Affective and Cognitive State Modelling within Human-Computer Interaction

Author
Samara, Anas

Affective and Cognitive State Modelling within Human-Computer Interaction

Anas Samara

School of Computing

This dissertation is submitted for the degree of
Doctor of Philosophy

Faculty of Computing, Engineering
and the Built Environment

December 2018

"Then Praise be to Allah, Lord of the heavens and Lord of the earth, Lord and Cherisher of all the Worlds!: To Allah be glory throughout the heavens and the earth: and Allah is Exalted in Power, Full of Wisdom!"

.

I would like to dedicate this thesis to my beloved parents, my sister, my brothers . . .
and to my lovely fiancée

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Anas Samara
December 2018

Acknowledgements

I have to to express my sincere gratitude and extend my thanks to my supervisors Dr Leo Galway, Dr Raymond Bond and Prof. Hui Wang, for their continuous support, remarkable cooperation, creative ideas and perfect supervision throughout my PhD course.
I would like to thank my PhD committee panel: Dr George Moore, Dr Paul McCullagh and Prof. Christopher Nugent for their follow-up, comments and beneficial feedback.
I acknowledge all my colleagues in the School of Computing, the Artificial Intelligence Research Group and the Smart Environments Research Group for their wonderful fellowship. Besides, I thank Miss Kate McMorris, Mr. Leonard O'Regan and Everyone who has helped me along the way of my PhD.
Lastly, I gratefully acknowledge the Vice Chancellor's Research Scholarships of Ulster University for the generous financial support of this PhD.

Abstract

There is an opportunity and necessity to enhance computer systems with automated intelligence in order to permit natural and reliable interaction similar to human-human interaction. The recent availability of unobtrusive input modalities, such as eye trackers and web cameras, has enabled the viable real-time detection of user's emotions and mental states. However, a key challenge is utilising such modalities to enable a computer to actively interact with users based on their emotions and mental states. Consequently, the aim of this PhD is to develop a user model that will be utilised to infer a user's emotional and cognitive state, which may be subsequently exploited to adapt the user experience in order to maximise the performance of the system and guarantee task completion.

An underlying framework for adaptive Human-Computer Interaction has been developed that comprises a Perception Component responsible for capturing and modelling affective and cognitive states, and an Adaptation Component, which drives the appropriate adaptation to the User Interface or User Experience. The research presented within this thesis primarily contributes to the Perception Component by probing the use of facial expressions as an input modality for computer systems. Additionally, eye-gaze tracking data has been investigated for assessing and modelling cognitive workload.

This work presented herein involves data collection from unobtrusive input modalities, as well as the development of machine learning algorithms to build user models from the collected data. Subsequently, user affective states and cognitive load have been modelled and investigated within various computer-based tasks. Moreover, the credibility of using facial expressions for modelling affective states and pupil size for modelling cognitive load has been explored and discussed. Intrinsically, pupil size variation can be used to model cognitive load during interaction with user interfaces, while facial expressions do not reflect the actual feelings of the user in that context.

Glossary

Active Interaction: refers to the case when a user interacts with a user interface, working and attempting to complete a task.
Adaptation Component: represents that motor that makes the adjustment and suitable changes into the user interface or the user experience.
Affective Computing: is the research domain that focus on studying and detecting Affective States of users.
Affective State: refers to the experience of feelings and emotional state of a user.
Area of Interest: refers to a sub region of the User Interface where most eye fixations are located on that particular boundary of the display.
Arousal: is the physiological and psychological state that activates the alertness, consciousness and attention as a reaction to stimuli.
Circumplex Model: a model that presents a human emotional state as a result of combining two dimensions which are Valence and Arousal.
Cognitive State: refers to an assessment of user's mental work load.
Cross Validation: is an evaluation technique that is commonly used to assess the model predictability by splitting dataset into a number of folds that one-fold used testing whilst the other folds used in training in each run, and then rotating this process over the folds and the results will be aggregated from the each run.
Eye-Gaze Tracking Data: refers to the data obtained from the infra-red eye tracker device.
Facial Expression: is a non-verbal representation of user emotional feeling, which originated from the movement of the facial muscles.
Facial Expression Recognition: is the process that attempts to automatically detect the valence emotional state through analysing features from the image of the face.
Fixation: represents the moment that a user eyes are relatively stationary on a specific object.
Human-Computer Interaction: involves the study, planning, design and uses of interfaces between computers and people.
Input Modality: is a channel through which a computer system acquires data about the user.
Passive Interaction: refers to the case when a user looking on a user interface without exerting effort.

Perception Component: represents the component that monitor human user's states whilst interacting with a machine.
Saccade: represents eye movements between fixations on the screen whilst interacting with a User Interface.
Ubiquitous Computing: refers to the era where computing becomes everywhere, on different devices, in which computer systems and user interfaces utilised within different aspects of daily life.
User Experience: refers to all users' aspects whilst interacting with a user interface.
User Interface: refers to the visible means where Human-Computer Interaction takes place.
User Model: refers to knowledge source of aspects of the user that are relevant to the system or related to a particular Human-Computer Interaction context.
Valence: is a representation of the intrinsic attractiveness or averseness of an emotion.

Abbreviations

ACT-R: Adaptive Control of Thought-Rational Model.
AU: Action Unit.
BDI: Belief Desire Intention Theory.
CKPLUS: Cohn-Kanade Plus dataset.
CLG: Command Language Grammar.
CMA: Circumplex Model of Affect.
CNN: Convolutional Neural Networks.
EOG: Electrooculography.
FACS: Facial Action Coding System.
GOMS: Goals Operators Methods Selection.
HAC: Human Action Cycle.
HCI: Human-Computer Interaction.
HPM: Human Processor Model.
HMM: Hidden Markov Model.
KDEF: Karolinska Directed Emotional Faces dataset.
KLM: Keystroke Level Model.
LSTM-RNN: Long Short-Term Memory.
NASA-TLX: NASA-Task Load Index Scale.
SAM: Self-Assessment Manikin Scale.
SOAR: State Operator and Result Model.
SVM: Support Vector Machine.
TEPR: Task-Evoked Pupillary Response.
UI: User Interface.
UX: User Experience.
WIMP: Windows, Icons, Menus, and Pointers Model.

Publications

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2017) *Affective state detection via facial expression analysis within a human-computer interaction context*. Journal Ambient Intelligence and Humanized Computing, 8 . pp. 1-10. [Journal article]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2018) *Adaptive User Experience Based on Detecting User Perplexity*. In: British HCI Conference 2018, Belfast, United Kingdom. ACM. 4 pp. [Conference contribution]

Samara, A, Galway, L, Bond, R and Wang, H (2017) *Human-Computer Interaction Task Classification via Visual-Based Input Modalities*. In: 11th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI) 2017, Villanova University, Philadelphia (Pennsylvania, USA). Spinger International Publishing AG. 6 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2017) *Tracking and Evaluation of Pupil Dilation via Facial Point Marker Analysis*. In: 1st International Workshop on Affective Computing in Biomedicine and Healthcare (ACBH 2017), Kansas City, MO, USA.. IEEE. 7 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2017) *HCIViewer - A Tool for Human-Computer Interaction Practitioners*. In: British HCI Conference 2017, Sunderland, United Kingdom. ACM. 4 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2017) *User Interaction Modelling for Adaptive Human Computer Interaction*. In: British HCI Conference 2017, Doctoral Consortium, Sunderland, United Kingdom. ACM. 4 pp. [Conference contribution]

Samara, Anas, Menezes, Maria Luiza Recena and Galway, Leo (2017) *Feature Extraction for Emotion Recognition and Modelling Using Neurophysiological Data*. In: 15th International Conference on Ubiquitous Computing and 8th Communications and Interna-

tional Symposium on Cyberspace Safety and Security, IUCC-CSS 2016, Granada, Spain, December 14-16, 2016, Granada, Spain. IEEE. 7 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2016) *Sensing Affective States using Facial Expression Analysis*. In: 10th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2016, Las Palmas, Gran Canaria. Springer. 12 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond R and Wang, Hui (2016) *Automatic Affect State Detection using Fiducial Points for Facial Expression Analysis*. In: Irish Human Computer Interaction Conference, Cork. iHCI. 1 pp. [Conference contribution]

Samara, Anas, Galway, Leo, Bond, Raymond and Wang, Hui (2015) *User Modelling for Adaptive Human-Computer Interaction.* In: Irish Human Computer Interaction, Dublin. iHCI. 1 pp. [Conference contribution]

Other Publications by the Author

Menezes, M. L. R., Samara, A, Galway, L, Sant'Anna, A, Verikas, A, Alonso-Fernandez, F, Wang, H and Bond, R (2017) *Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset.* Personal and Ubiquitous Computing, August . pp. 1-11. [Journal article]

Mendoza-Palechor, F., Menezes, M.L., Sant'Anna, A., Ortiz-Barrios, M., Samara, A. and Galway, L., 2018. Affective recognition from EEG signals: an integrated data-mining approach. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-20. [Journal article]

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Overview

As humans, our abilities and performance while carrying out different tasks differ due to several factors, such as health conditions, mental processing capabilities, emotional feelings, task nature and the desire to achieve it. Additionally, understanding human feelings and states that may entail information about achieving a certain task can be an intricate endeavour. Subsequently, this research aims to examine the feasibility and the possibility of making machines more perceptive devices that can recognise innate human factors, which have an impact in human's performance and effectiveness.

1.2 Studying Human-Computer Interaction

Human-Computer-Interaction (HCI) is defined by (Hewett et al., 1992) as *a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them*. The interaction styles between users and computers have passed through a number of different phases, starting from simple command line interfaces, to interfaces that adopt the now ubiquitous Windows, Icons, Menus, and Pointers (WIMP) model. The relationship between computers and humans has progressed through three main eras (Weiser and Brown, 1996), (1) the Mainframe Era, (2) the Personal Computer Era, and now (3) the Ubiquitous Computing Era. Within the Ubiquitous Computing Era, there has been a dramatic growth in the variety of computer systems and user interfaces utilised within different aspects of daily life (Dumas et al., 2009; Duric et al., 2002; Karray et al., 2008). Therefore, there is an opportunity and

a necessity to enhance computer interfaces with automated *intelligence* in order to permit interaction that is similar to natural human-human interaction (Harper et al., 2008).

Within the Ubiquitous Computing Era in which we currently live, there has been a dramatic growth in the variety of computer systems and user interfaces utilised within different aspects of daily life (Dumas et al., 2009). Therefore, a lot of effort is made towards enhancing the interaction quality between the human and the computer, which attempts to determine the most feasible and intuitive form of interaction, especially when humans deem computers as *social agents* (Pantic and Rothkrantz, 2003). Consequently, a range of technologies have been explored within the research literature to endow computers with the ability to perceive more about the users, such as measuring different states of the user including stress and mental workload (Ahn and Picard, 2014; Hariharan and Philipp Adam, 2015; McDuff et al., 2014; Neoh et al., 2015).

Consequently, many HCI researchers utilised different methodologies as an attempt to improve the interaction between humans and computer systems, and they suggest a range of solutions to the problem resulting from the gap between the core language carried out by the computer and the task language carried out by the user, i.e. the difference between the user's formulation of actions to achieve a goal and the actions permitted by the system (Dix et al., 2004; Kirakowski and Corbett, 1990; Saffer, 2009).

1.3 Basic Components of HCI Study

HCI research is typically focused on producing interaction models that are characterised by *consistency, observability, predictability, reachability, visibility, simplicity, flexibility, learnability, operability, usability* and *accessibility* (Dix et al., 2004; Kirakowski and Corbett, 1990; Stephanidis, 2001). The HCI domain attempts to design, build and improve interaction between humans and computers. Moreover, HCI is technology driven with the aim of deploying new interaction paradigms to reduce the barrier between the users' tasks and the capabilities of the computer. This section discusses the basic components for any HCI study, which include: *(1) User Model; (2) Application Model; (3) Interaction Model*.

1. The *User Model* represents a knowledge source that contains explicit assumptions on all aspects of the user that are relevant to a system behaviour (Sullivan and Tyler, 1991). Furthermore, it helps the system to predict the characteristics that are associated with a user (Benyon and Murray, 1993). In a broader perspective, there exists a wide range of generic user characteristics, such as language, culture, demography, preference, emotion and many others, that can be utilised to model a user.

2. The *Application Model* represents an abstraction that defines the functional aspects of the system, and the associated operations (Benyon and Murray, 1993). Furthermore, it defines the logical activities that help users to achieve their goals. This model should describe the actual application in terms of its objects, corresponding attributes and associated relationships.

3. The *Interaction Model* provides a representation for the designated interaction between a user and an application. This is typically based on one of the most famous models used by HCI researchers, i.e. the Human Action Cycle (HAC) which was suggested by Donald Norman (Norman, 2002). The benefit of such modelling is to help designers to understand complex human behaviour (Dix et al., 2004). In terms of the HAC, there exists four basic parts to any action (Norman, 2002):

   (a) The Goal Formation, in which the user formulates what he or she wants to do, framed in terms of the domain.

