enhancement Using Real-time Photometric Stereo and Reflectance Transformation', *Rendering Techniques*.

McNamara, N. and Kirakowski, J. (2006) ' Functionality, usability, and user experience: three areas of concern', *interactions*, vol. 13, no. 6, pp. 26-28.

Mecca, R. and Durou, J.D. (2011) 'Unambiguous photometric stereo using two images', Image Analysis and Processing–ICIAP, Ravenna, 286-295.

Microsoft *How-Old.net*, [Online], Available: http://how-old.net/# [25 June 2015].

Moghaddam, B. and Yang, M.H. (2000) 'Gender classification with support vector machines', Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, 306-311.

Moore, D.M. (1994) *Visual literacy: A spectrum of visual learning*, Educational Technology.

Ng, C.B., Tay, Y.H. and Goi, B.M. (2012) 'Vision-based human gender recognition: A survey', *arXiv preprint* , vol. arXiv:1204.1611.

Niu, Z., Shan, S., Yan, S., Chen, X. and Gao, W. (2006) '2D cascaded Adaboost for eye localization', Proceedings of the 18th International Conference on Pattern Recognition, 1216-1219.

Ojala, T., Pietikäinen, M. and Mäenpää, T. (2000) 'Gray scale and rotation invariant texture classification with local binary patterns', Proceedings of Computer Vision–ECCV, 404-420.

Onn, R. and Bruckstein, A. (1990) 'Integrability disambiguates surface recovery in two-image photometric stereo', *International Journal of Computer Vision*, vol. 5, no. 1, pp. 105-113.

Oulasvirta, A. (2009) 'Field experiments in HCI: promises and challenges', in *Future Interaction Design II*, London: Springer London.

Pan, Y., Shen, P. and Shen, L. (2012) 'Speech emotion recognition using support vector machine', *International Journal of Smart Home*, vol. 6, no. 2, pp. 101-107.

Peres, S.C., Best, V., Brock, D., Shinn-Cunningham, B., Frauenberger, C., Hermann, T., Neuhoff, J.G., Nickerson, L.V. and Stockman, T. (2008) *HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces*, Burlington: Morgan Kaufmann.

Perronnin, F. and Dance, C. (2007) 'Fisher kernels on visual vocabularies for image categorization', Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

Perronnin, F., Sánchez, J. and Mensink, T. (2010) 'Improving the fisher kernel for large-scale image classification', 11th European Conference on Computer Vision, Crete, 143-156.

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J. and Worek, W. (2005) 'Overview of the face recognition grand challenge', Proceedings of the IEEE computer society on computer vision and pattern recognition, 947-954.

Phillips, P.J., Wechsler, H., Huang, J. and Rauss, P.J. (1988) 'The FERET database and evaluation procedure for face-recognition algorithms', *Image and vision computing*, vol. 16, pp. 295–306.

Phung, S.L. and Bouzerdoum, A. (2007) 'A pyramidal neural network for visual pattern recognition', *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 329-343.

Pieters, R., Warlop, L. and Wedel, M. (2002) ' Breaking through the clutter: Benefits of advertisement originality and familiarity for brand attention and memory', *Management Science*, vol. 48, pp. 765-781.

Rai, P. and Khanna, P. (2014) 'A gender classification system robust to occlusion using Gabor features based (2D) 2 PCA 25', *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1118-1129.

Rautaray, S.S. and Agrawal, A. (2015) 'Vision based hand gesture recognition for human computer interaction: a survey', *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1-54.

Reddy, V.R. and Chattopadhyay, T. (2014) 'Human activity recognition from kinect captured data using stick model', Human-Computer Interaction, Advanced Interaction Modalities and Techniques, 305-315.

Reynolds, D. (2009) 'Gaussian Mixture Models', *Encyclopedia of Biometrics*, pp. 659-663.