   (b) The Execution Stage, in which the user formulates an intention on how to perform the action.

   (c) The Evaluation Stage, in which the user validates the achieved results against the original goals.

   (d) The Real World where the objects are manipulated by the user.

Consequently, this research project focuses on the User Model because it represents a key for intuitive and intelligent interaction, as it endows machines to perceive and understand humans or at least be aware of relevant human aspects that influence computer-based task completion.

1.4 Research Aim and Objectives

The overarching aim of the thesis is to examine the exploitation of visual-based input channels, particularly webcam and eye tracker, for detecting and recognising spontaneous user feelings and experiences. Subsequently, the investigations and experiments for given studies conducted throughout the presented herein, provide a set of techniques that facilitate the generation of a User Model that represents the *affective* and *cognitive* states of the users interacting with a User Interface (UI).

In order to support the overarching aim of the thesis, a number of studies were carried out, including studies that examine and explore facial-based features for recognising facial

expressions and associated user affective states. Subsequently, these studies validated the relationship between facial expressions and affective states in HCI contexts. Therefore, a review has been carried out of the Affective Computing literature to explore the appropriate set of emotional states that should be used, as well as to examine the most feasible methods that can be used for feature extraction and representation for facial expression analysis together with the effective machine learning techniques for facial expression recognition and detection. Moreover, a Data Collection Study has been developed and implemented to collect relevant features via different channels to be used for analysis and validation. Alternatively, studies were conducted that investigated eye tracking data in relation to actual cognitive states of the users, whilst interacting with different computer-based tasks. In other words, checking the pupil size dilation (i.e.variation) and response according to the cognitive states within HCI contexts. In addition, the relationship between facial-based features and eye tracking-based features was also examined using correlation tests,likewise the association between the affective states and cognitive states.

The studies presented in this thesis went through a number of stages, whereby each stage facilitates the establishment of a component for a framework for adaptive HCI, and also each component may contribute to relevant research areas such as machine learning, feature extraction and representation, statistical analysis, data collection, self reporting assessment tools and exploitation of visual-based input channels. Additionally, the outputs of this research including feature extraction and representation, analysis approaches, and generated dataset, would be beneficial for both students and researchers who work within HCI research themes as well as other disciplines related to *Machine Learning, Affective Computing, Cognitive Computing* and *User-Centered Design*.

Subsequently, the purpose of the studies is to test the behaviour of the facial expressions and eye gaze behaviour within different HCI contexts. Accordingly, the User Model symbolises the vital role of adaptive HCI, which leads to the intelligent systems that are able to select the appropriate way to make the suitable adaptation according to the application domain, in order to achieve the an intelligent adaptive form of HCI.

1.5 Approach and Methodology

The general strategy followed within the research presented herein is to investigate the relationships between different types of data, which was acquired within the context of Human-Computer Interaction, specifically data relating to user interactions during common computer-based tasks that represent the most usage by ordinary user. Subsequently, the multi-modal dataset collected during the PhD is composed of visual-based recordings, which

are webcam and infra-red eye tracker frames, and associated self-reported scores that entail information about the tasks, user emotional states and cognitive load i.e. mental processing.

The data can be viewed as two categories. First category is the subjective self-reporting, which includes: (1) task questionnaire composed of three questions about knowledge in the task and relevant user competence, and difficulty level of the task; (2) a user state questionnaire using Self Assessment Manikin (SAM) scales responses that are three sub-scales about *Valence*, *Arousal* and *Dominance*; (3) a task load questionnaire using NASA Load Index (NASA-TLX) scores that composed of six sub-scales. The second category represents the electronically populated features from visual-based input channels, which are facial-based features and eye gaze-based features. On one hand, facial-based features that are point-based features, which are points in the Cartesian coordinate space, and referred to as *fiducial points*. Additionally, distance-based features that were calculated from extracted fiducial points. On the other hand, eye gaze-based features, including different statistical functions that were used for features generation from spontaneous samples features to descriptive features, where each one of them represents a set of values captured within a particular window of time.

Consequently, data collection protocol has been designed and implemented to populate and prepare a dataset that facilitates the investigations and helps out in examining the hypothesis derived out to answer the research questions. Additionally, a number of classification experiments and statistical analysis have been conducted and applied on benchmark datasets as well as applied on the collected dataset.

1.6 Research Questions

At the outset of the research, the following research questions were generated that the various studies aimed to answer:

1. *What are the emotional states that can be detected and utilised for the purpose of modelling the user within an adaptive HCI context?*

2. *What is the best method that can be used as input channel to capture the user state that would be useful for adaptive HCI?*

3. *What are the most efficient machine learning and statistical approaches to modelling user states within an adaptive HCI context?*

1.7 Key Research Contributions

Despite the fact that detecting and recognising affective and emotional states via facial expression analysis has been investigated, still it is considered a challenging task in terms of accuracy versus the quality of the images within different real world conditions. Subsequently, one of the outcomes of current research project contributes to the advancements in facial expression recognition process. The yielded approach facilitates achievement of a robust and highly accurate recognition model, specifically in terms of feature representation, whereby a novel feature representation based on Euclidean distance-based features have been generated from facial point markers together with exploiting these features for facial expression recognition as given in Chapter 4. Moreover, the design and development of an ensemble-based hierarchical machine learning technique that decomposes the classification decision into micro-decisions made by binary classifiers. Therefore, a novel classification technique that is a hierarchical structure of a set of classifiers working together have been implemented and validated, as presented in Chapter 4 is able to provide a classification mechanism that is extensible over a number affective states, taking into account classification accuracy.

Additionally, modelling user affective states within different HCI contexts has not been explored in the literature sufficiently. Consequently, user states and feeling during typical HCI contexts have been explored and identified within the research presented herein. Validated classification techniques and feature extraction methods were applied on the collected dataset in order to automatically classify emotions via facial expressions by analysing video frames that have been acquired whilst users attempted and interacted with each of the computer-based tasks. In Chapter 4, conclusions and interesting findings on the relationship between facial expression and UIs under various tasks have been drawn as well as the relationship with self-subjective reports. Consequently, the aforementioned studies demanded an appropriate dataset that provides a representation of typical HCI contexts. Therefore, multi-modal feature sets have been populated through a Data Collection Study given in Chapter 3, where features from different input modalities were extracted and collected, which can be used to reason about users' affective and cognitive states whilst interacting with common computer software and attempting to complete typical computer-based tasks.

Furthermore, additional studies were conducted that exploited visual-based features, which were captured from a webcam and an eye-tracker for other uses. For instance, computer-based task classification through exploiting facial-based features. As presented in Chapter 6, classification experiments revealed the possibility to distinguish between different computer tasks through means of facial expressions. Moreover, a novel approach to track pupil size variation through features based on facial points markers was presented. The

relationship between pupil size and features of the eye boundary area were investigated, whereby correlation analysis between eye-tracking data and features populated from facial expression analysis showed the possibility to track pupil size variation through features populated from a webcam frames of the eye area as given in Chapter 6.

1.8 Thesis Document Structure

The current thesis document consists of eight Chapters and four Appendices. Chapter 1 provides an introduction to the research area and briefly describes the motivation for the work, an overview of the methodology, the research questions together with key contributions achievements.

A review of the literature is given in Chapter 2, where a background is given on the research approaches that handled HCI models, in addition to a review of related research conducted on user modelling. Moreover, the input channels that can be used for modelling user states, machine learning methods and state-of-the-art techniques and assessment tools have been reviewed and discussed. Further details on the conducted literature review have been given in a tabular form in Appendix A and Appendix B.

Chapter 3 provides a description of the Data Collection Study carried out in the current project, which includes the design, implementation and data exploitation schemes.

Chapter 4 presents the work of facial expression analysis, which contains the details of facial-based data feature extraction methods across several classification techniques along with the developed hierarchical machine learning technique. Moreover, the use of pre-trained models and the use of recent image classification techniques for facial expression detection have been addressed.

Chapter 5 and Chapter 6 focused on eye-tracking data. On the one hand, Chapter 5 provides analysis of several techniques utilising self-reported scores against eye-tracking data, notably the NASA-Task Load Index and the Self-Assessment Manikin Scales. On the other hand, Chapter 6 presents an experiment where both of facial-based features are combined with eye gaze-based features and correlation analysis has been conducted between the both.

Finally, Chapter 7 outlines the main aspects of an adaptive form of HCI, considering the major facets of affective and cognitive state perception of users, and identifies and outlines the components for an adaptive HCI framework. Moreover, it concludes the thesis with the main items of the work along with the main findings derived from the experiments, such as the relationship between facial expressions and actual affective states, the correlation between pupil dilation and the cognitive load, the association between facial-based features captured

from a webcam and eye-gaze features captured from an infra-red camera. In addition, suggestions for future work in the domain of Affective Computing or related research areas are presented.

Chapter 2

Literature Review

2.1 Overview

This chapter presents a review of the literature into the work done on the related topics and relevant areas, which are: HCI Research Approaches, HCI Models, Human Factors and Cognitive Models, User Modelling and Adaptive HCI, and Input Perception Modalities. In addition, usage of Visual-Based Data has been discussed including Facial Expression and Eye Gaze-Based Tracking Data, along with key aspects of the state-of-the-art on Machine Learning and Classification Methods. Moreover, the review discussed Subjective Assessment Tools that can be used in order to gather information from user's experiences and feelings. It concludes with a synthesised summary of the research literature, identifying the research gaps.

2.2 Human-Computer Interaction Research Approaches

Human-Computer Interaction (HCI) research aspects can be handled at different levels: (1) *Physical Aspects*, which investigate the mechanics of interactions between computer and human in which the mechanics of interactions are investigated (Chapanis, 1965); (2) *Cognitive Aspects*, which are concerned with the design, analysis and evaluation of complex systems to maximise the usability of the interface (Hollender et al., 2010); (3) *Affective Aspects*, which endeavour to provide the computer with perceptions and inferences of the user's emotions and attitudes (Picard, 2000a). With regard to this aspect, emotion perception is considered to be an active part of intelligence (Mavrikis et al., 2015; Picard, 2003; Sebe, 2009). According to Picard, computer systems need to have logical reasoning abilities about the user in order to interact intelligently (Picard, 2000a). In other words, it should recognise

the user's affective expressions, in particular, indications of frustration, fear, or dislike, and respond intelligently (Picard, 2000a, 2003). Subsequently, the foundations of HCI research up to the current progression on this area will be elaborately discussed in the following section.

2.3 Traditional HCI Models

The progression of HCI models is ongoing as technology keeps moving forward and new interfaces, needs, activities and uses arise. In early HCI research, HCI models were generated that utilised a formal grammatical description as a predictive tool to compare alternative designs for user interface ease of use (Reisner, 1981). This tool was employed to identify design inconsistencies from a human point of view during the design cycle. However, this type of model is particularly intended for determining a user's knowledge and competence, rather than measuring a user's performance during interaction. In addition, Moran (Moran, 1981) introduced a HCI model entitled Command Language Grammar (CLG) as a framework that represents user interface aspects at four levels of abstraction: *task level, semantic level, syntactic level*, and *interaction level*, where each level describes the actions that existed in the system in order to accomplish tasks throughout the whole system, from different points of view according to the level of abstraction. Nevertheless, this model does not consider any aspects related to the user's capabilities, limitations and situational context.

Regardless, this model paved the way for the Human Processor Model (HPM), which is one of the first HCI models that takes human aspects into consideration (Card et al., 1983). Subsequently, the HPM model aims to understand and measure user's performance within a HCI context according to the user's mental processing and memorisation capabilities. Utilising this perspective, the Goals Operators Methods Selection (GOMS) architecture was introduced in 1983, which comprised four primary components: *(1) Goals; (2) Operators; (3) Methods; (4) Selection* (Biswas et al., 2012). For example, the GOMS model defines the usability approach for accomplishing a *Goal* using available *Methods*, which are composed of *Operators* that take into account *Selections* (rules) in order to choose the proper way available in the system to achieve that *Goal* (Card et al., 1983). Subsequently, GOMS permits usability designers to break tasks into sub-tasks and goals in order to define the actions to be undertaken to achieve a desired target goal (John and Kieras, 1996). Consequently, the GOMS model is deemed as an important milestone in the history of HCI modelling as it introduces and utilises the perspective of the users during interaction. Moreover, it is considered a template for other models that use the same approach, albeit with different activities and levels of complexity.

Examples of the GOMS family of HCI models include the Keystroke Level Model (Card, Moran and Newell, 1980), Natural GOMS Language (Kieras, 1994), Cognitive Perceptual Motor GOMS (John et al., 2002). Such GOMS-based models focus on the procedures and operators that users employ in order to achieve their goals during interaction with a computer. Furthermore, GOMS-based models can be used to measure system aspects, such as procedure and operation speed and complexity. In particular, the GOMS family of models are efficient for modelling optimal behaviour as a fixed plan before interaction begins, i.e. the goals and plans are determined prior to execution. However, the assumption of following an optimal plan cannot be guaranteed in most real-world interaction scenarios (Biswas et al., 2012). Furthermore, the GOMS family of models do not address human aspects such as mental workload, user preferences, habits, or physical abilities (Sharp et al., 2011).