Rozado, D., Agustin, J.S., Rodriguez, F.B. and Varona, P. (2012) 'Gliding and saccadic gaze gesture recognition in real time', *ACM Transactions on Interactive Intelligent Systems*, vol. 10.

Rozado, D., Rodriguez, F.B. and Varona, P. (2012) 'Low cost remote gaze gesture recognition in real time', *Applied Soft Computing*, vol. 12, pp. 2072-2084.

Sánchez, J., Perronnin, F. and Campos, T.D. (2012) 'Modeling the spatial layout of images beyond spatial pyramids', *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2216-2223.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M. and Strope, B. (2010) 'Your Word is my Command”: Google Search by Voice: A Case Study', in Neustein, A. (ed.) *Advances in Speech Recognition*, Springer US.

Scharstein, D. and Szeliski, R. (2003) 'High-accuracy stereo depth maps using structured light', Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I-195.

Shan, C. (2010) 'Learning local features for age estimation on real-life faces', Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis, 23-28.

Shan, C. (2012) 'Learning local binary patterns for gender classification on real-world face images', *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431-437.

Shneiderman, B. (1992) *Designing the user interface: strategies for effective human-computer interaction.*, Reading: Addison-Wesley.

Siegert, I., Böck, R., Philippou-Hübner, D. and Wendemuth, A. (2012) 'Investigation of hierarchical classification for simultaneous gender and age recognition', *Young*, vol. 27, no. 28.

Simonyan, K., Parkhi, O., Vedaldi, A. and Zisserman, A. (2013) 'Fisher Vector Faces in the Wild', Proceedings of British Machine Vision Conference, 8.1–8.12.

Singh, K.R., Zaveri, M.A. and Raghuwanshi, M.M. (2010) 'Illumination and pose invariant face recognition: a technical review', *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 2, pp. 29-38.

Sledd, A. and O'Malley, M.K. (2006) ' Performance enhancement of a haptic arm exoskeleton', *IEEE 14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 375-381.

Smith, M.L. (1999) ' The analysis of surface texture using photometric stereo acquisition and gradient space domain mapping', *Image and vision computing*, vol. 17, no. 14, pp. 1009-1019.

Smith, M.L. and Smith, L.N. (2005) 'Dynamic photometric stereo—a new technique for moving surface analysis', *Image and Vision Computing*, vol. 23, no. 9, pp. 841-852.

Solomon, F. and Ikeuchi, K. (1996) 'Extracting the shape and roughness of specular lobe objects using four light photometric stereo', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 4, pp. 449-454.

Stalder, U. (2011) 'Digital Out-of-Home Media: Means and Effects of Digital Media in Public Space', *Pervasive Advertising*, pp. 31-56.

Sun, J., Smith, M., Farooq, A. and Smith, L. (2009) 'Concealed object perception and recognition using a photometric stereo strategy', *Advanced Concepts for Intelligent Vision Systems*, pp. 445-455.

Sutcliffe, A. (2006) 'Grand challenges in hci: the quest for theory-led design' Springer London.

Tawari, A., Chen, K.H. and Trived, M.M. (2014) 'Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation', Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems, 988-994.

Timm, F. and Barth, E. (2011) 'Accurate Eye Centre Localisation by Means of Gradients', International Conference on Computer Vision Theory and Applications, 125–130.

Tivive, F.H.C. and Bouzerdoum, A. (2006) 'A gender recognition system using shunting inhibitory convolutional neural networks', International Joint Conference on Neural

Networks, Vancouver, 5336-5341.

Türkan, M., Pardas, M. and Cetin, A.E. (2007) 'Human eye localization using edge projections', International Conference on Computer Vision Theory and Applications, Barcelona, 8-11.

Ullah, I., Hussain, M., Aboalsamh, H., Muhammad, G., Mirza, A.M. and Bebis, G. (2012) ' Gender recognition from face images with dyadic wavelet transform and local binary pattern', Advances in Visual Computing : 8th International Symposium, 409–419.