2.4 Human Factors and Cognitive Models

An initial attempt at generating models that address human behaviour is the State Operator And Result (SOAR) model proposed in (Laird, 1987). SOAR models human cognition as a rule-based system that exploits a learning technique to convert sequences of operations into production rules that will be utilised in situations where a user is not able to complete a task due to insufficient knowledge, referred to as an *impasse state*. Subsequently, the production rules can be employed in similar situations through the use of a Chunking Mechanism (Newell, 1992). In a similar manner, Bratman (Bratman, 1987) described the Belief Desire Intention (BDI) theory of human practical reasoning, which discussed each component of this architecture along with the underpinning rationale. Furthermore, the BDI theory captures the main components of practical reasoning, including intention handling, execution, option generation and deliberation. However, as the associated BDI architecture describes a higher level of abstraction, it is not necessarily practical for use in rational reasoning systems (Rao and Georgeff, 1995). Regardless, BDI theory inspired the development of other models and architectures such as the BDI software architecture presented in (Rao and Georgeff, 1995), which models an intelligent agent in terms of Beliefs, Desires and Intentions. However, the BDI software architecture does not consider inter-agent interaction, and does not support the integration of agent learning within multi-agent systems (Georgeff et al., 1998).

Besides these models, there are models known as Cognitive Models, which are utilised to investigate human behaviour during interaction with computers, particularly the prediction and simulation of mental processes during a task (Biswas and Robinson, 2010). In (Anderson, 1993), Anderson presented the Adaptive Control of Thought-Rational (ACT-R) model that helps to derive assumptions about human cognition. ACT-R describes how humans apply

knowledge to solve problems and achieve their goals. Moreover, it shows that human memory is divided into *declarative* and *procedural* memory. On the one hand, *declarative* memory stores the goals and related facts, whereas, on the other hand, *procedural* memory includes procedures and related rules for operation.

In fact there are many models and architectures designated for diverse forms of interactions within various environments with different types of constraints, such as the Cognitive Architecture for Computational Modelling of Human Performance (EPIC) (Kieras and Meyer, 1994), the tool that supports reasoning about behaviour (CORE) (Howes et al., 2001), the methodology framework for modelling human performance (APEX) (Freed et al., 1999), the computational cognitive architecture entitled Connectionist Learning with Adaptive Rule Induction On-line (CLARION) (Sun, 2006) and the Man-Machine Integration Design and Analysis System (MIDAS) (Gore, 2011). Table A.1 in Appendix A presents an comparative overview of a number of models surveyed within the literature from a range of disciplines that are potentially relevant to HCI.

2.5 Adaptive Human-Computer Interaction

Generally speaking, *adaptability* can be used to refer to systems that enable users to choose different, preferential ways to accomplish tasks (Kirakowski and Corbett, 1990). Furthermore, *adaptation* is used in many disciplines such as behaviour adaptation in human-human interaction, process and device interaction in multi-agent systems, as well as adaptation in intelligent HCI. Consequently, enhancing HCI to be *adaptive* and *intelligent* empowers a solid integration between humans and the computerised resources, resulting in the achievement of effective, efficient, comfortable, and reliable HCI (Harper et al., 2008). In this context, another depiction for adaptation is the system that guarantees reliable system functioning through error-free or error-tolerating operation, and provides system state diagnosis to support human decision making (Balint, 1995). Moreover, in the same paper, Balint presented different possibilities for adaptation and suggested categorisation according to type, applicability, control and usability (Balint, 1995).

Therefore, within the research literature, different approaches such as giving help adjuncts, changing tasks organisation, or modifying the interface and others have been suggested for adaptation, due to the dependency between the adaptation approach utilised and the domain in which the software is being employed. Subsequently, Cheng et al. presented a software engineering study for self-adaptive systems; they identified four sets of dimensions that should be modelled in a way such that each dimension represents a specific aspect of the self-adaptive systems (Cheng et al., 2009): (1) Goals and Objective-based dimensions

of the system, which are: *Evolution, Flexibility, Duration, Multiplicity* and *Dependency*. (2) Changes, which refer to causes of adaptation, including *Source, Type, Frequency* and *Anticipation*. (3) Mechanisms that show adaptation process aspects, which include *Type, Autonomy, Organization, Scope, Duration, Time-lines* and *Triggering*. (4) Effects, which identify the impact of adaptation on the system, which includes *Criticality, Predictability, Overhead* and *Resilience*.

Additionally, Qureshi and Perini stated that the main requirement for an adaptive system is by defining explicit alternatives in the operations used to achieve the goals; in other words, adaptive systems should include variability in both system operations and behaviours, and evaluation criteria (Qureshi and Perini, 2009).

2.6 User Affective States Modelling

A User-Model is defined by Benyon and Murray as *"a representation of the knowledge and preferences which the system believes that a user possesses"* (Benyon and Murray, 1993). Arguably, the aforementioned HCI models cannot be regarded as User Models, Interaction Models, or Application Models according to the basic structure of HCI given in the Introduction Chapter, due to the links and Interrelationships between characteristics and aspects in the components of these models. Therefore, it is difficult to investigate the User Model apart from Interaction Model and Application Model within a HCI context, because the User Model is intrinsically linked to the other two models (Biswas and Robinson, 2010; Biswas et al., 2012; Oviatt, 2003). Nevertheless, this section discusses methods used to model user characteristics, specifically those aspects related to both the affective state and the cognitive states of the user. Moreover, the discussion will focus on the approaches, techniques, assessment methods and input perception modalities used to detect and model these characteristics, as specified within the research literature.

In the context of HCI, the detection of user emotion could be just as important for adaptation and intelligent interaction with the next generation of machines (Picard, 2000a). Consequently, several studies focused on detecting different sets of emotions (Barakova et al., 2015; Busso et al., 2004; Ekman, 2003; Hariharan and Philipp Adam, 2015; Kneer et al., 2016; Mohammad et al., 2015; Neoh et al., 2015; Shan et al., 2007). However, within the domain of Affective Computing, there is no agreement on a definitive set of human emotions to recognise. Hence, several research efforts have defined different sets of emotion labels that are domain and application specific. Table B.1 in Appendix B provides a summary of user modelling studies surveyed within the Affective Computing literature, which outlines the sets of emotions that have been defined and detected within the existing research studies

and provides relevant details of each of the studies, such as the context and information sources used to infer and predict each kind of emotion. For example, Ekman and Friesen (Ekman and Friesen, 1971) revealed six emotions that can be inferred from facial expressions, including *surprise, fear, happiness, sadness, anger* and *disgust*. Whereas in Shan et al. (Shan et al., 2007), a different set of emotions were detected from body gestures, including *anger, anxiety, boredom, disgust, joy, puzzlement* and *surprise*. By contrast, in Busso et al. (Busso et al., 2004), only four emotions were identified, which included *sadness, anger* and *happiness*, along with a *neutral state*. Subsequently, 102 labels for different affective states were compiled from 46 studies, and can be visualised in the word cloud given in Figure 2.1.

Fig. 2.1 Word cloud for used labels of different affective states compiled from 46 studies, where the size of the label is related to the number of occurrences of that label across the surveyed studies.

(Russell, 1980) used an alternative endeavour to model emotional states. Which is considered a renowned model of affective states, entitled the Circumplex Model of Affect (CMA) given in Figure 2.2, which states that instead of defining labels as discrete emotions, affective state is a result of two neurophysiological dimensions: (1) the *Pleasant-Unpleasant* continuum, known as the *Valence axis*; (2) the *Activation-Deactivation* continuum, known as the *Arousal axis* (Posner et al., 2005; Russell, 1980).

Furthermore, in (Soleymani et al., 2012) three classes were defined for both *arousal* and *valence*, with the *arousal classes* containing the states *calm, medium aroused* and *activated*, whereas the *valence classes* contained the states *unpleasant, neutral* and *pleasant*. Other studies, such as the work carried out in (Joho et al., 2009), used two approaches to model

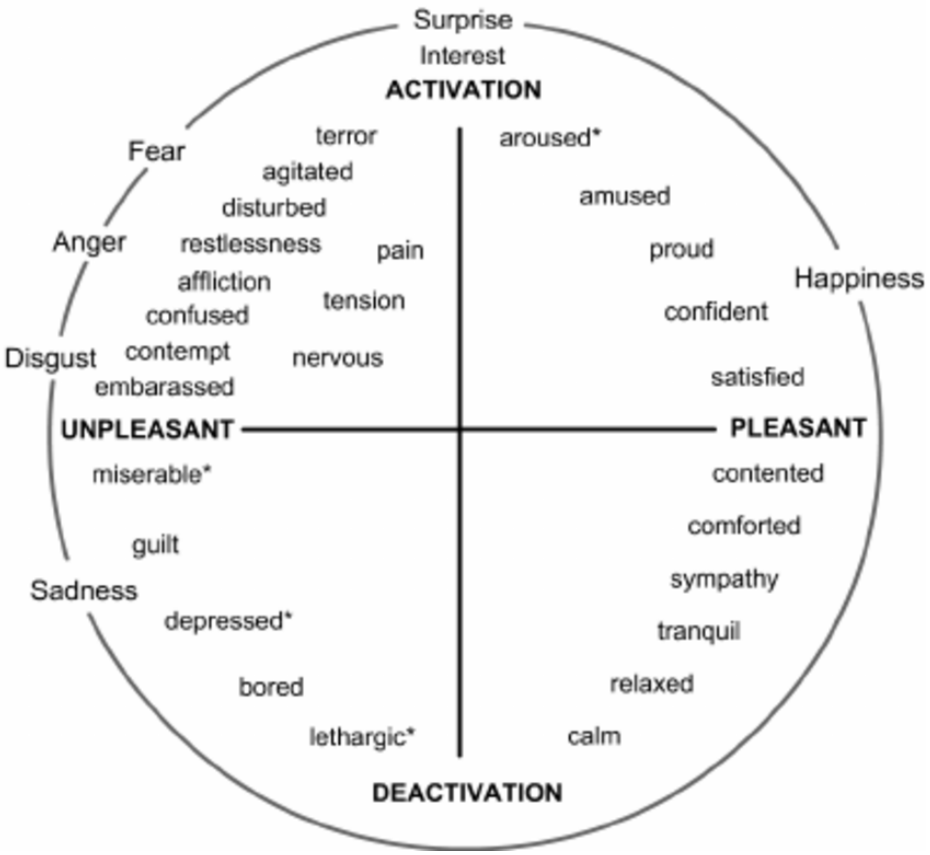


Fig. 2.2 Valence (*pleasant-unpleasant* continuum) versus Arousal (*activation-deactivation* continuum). The two dimensional spaces of the Circumplex Model (Russell and Lemay, 2000).

affective expressions; the first model, referred to as *Pronounce Level*, uses three categories to classify facial expressions according to how much they are pronounced: *No (Neutral), Low (Angry, Disgust, Fear, and Sad)* and *High (Happy and Surprise)*. The second model, referred to as the *Facial Expression Change Rate*, represents how often the detected facial expressions changed from one category to another, which subsequently indicated the impact on affect of the content of a video clip.

From the aforementioned literature and as provided in Table B.1 in Appendix B, the user modelling process is changeable and diverse, different emotional states and labels are addressed and defined, therefore, user modelling process can be handled in different ways according to the application target and environment in which the application is used.

User-Model for Adaptive HCI

Affective state detection in the context of HCI is a key for intelligent adaptation, which subsequently generates an interaction style that reasons in context with the user's goals, attitudes, plans and capabilities (Picard, 2003; Sullivan and Tyler, 1991). Also, achieving an intelligent and adaptive form of HCI depends on the software application (i.e. the user interface (UI)) and the domain in which the software is being employed, together with other constraints related to the UI functions, and both task characteristics and requirements (Jameson, 2003; Saffer, 2009). Subsequently, the system becomes able to provide error-tolerant operations as well as providing system state diagnosis to support human decision making (Balint, 1995). Therefore, a more robust integration between humans and computers may be obtained, resulting in the achievement of an effective, efficient, comfortable and reliable HCI, where the systems have the ability to make suitable adaptations in relation to contextual information about both the task and the user's affective and cognitive states (Karray et al., 2008; Oviatt, 2003).

Affective State Perception

Generally speaking, *Affect* as a concept describes the feeling that a human experiences whilst progressing in everyday activities (Stangor et al., 2014). Additionally, *Affect* is a broad term that can be related to emotions, mood and any reaction yielded instinctively due to a response caused by specific incident or stimuli (Stangor et al., 2014; Zajonc, 1980).

On the other hand, *Affective State* (i.e. emotion) represents a composition of more complicated psychological and physiological constructs (Harmon-Jones et al., 2012). *Affective State* is associated with a perceptible biological changes, which causes the human to feel it, as well as the opportunity of the appearance of such changes that make them apparent for other

humans so that they might observe the feeling that the person is experiencing (Davidson, 1992). Subsequently, Oatley and Johnson-Laird described affective states as cognitive states that coordinate nervous system processes, which make the biological changes to achieve certain goals through transition between different plans, together with communicating and conveying these goals to others (Oatley and Johnson-Laird, 1987).

Affective states have been studied by many cognitive psychologists because of the cause-effect relationship between emotional state and other cognitive processes, such as working memory, attention and reasoning operations (Harmon-Jones et al., 2012; Oatley and Johnson-Laird, 1987; Zajonc, 1980). User's *Affective State* has a profound impact on the cognitive processes, brain activation and attention, for example the *Happy* emotional state strengthens cognition and broadens the cognitive scope, whereas the *Anxiety* emotional state has the opposite effect (Harmon-Jones et al., 2012; Oatley and Johnson-Laird, 1987).

Subsequently, several theories has been developed by psychologists to explain the relationship between *Affective State* and cognitive abilities, in spite of the fact that many of these theories do not address the type of a particular *Affective State* within a specific context such as critical thinking or learning (Graesser et al., 2008). For instance, Frijda and Parrott proposed the *"Action-Readiness Theory of Emotions"* that describes what they call *UR-Emotions* - that are multicomponential emotions - a set of stimulus-response pairs, which are universal and based on the biological changes and represents states of readiness for specific actions (Frijda and Parrott, 2011).

Additionally, the most common employed theory for *Affective State* is the *"Core-Affect Theory of Emotions"* introduced by (Russell, 2003) and elaborated on by (Yik et al., 2011). This theory advises that emotion is a state resulting from core affect and comprises a set of underlying dimensions composed of a *Valence* component that describes the pleasure level, and an *Arousal* component that refers to the agitation intensity and the activation level (Russell, 2003; Harmon-Jones et al., 2011).

Cognitive State Perception

According to (Sweller, 1988) the *Cognitive Load Theory* states that there are three types of cognitive load:

1. *Extraneous Cognitive Load*, which is related to the load occurred because of the way of presenting the instructions and material. Thus, this type of cognitive load is more relevant to HCI and it can be managed and controlled by the designers in terms of the UX design (Chandler and Sweller, 1991). Therefore, the objective should be to avoid extraneous cognitive load.

2. *Intrinsic Cognitive Load*, which is related to the difficulty level associated with the instructions, which subsequently related to relevant information and experience level of the person for that particular task (Bannert, 2002).

3. *Germane Cognitive Load*, which is related to the processing, construction and automation behaviour (i.e. schemata) that organises categories of information, and mental structure of the thoughts, and the relationships among them (Sweller et al., 1998). The difference between this type and the *Extraneous* and *Intrinsic* is that Germaine was suggested to explain the effect of material presentation variability, such as the problem solving situations, where the learner (i.e user) formulates a schemata to identify variability and distinguishes between relevant and irrelevant features for that particular task.

Regardless the cause of the cognitive load, it is important to maintain the amount of cognitive load within the total cognitive capacity in order to avoid cognitive ageing, which results from working-memory capacity reduction due to irrelevant information. This consequently slows down mental processes, and thus affects overall performance (Van Merriënboer and Sweller, 2005). Therefore, measuring the amount of cognitive load in the HCI context would be a practical and beneficial pillar for the *Perception Component* and the achievement of adaptive form of HCI.

2.6.1 Assessment Tools of User's States

User modelling involves the development of tools for the assessment, evaluation and prediction of user's feelings and mental workload (Hart, 2006; Saneiro et al., 2014). Researchers seek to increase user performance through effective, efficient, satisfactory, safe and conformable ergonomics (Rubio et al., 2004). Principally, there are three categories of mental workload evaluation tools according to (Rubio et al., 2004): (1) *Performance-based measures* that rely upon the hypothesis that increasing task difficulty leads to an increase in demands and eventually decreases performance. (2) *Physiological measures*, such as eye activity, cardiac activity, respiratory activity, speech activity, and brain activity (Miller, 2001), which utilise changes in the physiological condition of the user according to the basis presented in (Johanssen et al., 1979) that mental workload affects human activity and causes a physical response. (3) *Subjective measures* based on self-rated assessment scales that are completed by users, which indicate their thoughts and impressions about tasks performed. Despite the subjective nature of these types of metrics, they are considered to be efficient, successful and are widely used due to their flexibility, ease of use and non-intrusive nature (Rubio et al., 2004).

Subjective Measures of Cognitive Workload

Interestingly, subjective measures could potentially provide results as precise as those obtained by physiological measurements, as discussed in (TATTERSALL and FOORD, 1996). Correspondingly, the most popular subjective measurement tools are: Subjective Workload Assessment Technique (SWAT), developed by Reid and Nygren (Reid and Nygren, 1988), and the NASA-Task Load Index (NASA-TLX), developed by the NASA Ames Research Centre (Hart and Staveland, 1988). These tools provide a summary and detailed analysis of aspects relevant to cognitive workload (Miller, 2001). In addition, these tools can also help to determine specific sources of problems related to cognitive workload (Hart and Wickens, 1990).

SWAT is composed of three scales: *(1) Time Load; (2) Mental Effort Load; (3) Psychological Stress Load,* where each scale has a number of options for response *(low, medium* and *high)* (Reid and Nygren, 1988). SWAT is a suitable tool for assessing cognitive workload in ground-based operations such as simulations and laboratory-based trials. In spite of the success and simplicity of SWAT, using three scales with only three dimensions for each may potentially decrease the accuracy of the available responses. It has been characterised as a tool with high inter-rater variability and a less sensitive tool when a low-level cognitive workload occurs (Hart and Wickens, 1990).

NASA-TLX is a cognitive workload assessment tool that is used to identify human performance aspects based on six independent subscales: *(1) Mental Demand; (2) Physical Demand; (3) Temporal Demand; (4) Performance; (5) Effort; (6) Frustration*. NASA-TLX incorporates features such as multidimensional rating, along with subscales for different measurements, and achieves a higher rate of user acceptability due to its ease of use. NASA-TLX correlates between high performance measures and low inter-rater variability (Hart and Wickens, 1990), Consequently, user interaction quality in HCI context is associated with mental workload and affective states (Iqbal et al., 2004). Therefore, research that seeks efficient and productive interaction within the workplace, particularly in relation to HCI, focuses on making suitable task demands that neither underload nor overload a user (Rubio et al., 2004). Subsequently, one of the important workload assessment instruments is the NASA-TLX, which is of considerable interest to researchers and subsequently used extensively in different, complex human-machine systems such as aviation, transportation, military and healthcare systems (Cain, 2007; Colligan et al., 2015).

In addition, (Bradley and Lang, 1994) developed a user self-reporting technique that employs a non-verbal approach, known as the Self-Assessment Manikin (SAM). As shown in Figure 2.3, SAM uses a collection of pictures to self-assess different measurements using three dimensions: (1) *Valence*, which refers to pleasure; (2) *Arousal*, which refers to

activation (i.e. agitation); (3) *Dominance*, which refers to the level of control that a stimulus evokes. The advantage of using the SAM scales is that it can be exploited in a wide range of applications, across different experience levels of users due to its pictorial nature, hence making it very easy to be understood and used.

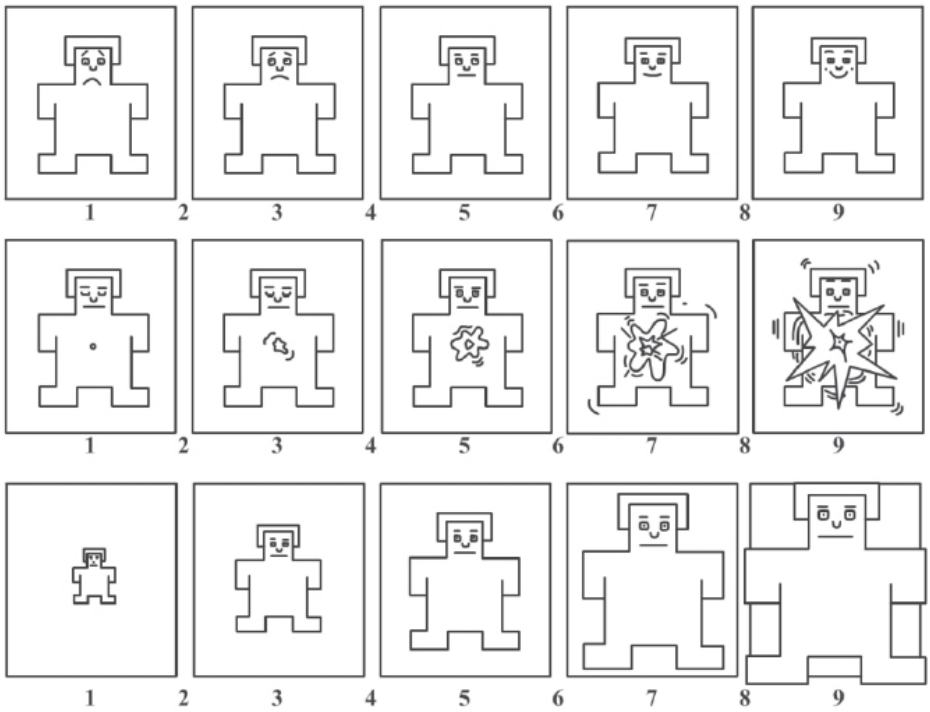


Fig. 2.3 Self-Assessment Manikin scales (SAM) for dimensions of Valence, Arousal, and Dominance (Bradley and Lang, 1994).

2.6.2 Perception and Modelling Techniques

The perception of the user's state is a key component in adaptive HCI (Picard, 2000b; Sullivan and Tyler, 1991). Therefore, several studies have been conducted to deal with various types of perception techniques to detect and predict users' affective states whilst they interacting with computer systems (Ahn and Picard, 2014; Hariharan and Philipp Adam, 2015; Lanatà et al., 2013; McDuff et al., 2014; Neoh et al., 2015; Soleymani et al., 2012). Additionally, within the research literature there are several machine learning methods that have been used for user states classification. For example, (McDuff et al., 2014) used Support Vector Machines (SVM) and Bayesian classifiers to recognise the stress state of users, employing facial expression as one of the inputs, and the accuracy of model they created was 85% using SVM and 80% using Naive Bayes. Moreover, (Oliver, 1997) built a real-time recognition

system that tracks lips and facial expressions in order to detect a set of emotions using Hidden Markov Models, where their system has a recognition rate of 95%. By contrast, Kapoor et al. (Kapoor et al., 2007) utilised facial expressions together with other input modalities such as mouse pressure and skin conductance response to predict frustration using a range of classification techniques including SVM, Gaussian Process, and K-Nearest Neighbour. Additionally, the use of facial expression features with Bayesian classifiers to determine affective scenes while watching videos was also proposed by Joho et al. (Joho et al., 2009). More information are provided in the survey presented in Table B.1 in Appendix B, where a number of conducted studies within this research area that utilised various types of perception input channels of for the purpose of detecting and recognising the user state.

2.6.3 Input Modalities

In human-to-human interaction, one can intuitively predict emotional state based on observations about a person's facial expression, body behaviour, and voice intonations (Karray et al., 2008). This ability is essential as humans often adapt their own behaviour based on such inferences. Therefore, from a computer engineering point of view, acquiring this kind of information about the characteristics of a user can be achieved implicitly in a variety of ways. Firstly, by interpreting user behaviour from mouse movements (Sun et al., 2014), keyboard keystrokes (Hernandez et al., 2014) and content viewing (Lew et al., 2006; Pazzani and Billsus, 2007). Secondly, from non-verbal affective perception channels, such as eye gaze tracking (Lanatà et al., 2013; Wang et al., 2006), detection of facial expressions (Afzal and Robinson, 2009; Afzal et al., 2009; Fasel and Luettin, 2003; McDuff et al., 2013; Pantic and Rothkrantz, 2000), body movement tracking (Schrammel et al., 2010), gesture recognition (Bretzner et al., 2001; Jaimes and Liu, 2005), speech and auditory analysis (Sobol-Shikler, 2009). Thirdly, from physiological aspects, using sensor-based detection such as Electrocardiogram (ECG) (Kim et al., 2004), Electroencephalogram (EEG) (Sourina and Liu, 2013) and Galvanic Skin Response (GSR) (Shi et al., 2007). Last but not least, a combination of the aforementioned modalities may be utilised within a multi-modal system (Caridakis et al., 2007; Duric et al., 2002; Gunes and Piccardi, 2007; Kessous et al., 2010).

Subsequently, *input modalities*, also referred to as *perception modalities*, refer to the natural human perceptual channels, which are the five human senses *(vision, hearing, touch, smell* and *taste)* (Obrenovic and Starcevic, 2004; Oviatt et al., 2004; Vertegaal and Vertegaal, 2003). Analogously, in the perception modalities for machines and computer systems include the keyboard, mouse, gesture, gaze, speech, touch and pen-based, which represent system inputs utilised to obtain information from the user that is relevant to the application and

associated context. Consequently, perception modalities can be categorised into three primary areas (Karray et al., 2008): (1) audio-based; (2) sensor-based; (3) visual-based.