Valenti, R. and Gevers, T. (2008) 'Accurate eye center location and tracking using isophote curvature', IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 1-8.

Valenti, R. and Gevers, T. (2012) 'Accurate eye center location through invariant isocentric patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1785-1798.

Vedaldi, A. and Fulkerson, B. (2010) 'VLFeat: An open and portable library of computer vision algorithms', Proceedings of the international conference on Multimedia, 1469–1472.

Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S. and Cabeza, R. (2013) 'Hybrid method based on topography for robust detection of iris center and eye corners', *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 9, no. 4, p. 25.

Viola, P. and Jones, M.J. (2001) 'Rapid object detection using a boosted cascade of simple features', Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I-511-518.

Viola, P. and Jones, M.J. (2004) 'Robust real-time face detection', *International journal of computer vision*, vol. 57, no. 2, pp. 137-154.

Wachs, J.P., Stern, H.I., Edan, Y., Gillam, M., Handler, J., Feied, C. and Smith, M. (2008) 'A gesture-based tool for sterile browsing of radiology images', *Journal of the American Medical Informatics Association* , vol. 15, no. 9, pp. 321–323.

Wang, X. and Kambhamettu, C. (2013) 'Gender classification of depth images based on shape and texture analysis', Proceedings of Global Conference on Signal and Information Processing, 1077-1080.

Wang, J.G., Li, J., Yau, W.Y. and E. Sung (2010) ' Boosting dense SIFT descriptors and shape contexts of face images for gender recognition', *Proceedings of IEEE Computer Society Conference on CVPRW*, pp. 96–102.

Wang, J.G., Sung, E. and Venkateswarlu, R. (2005) 'Estimating the eye gaze from one eye', *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 83-103.

Wobbrock, J.O., Rubinstein, J., Sawyer, M.W. and Duchowski, A.T. (2008) 'Longitudinal evaluation of discrete consecutive gaze gestures for text entry', Proceedings of the 2008 symposium on Eye tracking research & applications, 11-18.

Woodham, R.J. (1980) 'Photometric method for determining surface orientation from multiple images', *Optical engineering*, vol. 19, no. 1, pp. 139-144.

Xia, B., Amor, B.B. and Daoudi, M. (2014) 'Exploring the Magnitude of Human Sexual Dimorphism in 3D Face Gender Classification', Proceedings of Computer Vision–ECCV 2014 Workshops , 697-710.

Xia, B., Amor, B.B., Huang, D., Daoudi, M., Wang, Y. and Drira, H. (2013) 'Enhancing gender classification by combining 3d and 2d face modalities', Proceedings of the 21st IEEE European Signal Processing Conference (EUSIPCO), 1-5.

Ylioinas, J., Hadid, A. and Pietikainen, M. (2012) 'Age classification in unconstrained conditions using LBP variants', 21st International Conference on Pattern Recognition (ICPR), 1257-1260.

Zafeiriou, S., Atkinson, G.A., Hansen, M.F., Smith, W., Argyriou, V., Petrou, M., Smith, M.L. and Smith, L.N. (2013) 'Face recognition and verification using photometric stereo: The photoface database and a comprehensive evaluation', *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 121-135.

Zhang, Z. (2012) 'Microsoft kinect sensor and its effect', *MultiMedia*, vol. 19, no. 2, pp. 4-10.

Zhang, X. and Gao, Y. (2009) 'Face recognition across pose: A review', *Pattern Recognition*, vol. 42, no. 11, pp. 2876-2896.

Zhao, W. and Chellappa, R. (2000) 'Illumination-insensitive face recognition using symmetric shape-from-shading', IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, 286-293.

Zhu, Z. and Ji, Q. (2005) 'Robust real-time eye detection and tracking under variable lighting conditions and various face orientations', *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 124–154.

Zhu, X. and Ramanan, D. (2012) ' Face detection, pose estimation, and landmark localization in the wild', IEEE Conference on Computer Vision and Pattern Recognition, 2879–2886.