To begin with *audio-based perception modalities*. *Audio-based perception modality* relates to information inferred from data acquired from different audio channels. Despite the fact that audio signals are not as variable as visual signals (Karray et al., 2008), audio-based modalities suffer from vulnerability to noise. As a result, the performance of such systems degrades dramatically within real-world, noise polluted environments (Chin et al., 2012).

The second category of input modalities is the *sensor-based perception modality*. Many different types of sensors can be used as perception modalities, ranging from the very primitive to the very sophisticated in terms of properties such as specification, mechanism and application (Fraden, 1998). The aim of some sensors is to emulate the human senses, such as haptic and pressure sensors (Robles-De-La-Torre, 2006), and olfactory sensors (Legin et al., 2005). Additionally, a range of biometric sensors has emerged that are utilised to physically detect human body fluctuations and reactions, such as the ECG (Kim et al., 2004), EEG (Sourina and Liu, 2013) and GSR (Shi et al., 2007). However, these types of perception modalities are often impractical, invasive and obtrusive, and time-consuming with regard to setup.

Thirdly, the *visual-based (vision-based) perception modality*, which represents the channels that provide data from seeing and perceiving the visual space of the world. These modalities are deemed the most predominant input channels in the domain of Affective Computing where the systems have the ability to recognise and interpret human affect (Afzal et al., 2009), due to the correlation of facial expressions with human emotion (Ekman, 1999).

Nevertheless, many of these technologies are noisy, intrusive and obtrusive, which may further produce a biased effect rather than detecting the actual states of the user (Hernandez et al., 2014). On the contrary, visual-based modalities have many advantages over the other modalities. Firstly, the human face mediates the perception of emotional expressions, and affects interpersonal behaviour (Bruce, 1992). In addition, visual-based inputs are more immune to noise than audio-based inputs (Bretzner et al., 2001). Also, they may potentially provide perceptions about the user in real-time in a way that is unobtrusive to the user, so the interaction can normally run without the need for any specialised equipment. Moreover, the process of detection is viable at a distance and without the need for physical contact (Shan et al., 2007). Furthermore, users often interact with computers as *social agents* in which visual interaction may be the most feasible and intuitive form of interaction (Pantic and Rothkrantz, 2003). Consequently, features extracted from visual-based inputs, i.e. facial expression and eye tracking data, are commonly utilised due to their ease of use, improved accuracy and the unobtrusive nature of these technologies (Lanatà et al., 2013).

2.6.4 Facial Expression

As one of the input channels that is commonly used as a visual-based perception modality, facial expressions are considered as one of the most relevant features that can provide an indication about a user's emotional state (Afzal et al., 2009; Fragopanagos and Taylor, 2005). Furthermore, they are considered instrumental in revealing mental states and clues to the user's feelings (Akakin and Sankur, 2010). Although facial expressions are considered the main cue for emotion recognition, the use of body expressions has also been employed to provided information about the intensity of the emotion (Kleinsmith and Bianchi-Berthouze, 2013). Moreover, other studies in neuroscience and psychology have suggested body expressions are as valuable as facial expressions in terms of providing an indication of emotional state (Kleinsmith et al., 2011). However, capturing body expressions may not be suitable in some workspaces and contexts.

Generally, facial expressions are generated from the movements of facial muscles, which can originate unconsciously (Vega et al., 2014). Within the research literature, face gestures are detected by analysing features from different regions of the face, primarily the mouth, nose, eyes, eyebrows, and forehead (Akakin and Sankur, 2010). Moreover, many facial recognition systems uses a facial expression taxonomy, notably the Facial Action Coding System (FACS), as suggested by Ekman, Friesen and Hager in (Ekman et al., 2002). Within the FACS, 46 Action Units (AU) have been identified for the face, where each AU corresponds to a specific facial behaviour. Moreover, apart from FACS system and action units, (Lopes et al., 2017) analysed facial expressions using Convolutional Neural Networks. In addition, the face reader software that is a known tool for facial expressions analysis and detection has deployed Active Appearance Model descriptors to extract features that used to train an artificial neural network (Den Uyl and Van Kuilenburg, 2005). Furthermore, another direction of this area of research that exploits 3D images for facial expression recognition such as the work of (Lemaire et al., 2013), in which they used Differential Mean Curvature Maps as features of depth images. Additionally, (Isezaki and Suzuki, 2011) proposed a method spatial and time-series analysis of facial expressions using real-time depth images.

Benchmark Datasets

Many research studies have focused on building annotated and robust datasets in order to develop computerised models for the automatic prediction of states of human emotion. For example, Afzal et al. (Afzal and Robinson, 2009) used simulated driving scenarios and a computer-based learning settings to induce different emotions that were subsequently detected from facial expressions. Furthermore, Ahn and Picard (Ahn and Picard, 2014)

showed that visual-based inputs are used to predict customer's desires by analysing the state shown in the facial expression of subjects during a beverage tasting experiment. Other studies have explored the use of facial expression within a learning context using pedagogical agents and tutoring systems using AU from the FACS framework (Whitehill et al., 2008). As provided in Table B.1 in Appendix B, several contexts exploited across different applications that are related to human states detection and modelling.

Besides, there currently exists a number of available datasets that commonly used for benchmarking, and evaluating facial expression recognition techniques. Firstly, the the CMU-Pittsburgh AU-Coded Face Expression Image Database, known as the CK dataset, which contains a number of image sequences from 182 subjects of several ethnic groups, who performing expression of the most common FACS action units (Kanade and Cohn, 2000). Subsequently, the Cohn-Kanade Plus dataset (CKPLUS), which is an extension of the CK version released in 2000, is deemed a benchmark for automatic facial expression analysis and detection (Ko, 2018). CKPLUS is comprised of 593 sequences taken from 123 subjects. The labelling process was carried out using the FACS coding system and only 327 sequences met the criteria to be labelled with a specific emotion. Consequently, the sequences are divided into seven groups: *angry (45), contempt (18), disgust (59), fear (25), happy (69), sadness(28)* and *surprise (83)* (Lucey et al., 2010).

In a similar manner, (Lundqvist et al., 1998) published the Karolinska Directed Emotional Faces dataset (KDEF), which consists of 4900 pictures captured from 70 subjects (equally divided between 35 males and 35 females), where each subject acted seven different affective states, which include: *afraid, angry, disgusted, happy, neutral, sad* and *surprised*. Each facial expression was captured in two sessions from 5 different angles: full left, half left, straight, half right, and full right.

Likewise, (Lyons et al., 1998) published the Japanese Female Facial Expression (JAFFE) database, which contains 213 images rated by 60 subjects as one of the states: *neutral, happiness, sadness, surprise, anger, disgust* and *fear*. Moreover, another multi-modal dataset entitled DEAP, is a dataset for Emotion Analysis that contains EEG signals, along with physiological and video signals from 32 subjects (frontal face video was recorded for only 22), who individually watched 40 one-minute music videos of different genres as a stimulus to induce different affective and emotional states (Koelstra et al., 2012). Furthermore, the Boğaziçi University Head Motion Analysis Project (BUHMAP) database was collected, annotated and released freely for academic research purposes. The BUHMAP database contains a number of videos for 8 different classes of visual signs including: *neutral* state of the face, *head-L-R* (shaking the head to right and left sides), *head-UP* (raise the head upwards while simultaneously raising the eyebrows), *head-F* (head is moved forward accompanied

with raised eyebrows), *sadness* (lips turned down and eyebrows down), *head-U-D* (nodding head up and down continuously), *happiness* (lips turned up when subject smiles), *happy-U-D* (the preceding two classes are performed together which are *head-U-D* + *happiness*) (Aran et al., 2007). Additionally, The AVEC challenge dataset represents a naturalistic dataset, which consists of a large number of emotionally coloured interactions (31 videos used for training, 32 videos for development, and another 31 videos for testing) between human participants and an emotionally-stereotyped character (Schuller et al., 2011).

Consequently, among the diverse types of datasets, CKPLUS and KDEF datasets are more suited to facial expression analysis and have been used as benchmarking sets for facial expression analysis, because they were labelled to the universal basic labels for emotional states, which commonly observed within different applications in Affective Computing research. Subsequently, this allows them to be exploited within different contexts besides the conversations, on the contrary of non-acted naturalistic datasets that were designed to be deployed in particular context. Additionally, the CKPLUS and KDEF datasets were widely used in the facial expression recognition literature, which makes benchmarking and comparative analysis of feature extraction and machine learning techniques more obvious and easy to comprehend.

2.6.5 Eye-Gaze Tracking Data

Eye tracking technology has been widely used as an assessment approach for eye gaze behaviour. Subsequently, eye-tracking is a vital visual-based channel that conveys a wide range of information. For example, conversation in human-human interaction is managed through eye contact, such as managing turn in conversation, displaying attentiveness and giving feedback (Ruhland et al., 2015).

In general, several techniques have been adopted for the analysis of eye-tracking data within the HCI domain (Sharma and Dubey, 2014). For instance, the possible correlation between eye-tracking data and the usability issues of websites was investigated by Ehmke and Wislon (Ehmke and Wilson, 2007). In addition, eye-tracking data is widely used in several areas related to the detection of mental states and cognitive workload (Morimoto and Mimica, 2005), and can efficiently provide an indication about the amount of human cognitive processing being applied to objects at fixation points along with contextual information it can provide about where a user is looking in a particular moment of time (Poole and Ball, 2005; Slaney et al., 2014).

Furthermore, eye gaze has been adopted for the control of the UI and exploited as a pointing apparatus to replace the mouse. In particular, eye gaze-based interfaces are typically deployed to people with severe disabilities (Biswas and Langdon, 2011). As well as several

studies identified the eye gaze patterns differences across different visualising techniques (Blascheck et al., 2014). Consequently, eye gaze metrics like *fixations* number and duration, and *saccades* have been exploited in visualisation analysis approaches such as analysis of information lookup tasks, pre-attentive and scanpath interruption, and specifying the area of interest as shown in the work given by (Goldberg and Helfman, 2011; Steichen et al., 2014).

Moreover, eye gaze patterns have been used to identify the nature of the user task, as presented in (Iqbal and Bailey, 2004). The authors showed that the pattern of eye movement changes across four different categories of task: *Reading, Mathematical Reasoning, Searching* and *Object Manipulation*. Measuring the amount of attention on the screen using eye tracking data is also investigated as part of applications within usability and media research, as presented in (Schiessl et al., 2003). In addition, eye tracking data was exploited in the assessment of task performance, user physiology and task difficulties, in conjunction to other physiological measurements and mood ratings (Bruneau et al., 2002).

Eye Gaze Measurements

Human eye is a sensory organ of spherical shape that reacts in response to environmental light changes and psychological factors (Atchison, 2017). The external parts of human eye are *pupil, iris* and *sclera*. The *pupil* is the aperture where through the light goes into the center of the eye. The *iris* is the texture that surrounds the pupil and gives the colour to the eye. The *sclera* is the white area of the eye that surrounds the iris, and the cornea is the transparent membrane that covers the frontal side of the eye (Morimoto and Mimica, 2005; Poole and Ball, 2005).

Many research studies have reported that tracking eye gaze behaviour using different measurements can reflect different aspects about the performance of the user (Chen and Epps, 2014). Within eye gaze tracking research, the most used measurements and metrics are *pupil dilation, blink rate, fixations* and *saccades* (Imotions, 2016; Iqbal et al., 2004; Poole and Ball, 2005; Recarte et al., 2008). With regard to the measurements *pupil dilation* and *blink rate*; they have been widely explored as they can provide a feasible representation of the actual cognitive and perceptual workload (Chen and Epps, 2014).

For instance, *pupil dilation* is strongly correlated with mental processing and cognitive workload (Iqbal et al., 2004). Primarily, pupil dilation is generated unconsciously from the movement of radial and circular muscles located in the iris (Marshall, 2000). Where the radial muscles pull the pupil outward, and the circular muscles pull it inward, causing the pupil expansion and constriction, as depicted in Figure 2.4.

Additionally, there are two main causes of pupil size change: illumination and mental and cognitive processes. On one hand, pupil size changes in response to illumination, in

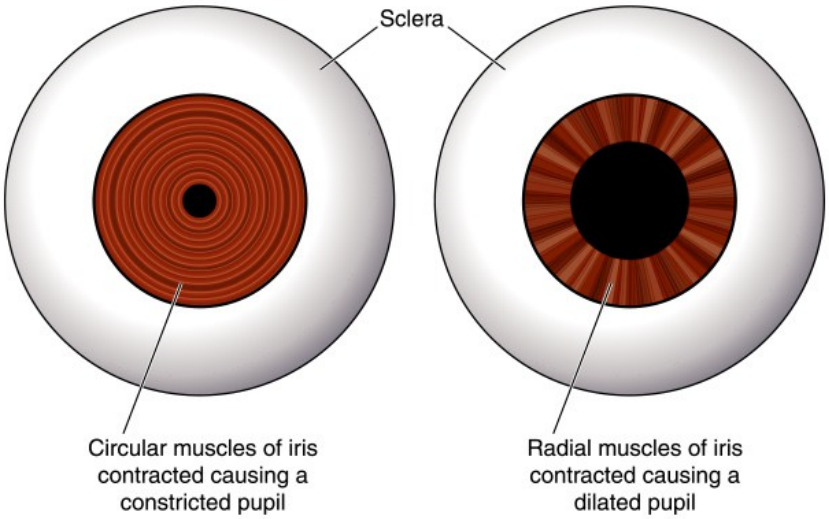


Fig. 2.4 Circular and radial of iris muscles contraction causing pupil constriction and dilation respectively (DeRemer, 2015).

order to control the amount of light that enters the eye (Lowenstein and Loewenfeld, 1962). For instance, increasing the lighting conditions causes the pupil to get smaller. On the other hand, pupil size dilates in response to attention, fatigue, mental effort and other physiological and psychological factors (Hoeks and Levelt, 1993; Tryon, 1975). Subsequently, it is widely reported that pupil size is strongly correlated with mental workload; the magnitude of pupil dilation is considered proportional to the mental effort required to process a task (Iqbal et al., 2004; Pomplun and Sunkara, 2003).

Furthermore, pupil dilation is considered as a manifest physiological metric of the mental processing workload, which is known as*Task-Evoked Pupillary Response* (TEPR), proposed by Beatty, who investigated pupillary response caused by cognitive workload (Beatty, 1982). Moreover, pupillary response can indicate mental workload and processing both within a task and between different tasks (Fehrenbacher and Djamasbi, 2017).

Other measurements can be extracted from eye tracking data, such as the *blink rate*, which can be utilised as an index for mental workload during cognitive tasks, whereby a higher *blink rate* indicates a higher mental effort (Recarte et al., 2008). Nonetheless, *pupil dilation* is used more often than the *blink rate* because it varies according to a wide range of processing activities such that the dilation proportionally changes with levels of cognitive processing (Buettner, 2013; Chen and Epps, 2014; Marshall, 2000).

Additionally, measurements such as *fixations* and *saccades* that relate to eye movement within a HCI context can provide useful information about user's attention, and can be used to indicate the pattern or entropy of eye movement during interaction with a computer system (Poole and Ball, 2005). Where *fixations* represent the moments that a user's eyes are relatively stationary on a specific object, and the *saccades* represent eye movements between fixations (Poole and Ball, 2005; Rozado et al., 2015).

Measuring and Tracking Technology

There are different techniques for tracking eye gaze behaviour. these techniques include scleral search coil method (Robinson, 1963), Electrooculogram (EOG) (a.E. Kaufman et al., 1993) and using optical infra-red-based cameras (Morimoto and Mimica, 2005). Firstly, the scleral search coil method that is one of early attempts in this direction, which works using contact lenses that have a search-coil placed with positioned magnets around the eye, so variation in the magnetic field induces a voltage, which is subsequently used to estimate eye movement (Robinson, 1963). Secondly, the EOG is another technique that is commonly used in clinical environments for eye movement tracking, which works using a number of electrodes that should be placed around the eye, the electrodes measure the amount of the standing potential resulted from the micro-currents flow caused by hyper-polarizations and depolarizations between the cornea and the retina of the eye (a.E. Kaufman et al., 1993; Bulling et al., 2011; Singh and Singh, 2012). However, scleral search coil method and EOG techniques are certainly intrusive, and at the same time can be used only to track the movements of the pupil without measuring variation in the pupil size.

Alternatively, optical camera-based eye-trackers may be used, which work through analysis of information from the light (i.e. infra-red) reflected by the eye, such as pupil reflections and corneal reflections (Hansen and Ji, 2010; Morimoto and Mimica, 2005). However, commercial infra-red eye-trackers currently available typically require special setup and calibration in order to be able to measure the gaze location and pupil diameter (Harezlak et al., 2014).

Subsequently, measuring pupil size has long been investigated in academia and industry, and implemented tool sets can be found as free or commercial packages (Li et al., 2006). The pupil size is measured either by counting the number of pixels of the pupillary area in the image, or using the relative movement of the pupil to the cornea reflection (Wang, 2011). Generally, determining the pupil size by analysing images for the eye can be grouped into two categories (Kirschbaum, 1998). For instance, one approach depends on the fact that the pupil is darker than the *iris* and the *sclera* that surround the pupil in the captured image of the illuminated eye. Therefore, they determine the pupil size according to the number of pixels

that are less than a predefined threshold (Ohno et al., 2002). The other approach is based on finding the best fit ellipse or circle of the pupil contour, so either the area, diameter or the major axis can be used as a measure of the pupil size (Hansen and Pece, 2005). Alternatively, a combination between the two approaches is considered more robust and practical (Wang et al., 2015; Winfield and Parkhurst, 2005).

In addition, eye gaze behaviour has been investigated using normal webcam, also referred to as *webcam eye-trackers*, which is a software application that works by manipulating frames captured by a webcam in order to track eye gaze behaviour, specially estimating the gaze location, which is sufficient for usability testing and product design (Papoutsaki et al., 2016).

2.7 Machine Learning and Classification Methods

There exists a wide range of machine learning and classification techniques that can be used for facial expression recognition. Within the research literature, facial expression recognition is an active area of research despite the number of successes in this domain (Akakin and Sankur, 2010; Den Uyl and Van Kuilenburg, 2005; Ghimire and Lee, 2013; Liew and Yairi, 2015; Liu et al., 2014; Lopes et al., 2017; Shan et al., 2009). A number of studies exist where machine learning techniques trained on features extracted using different approaches achieved a reasonable classification accuracy. For example, the work of Liew and Yairi (Liew and Yairi, 2015) achieved a classification accuracy of 91.2% on the CKPLUS dataset using a SVM trained on Histogram of Oriented Gradients. Likewise, in the study reported by (Ghimire and Lee, 2013), an accuracy of 97.35% was achieved using SVM applied on geometric features on the CKPLUS dataset. In addition, (Shan et al., 2009) achieved an accuracy of 95.1% using a SVM model on Local Binary Pattern features using the same CKPLUS dataset. Moreover, (Akakin and Sankur, 2010) reported a classification accuracy of 94.04% using a SVM on features learned by Adaboost that extracts a set of discriminating spatio temporal features from the BUHMAP video database. Using the KDEF dataset, as reported by (Den Uyl and Van Kuilenburg, 2005), a classification accuracy of 89% was achieved using face reader software that deploys Active Appearance Model descriptors in order to train an artificial neural network. Furthermore, recent progress with the evolution of Deep Learning techniques has produced a classification accuracy of 96.76% on CKPLUS using Convolutional Neural Networks, as shown in the work presented by Lopes et al. (Lopes et al., 2017). Moreover, the work of (Liu et al., 2014) achieved an accuracy of 96.7% on the CKPLUS dataset using a boosted Deep Belief Network that is based on a composition of a set of weak classifier.

Support Vector Machines

SVM technique seeks to find an optimal hyperplane that separate the classes of data, as illustrated in Figure 2.5, is considered one of the most robust and efficient techniques that is commonly used for classification of expressions from facial data.

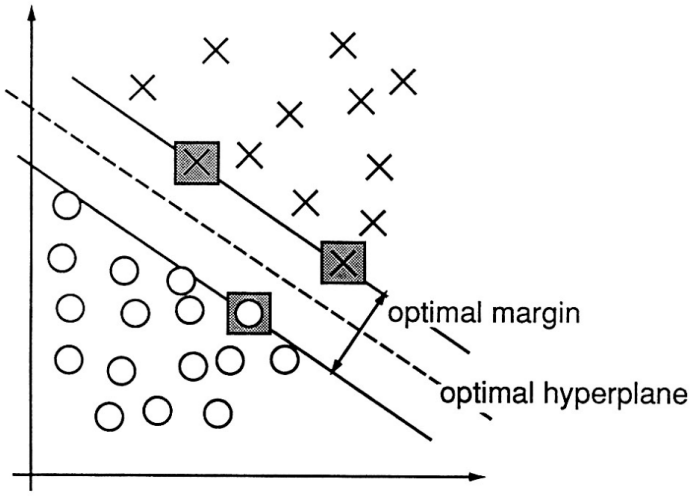


Fig. 2.5 Illustration for Support Vector Machine classifier on two dimensional space data, where an optimal hyperplane is the one that separates the two classes of data with maximum margin (Cortes and Vapnik, 1995).

SVM evolved initially from the idea of (Vapnik and Lerner, 1963) in their early publication, entitled *Pattern Recognition using Generalized Portrait Method*. The basic idea of SVM is to find the optimal hyperplane for linearly separable patterns, where the support vectors represent the closest data samples to the decision surface that separates between different data classes, which are the square shaded in Figure 2.5. Subsequently, the current, practical version of SVM was originally proposed by (Boser et al., 1992), in which SVM was extended using Kernel functions in order to be employed for patterns that are not linearly separable, through means of transformations and mappings of the original data into new spaces. In other words, the original data samples are mapped into a higher dimensional space such that the features become linearly separable. Figure 2.6 shows an example of data transformation using a Kernel function. However, SVM performance on non-linearly separable data depends on the kernel function employed. Therefore, the use of Kernel functions such as Linear, Polynomial and Radial Basis functions, transforms the input data into a higher dimensional space that is separable, which allows SVM to perform well in real life datasets (Xu et al., 2009).

As a result, SVM as an instance-based learning technique (Domingos, 2012), is a relatively simple technique that is efficient for linearly and non-linearly separable data, and

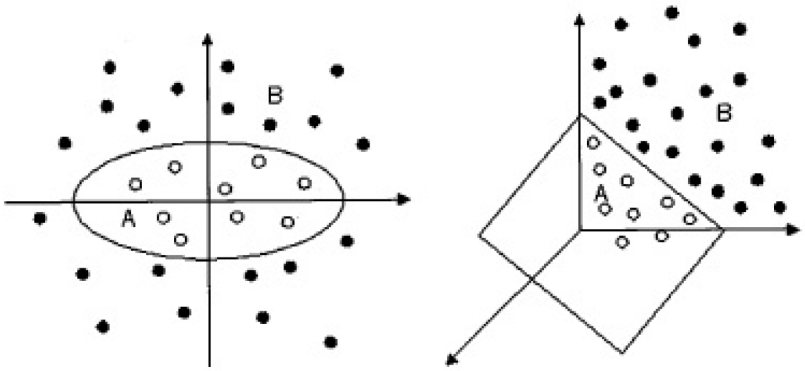


Fig. 2.6 Non-linear data transformation into a linearly separable data in a higher dimensional space using Kernel functions (Hofmann, 2006).

deemed a robust tool for classification and regression, as well as it is currently among the best tools that can be deployed for different classification tasks ranging from text to genomic data (Burbidge and Buxton, 2001; Jannsen, 2008; Menezes et al., 2017). Moreover, for supervised learning, SVM has been successful and widely used for pattern recognition and generating prediction models, which produces good classification accuracy rates, and it is able to deal with high-dimensional data (Jannsen, 2008; Xu et al., 2009). In addition, it can handle the *Curse of Over-fitting* that other classifiers suffer from by tuning Kernal parameters such as applying regularisation, and adding structural constraints on the decision surface (Awad and Khanna, 2015). Which thereby makes it a powerful predictive model along with good generalisation properties (Jannsen, 2008).

SVM has been implemented by many people like LIBSVM (Chang and Lin, 2001), and SVMLight (Joachims, 1999). In addition, it has been included within many popular data analysis and machine learning tools, such as MATLAB and WEKA (Mikut and Reischl, 2011).

2.7.1 Convolutional Neural Networks

The rise of Deep Learning, as a branch of machine learning, in which the architecture is based on multiple levels of neural networks, produces high levels of classification performance in different application domains, such as Image Recognition (He et al., 2016), Recommender Systems (Zhang, Yao and Sun, 2017), Natural Language Processing (Goodfellow, Bengio and Courville, 2016), etc. Deep Learning is a structure of layered neurons and biases associated with weights and parameters, and it works by tuning the parameters and adjusting the weights of internal layers according to the representation of the previous layers using the backpropagation algorithm (LeCun, Bengio and Hinton, 2015; Karpathy, 2016).

A Convolutional Neural Network that shows outstanding performance in image classification is the most popular and successful architectures of Deep Learning, which in simple words is a neural network composed of more than one layer (Krizhevsky, Sutskever and Geoffrey E., 2012; Schmidhuber, 2015). Convolutional layers are fully or sparsly connected with pooling layers are ended by a classification layer (Ciresan, Meier and Masci, 2011). Subsequently, Convolutional Neural Networks have shown their efficiency within computer vision and image classification tasks by learning and training the network parameters from regions (i.e. patches) of the input image (i.e 2D images) directly, without the need of preprocessing and feature extraction (Agarap, 2017), as can be depicted in Figure 2.7.

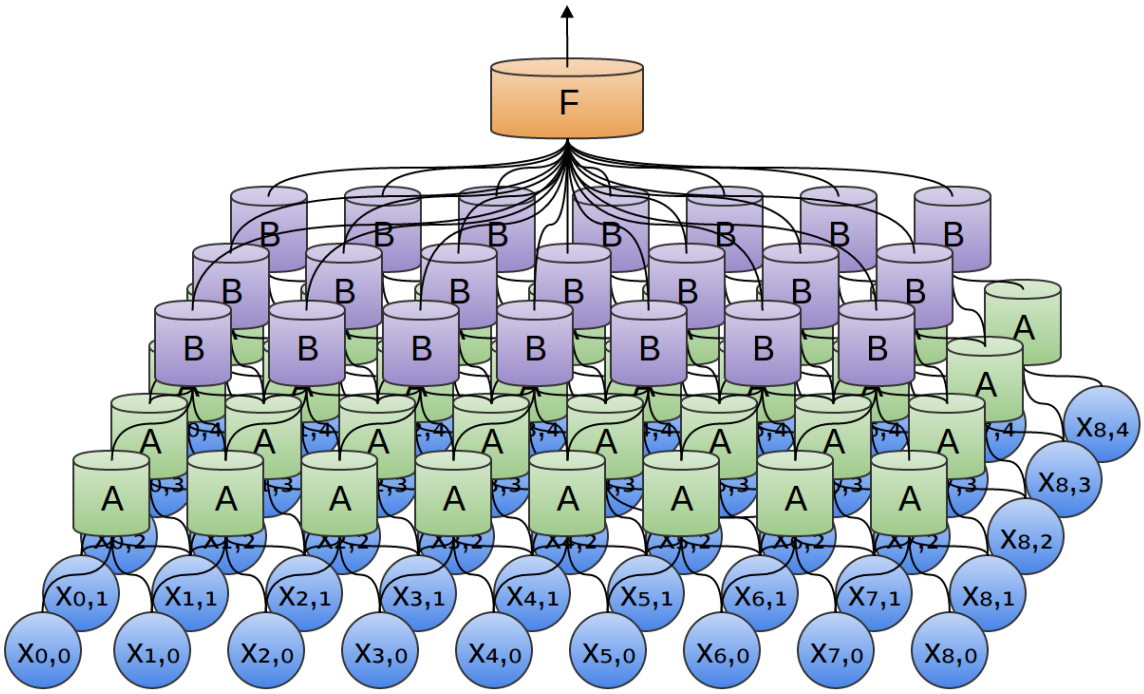


Fig. 2.7 Convolutional Neural Network structure of two dimensional input i.e. image (Olah, 2014).

Several architectures for Convolutional Neural Networks have been developed and are currenly considered as state-of-the-art models, such as *QuocNet* (Le et al., 2011), *AlexNet* (Krizhevsky et al., 2012), *Inception GoogLeNet* (Szegedy et al., 2015), and *Batch Normalization-Inception-v2* (Ioffe and Szegedy, 2015). Subsequently, in the *ImageNet Large Scale Visual Recognition Challenge*, which has been running since 2010, whereby the challenge is to improve classification accuracy on a dataset that contains millions of images for hundreds of different object types (Deng et al., 2009; Russakovsky et al., 2015), a *Deep Convolutional Neural Network* using the *AlexNet* architecture achieved the highest

classification accuracy and won the *ImageNet Large Scale Visual Recognition Challenge*, which was implemented by (Krizhevsky et al., 2012).

Nevertheless, with the continuous improvement of machine learning techniques that become able to achieve accurate mapping between input data and output labels; this reason leads to the ability to increase and utilise the generalisation capabilities of classifiers using transfer learning (Torrey and Shavlik, 2009). *Transfer Learning* is an inductive learning where a model trained on certain data to solve one problem is re-purposed to solve another problem and applied on different data that has common properties with the original data that has been used for training (Goodfellow et al., 2016). Consequently, pre-trained CNN on a set of images can be exploited for image classification of a different set of images, whereby network representation, along with weights and parameters, can be re-purposed and reused for other vision classification tasks (Donahue et al., 2014).

Subsequently, (Pratt, 1993) firstly introduced the idea of transfer learning in his algorithm entitled *Discriminability-Based Transfer between Neural Networks*. His investigation showed that using an initialised network containing weights and parameter values obtained from a network that already trained on another source of data would both improve and speed up the learning process. Additionally, (Donahue et al., 2014) proposed a pre-trained Deep Learning architecture, namely *Deep Convolutional Activation Feature for Generic Visual Recognition* (DeCAF), whereas weights and activations were extracted from training on the ImageNet dataset (Deng et al., 2009).

2.8 Summary

HCI models such as HPM and GOMS embody a set of memories and processors with a set of operations that are useful during the design stage of the UI of a software. However, they do not facilitate building intelligent computer interfaces that simulate human intelligence similar to human-human interaction. Also, HCI models given in Table A.1 in Appendix A designed to target different segments, tasks, methods, systems, environments, and user aspects, such as intentions, goals, procedures, knowledge, competence, skills, performance, and processing abilities (Fischer, 2001). Whilst they are beneficial for guiding HCI designers with taking human factors into account when developing UX, they are typically overcomplicated and more suited to specialised application domains (Biswas et al., 2012).

Furthermore, with reference to adaptive HCI, there is no exact theoretical definition of an intelligent adaptive interface (Karray et al., 2008). However, the main factors that should be considered are: task difficulty, ability, and the desired motive achieved after successful completion (Fairclough et al., 2013). Therefore, adaptation suitability could be measured

and evaluated by looking at how the adaptation approach changes the situation into a positive interaction after it was considered distressful and discomforting, which inevitably results in the user giving up the task before it has been completed .

Consequently, the significance of user modelling as an important factor for the next generation of HCI because it reasons in context with the user's goals, plans, attitudes and capabilities, which may be used to adapt the system accordingly (Sullivan and Tyler, 1991). As a result, a more intuitive form of interaction is sought to foster effective task completion in contrast to conventional non-adaptive applications (Saffer, 2009; Sebe, 2009). However, the idea of UI adaptation often violates traditional HCI design principles such as interface and layout consistency (Kirakowski and Corbett, 1990). Thus, a trade-off exists between complying to HCI design principles and the benefits experienced from adaptive HCI (Gajos et al., 2006). Nevertheless, existing research is limited in terms of the number of states detected and the application targeted.

With regard to exploiting input perception modalities to build a user model; visual-based modalities potentially provide an opportunity for endowing a computer with intelligence that may facilitate natural and intuitive HCI similar to human-human interaction (Harper et al., 2008). Consequently, it is anticipated that employing and facial expressions and eye tracking data as input perception modalities is a suitable and potentially effective approach for modelling affective aspects via facial expression and cognitive aspects via eye tracking data within the context of user interaction in HCI (Chen and Epps, 2014; Saneiro et al., 2014). Furthermore, tracking pupil dilation can potentially facilitate measuring stress and cognitive workload. Both will potentially permit adaptive and intelligent responses during user interaction. Moreover, it can be used for analogous situations where empowering machines with this sort of perception about the user in real time is crucial (Steichen et al., 2014).

Moreover, one can intuitively infer that machine learning methods and techniques, which support pattern recognition and building prediction constructs, are required for user modelling, through means such as image classification, facial expression recognition and pattern learning (Saneiro et al., 2014). Nevertheless, advantages and limitations associated with different techniques makes the selection of a machine learning technique a controversial issue. For example, SVM is very good for creation of a generic pattern learner model and use of high dimensional data, however, its performance decreases dramatically as the amount of noise in the data increases (Cristianini and Shawe-Taylor, 2000). On the contrary, Deep Learning techniques require a long time to train, in conjunction with a large dataset containing a great number of instances, in order to achieve good classification results without over-fitting (Krizhevsky et al., 2012). Correspondingly, many studies within the literature reported a high

classification accuracy, however, the results are not necessarily directly comparable with each other, as a number of different evaluation schemes have been adopted, such as number of folds in cross validation, or using splitting of different percentages, or even providing training data apart from testing data. In addition, to the number of labels/classes (i.e. expressions or states) that are included in the training and evaluation.

Therefore, subsequent chapters of current thesis will talk about a Data Collection Study, which could be used for the analysis of affective states, along with the assessment and self-reports made by the subjects themselves, such as using the SAM scales (Bradley and Lang, 1994), to typify the awareness of subjects' feelings. In addition to classification techniques alongside with feature extraction methods for facial expression analysis, as well as statistical analysis for eye gaze tracking data. Moreover, details of perception component of an adaptive system that potentially facilitate the achievement of a novel form of HCI.

Chapter 3

Human-Computer Interaction Data Collection and Exploitation

3.1 Overview

The aim of current research is to model user affective and cognitive states during interacting with different UIs. The hypothesis attempted in this chapter is to find out if suitable data can be collected during computer-based tasks that represents a daily use of computers. In this Chapter a detailed description is given about the data that have been collected and used in the experimental work pieces within this PhD research project. This involves the design of the procedure and the protocol together with the material and stimuli used, as well as the types of the collected data which can be grouped into subjective self-reporting data, and automatic recordings via visual-based input perception channels.

Additionally, technical details are given with the design specification of the software tool *'HCI-Viewer'*. This software tool was used to view different sceneries simultaneously of the recorded sessions in the Data Collection Study. This software tool has been developed and employed throughout the Data Collection Study, which permits participants to watch the recorded data during the sessions whilst provided self-reports were given. Thus, the implementation and the benefits of this tool is provided in this chapter.

3.2 Methodology

Available datasets in the literature are either general purpose datasets, which abstract the emotional state via facial expression activation such as CKPLUS dataset, or datasets that target a particular context such as AVEC challenge dataset that represents an interaction

between a human and stereotyped characters. The study presented in this chapter aims to generate a purposeful dataset that is suitable for HCI research in studying users' affective and cognitive states whilst they are interacting with typical UIs in different HCI contexts. The work given in this regard involves the design and implementation of a data collection study undertaken by 42 participants from Ulster University of various levels of expertise, including Undergraduate students, Postgraduate students and Staff members. The participants attempted to complete four predetermined tasks that represent the most frequently used interfaces among ordinary computer users. Each participant conducted four tasks that involved use of the operating system of the computer, online shopping using a web browser, spread-sheet manipulation and briefly playing a video game. Correspondingly, the study comprises different tasks to represent materials that induce different affective and cognitive states. Subsequently, the selected tasks potentially facilitate aspects of difficulty in users that will lead to a range of emotions and increase in cognitive load. Consequently, the facial images and eye gaze tracking were recorded when participants carried out the tasks. Additionally, participants were asked to self-report their actual emotional states using SAM scales and task demands using NASA-TLX scoring tool after each individual task. The data collection protocol has been approved by the Ethics Filter Committee of the Faculty of Computing and Engineering at Ulster University (FCE 20150617 15.31).

3.3 Data Collection Study Methods

Within the experiments outlined and discussed within this thesis, a Data Collection Study was initially conducted in order to collect features from different input modalities, which is used to reason about users' affective and cognitive states whilst interacting with common computer software and attempting to complete typical computer-based tasks.

A total of 42 participants took part in the study, whereby participants were either staff or students at Ulster University. There was no specific inclusion or exclusion criterion, other than being a current student or staff member of the university, as the study is interested in identifying affective and cognitive states in a generic HCI context, hence participants' experiences could vary from novice to expert computer users.

3.3.1 Material and Stimuli

The material for the tasks used throughout the study can be classified into four main categories: (1) basic operating system tasks; (2) online shopping tasks; (3) Excel spreadsheet manipulation tasks, and (4) game-based tasks. Consequently, the themes of these had been

chosen according to a study of computer usage statistics carried out by Thomas Beauvisage (Beauvisage, 2009), which presented the average distribution of individual weekly computer usage in France. However, the author assured that similar usage behaviour exists in other parts of the worldwide that have mature Internet markets such as North America and Western Europe.

Notably, the study showed that the four categories identified; occupy the highest percentages of time spent on computer usage for both households and individuals. Where Beauvisage's work dealt with data recorded over 19 months from 661 households and 1434 users at home.

Accordingly, the set of selected tasks represent active interaction with a computer in which the participant has a predefined task to carry out within a limited time of 5 minutes at most, and they were requested to do as follows:

1. **Basic operating system task**: the participant was asked to change the desktop background, screen saver, time-zone, and add a new input language to the system within the predefined time limit. Figure 3.1 shows a snapshot taken from one of the recorded sessions while the participant attempting this task.

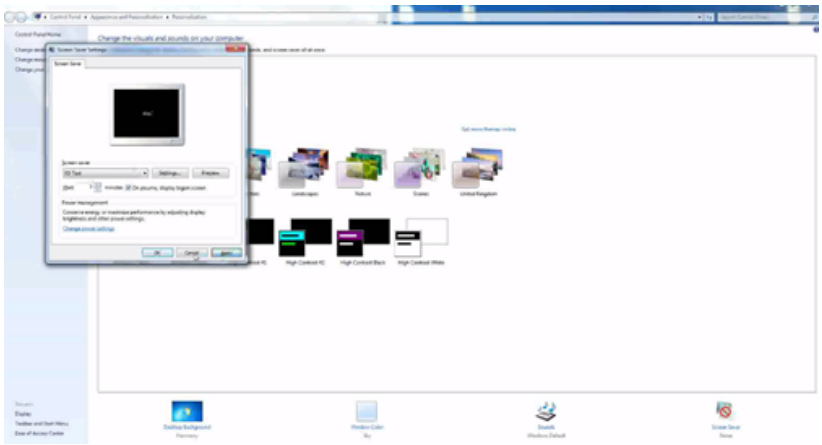


Fig. 3.1 Operating System Task, where the user tried to change the desktop background and the screen saver.

2. **Online shopping task**: the participant was asked to search online for a Tablet-PC with specific properties using their preferred Internet browser application. As shown in Figure 3.2, one of the participants used Amazon while looking for the requested tablet.

3. **Excel spreadsheet manipulation task**: the participant was asked to modify an existing Excel spreadsheet to insert new data into the existing records, sort the data in

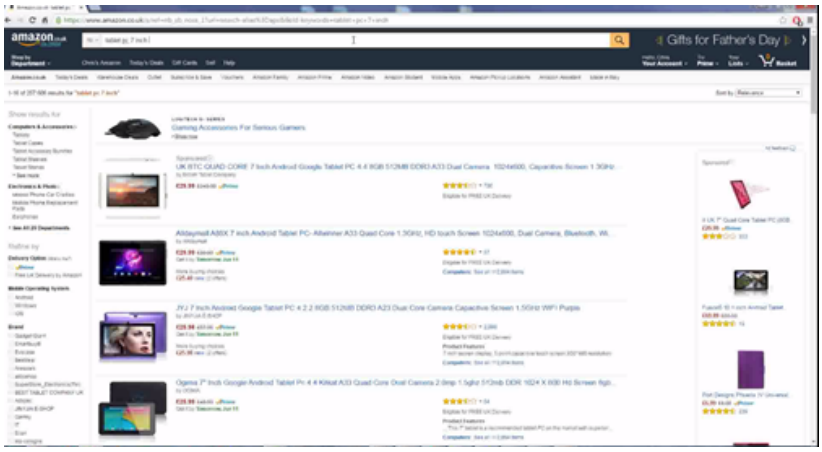


Fig. 3.2 Online Shopping Task, where the user tried to use Amazon looking for a Tablet-PC with specified properties.

ascending order, use an aggregation function (i.e. *Average*), and draw a line graph of the data, as the snapshot taken from one of the recordings of this type of task is given in Figure 3.3.

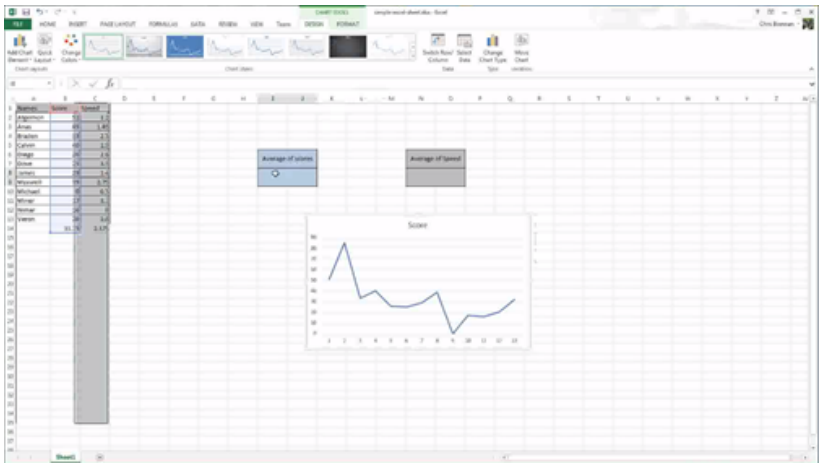


Fig. 3.3 Excel Spreadsheet Task, where the user worked on the spreadsheet and was able to generate a graph from the data.

4. **Game-based task**: the participant was asked to play a version of the arcade game Pacman (namely Deluxe Pacman 2 - Release v16 (Roy, 2014)) for a total of 3 minutes. A snapshot of one of the recordings of a participants playing Pacman is given in Figure 3.4.

Fig. 3.4 Game-Based Task (i.e. Pacman).

3.3.2 Session Recording Setup and Procedure

Firstly, each participant was given an information sheet describing the flow of the experiment, and then he/she was asked to sign a consent form that his/her participation in the study is completely voluntary. Moreover, the participant was asked to provide an estimate of his/her competency in the task, and the expected difficulty level before carrying out the task. In addition, the participant was asked again to report the actual difficulty level afterwards.

During the session, the video of participant's face was recorded using a typical webcam placed at the top of the screen that captures 30 Frames Per Second (FPS). In addition, an infra-red eye tracker, namely Eye-Tribe, was calibrated for each single participant, which works in a sample rate of 30 samples per second. Moreover, each frame hold information about the pupil size and the X, Y coordinate fixation of the left and right eyes. Additionally, the screen where the interaction between the participant and the UI was recorded to examine usability issues and other aspects related to the interface.

Furthermore, webcam recording, eye tracker recording and screen recording are synchronised together, whereby they triggered to start recording at the same moment. These corresponding software tools that were used for recording kept populating the frames until the session was over. Thus the three recordings started from the beginning of the session and lasted during the whole time of the session.

3.3.3 Self-Assessment Reporting by Participants

At this stage, the participant commenced working on each task within the designated time. Upon task completion, the participant was asked to fill out the subjective workload rating scale, the NASA-Task Load Index, which comprises six sub-scales related to the task

which are: *Mental Demand, Physical Demand, Temporal Demand, Performance, Effort* and *Frustration* (Hart and Staveland, 1988).

Additionally, participants were also asked to use the non-verbal pictorial assessment tool, the SAM scales, in order to report their feelings and affective states, and rate the amount of Valence, Arousal and Dominance whilst attempting different tasks during sessions (Bradley and Lang, 1994). The subject self-reporting and information sheets are given in Appendix C.

3.4 A Tool for Human-Computer Interaction (HCI-Viewer)

In order to give HCI researchers, UX and UI designers and practitioners the ability for to infer different aspects about the interface itself, a software tool referred to as HCI-Viewer designed, implemented, developed and used to view the participant's facial expressions conjointly with eye gaze behaviour when carrying out the interaction with the interface.

The HCI-Viewer tool aims to display the different recordings concurrently. The tool potentially provides the researcher with insights into information that entails different aspects of the graphical interface of the software application used during each of the four tasks, conjointly with contextual information about the eye fixations, as well as the pupil size variation during a recorded session. At the same time, it can envision the participant's emotions by showing the video of the participant's face whilst he/she is interacting with a particular software application.

Figure 3.5 illustrates the mock-up design of the primary user interface of the tool. As shown, a recording of the participant's face is presented alongside the recording of the screen where the interaction with the software application's user interface was being conducted. In addition, the fixations of the participant's eyes on the user interface are overlayed synchronously with the video playback. Several associated measures are also displayed, together with instantaneous pupil size and the average of the pupil size during the session. Furthermore, a graph plot of the changes in pupil diameter throughout the session is also displayed by the tool.

Moreover, the HCI-Viewer supports the ability to change a number of default options related to eye-gaze measurements: (1) the window length used during the aggregation of the measurements; (2) the rate that represents the denominator used to calculate the fixation rate; (3) the area of interest, which is the circle size that surrounds the fixation point, which is useful in tracking the fixation across different visualisation types and sizes.

The technical details about the implementation of HCI-Viewer, the development kit, the exploited application packages, and the logic implemented within the tool are given in Appendix D.

Fig. 3.5 Mock-up design of the HCI-Viewer tool interface.

3.5 Conclusion

As previously mentioned, 42 participants took part in the Data Collection Study, during which four sessions were recorded, where each session corresponded to a single user interaction task. The participants were members of Ulster University of different levels and expertise with computers. As presented in Table 3.1, before the experiment took place, each participant was asked to self-report his/her level of competency on the domain of the given task on the range from [0-10]. Subsequently, the average competency level of Online Shopping Task (Online) was the highest "8.9", followed by the Operating System Task (OS), then Excel Spreadsheet Task (Excel) and Game-Based Task (Pacman). In addition, participants were asked to estimate the difficulty level of the task based on the description before attempting doing it (i.e. Pre-Difficulty), as well as after doing the task (i.e. Post-Difficulty). Accordingly, the reported difficulty level afterwards is greater across the four tasks.

Furthermore, the participants filled-in the NASA-TLX after each session was completed, which includes the *Mental, Physical, Temporal, Performance, Effort* and *Frustration* level during each task. Table 3.2 shows the average of each score obtained from the NASA-TLX, where the Game-based Task showed the highest average values of *Mental, Temporal, Effort*, and *Frustration* scores. However, the *Physical Demand* scores were not reported properly by all participants, claiming that there is no physical demand in HCI contexts.

Table 3.1 Self-reported scores for the competency level of the participant against given tasks.

Task	Competence	Pre-Difficulty	Post-Difficulty
OS	7.5	2.7	3.6
Online	8.9	3.1	5.1
Excel	6.7	3.5	3.6
Pacman	5.7	5.8	7.0

Table 3.2 Self-reported scores average for NASA-TLX across the given tasks.

Task	Mental	Physical	Temporal	Performance	Effort	Frustration
OS	7.6	0.0	9.4	4.7	7.6	7.0
Online	8.6	0.1	10.9	6.5	9.0	9.6
Excel	9.5	0.0	8.6	4.6	8.0	8.3
Pacman	12.6	0.5	13.1	8.4	12.8	12.0

In addition to the NASA-TLX, the SAM scale was used by the participants after each session was completed to report the level of *Valence, Arousal* and *Dominance*. Subsequently, there is a slight variance in the scores obtained across the tasks, however, the average of the *Arousal* scores is higher than the average of the *Valence* and the average of the *Dominance* as shown in Table 3.3.

Consequently, the collected dataset comprises data from different channels, which are video recording of the participant face using normal webcam and the eye gaze tracking data using an infra-red eye tracker, together with several scores using different self-reporting tools. 15 participants were wearing glasses produced 60 recorded sessions. In addition, 33 recorded sessions have invalid eye gaze tracking data that were excluded from the eye gaze analysis experiments. Therefore, this dataset represents the entry point of the research presented in the thesis, which facilitates the validation of hypotheses and the examination of different techniques along the exploration of the methods and approaches adopted to answer the identified research questions.

Table 3.3 Self-reported scores average for the SAM scales across the given tasks.

Task	Valence	Arousal	Dominance
OS	4.9	6.2	5.3
Online	5.2	6.0	5.7
Excel	5.3	5.8	5.7
Pacman	5.9	6.7	5.7

Chapter 4

Facial Expression Analysis and Emotion Detection

4.1 Overview

As discussed in the Literature Review in Chapter 2, there is a range of approaches and techniques used for the purpose of user modelling, with contrasts in performance and efficiency among these approaches. Particularly within the Affective Computing domain, different technologies have been used to detect relevant human emotions and states. However, many of these technologies are noisy, intrusive and obtrusive, which may further produce a biased effect rather than detecting the actual emotions (Hernandez et al., 2014). Therefore, an intelligent system should aim to capture user states throughout the user experience without creating additional stress or bias. Consequently, the studies presented in this chapter investigate and validate the hypothesis that visual-based channels are feasible to capture user' affective states within HCI contexts, in particular facial expression analysis. In addition, facial expression analysis experiments will investigate machine learning techniques trained using data captured from a webcam, in order to construct a User Model to represent affective states.

Facial expressions, which are deemed the most effective input channel in the domain of Affective Computing, are generated from the movements of facial muscles from different regions of the face; primarily the mouth, nose, eyes, eyebrows, and forehead wrinkles. Human emotional states (i.e affective states) are associated with psychological and physiological changes make them sometimes visible and able to be seen or noticed (Davidson, 1992; Harmon-Jones et al., 2012). Subsequently, due to the correlation between facial expressions and human emotions, it is expected that automatic facial expression analysis will endow

computer systems with the ability to recognise human affective states. There are two parts to this chapter. In the first part a number of studies of facial expression analysis is presented. The use of facial point distance descriptor as a representation of facial expressions is considered, and a range of supervised machine learning techniques are investigated. The experimental results indicate a higher level of classification accuracy and robustness is achievable, in comparison to using standard Cartesian coordinates from the facial landmark points. On the basis of the experimental results, the best facial expression classifiers were selected for further investigation.

In the second part an investigation into the use of facial expression analysis for detecting human affective states in HCI contexts is presented, where users interacting with different UIs and attempting different tasks. Results are discussed and findings are drawn on facial expression analysis and emotional state recognition alongside with the detection of users' states in relation to HCI.

4.2 Methodology

The experiments presented in this chapter aim to investigate the use of facial expression to recognise user's affective states whilst interacting with different UIs. Subsequently, two facial-based feature representations, points-based and distance-based, have been examined because points-based feature representation is not robust enough to recognise facial expressions of individuals whose faces were not included within the training set. Subsequently, an alternative approach (i.e. distance-based) has been proposed and tested. Moreover, the two feature descriptors were exploited against a range of classification techniques, including single classifiers and an ensemble classification technique incorporating hierarchically structured SVM classifiers. The hierarchical structure classification model decomposes the classification problem into smaller micro-decisions that are made by specialised classifiers, with each one targeting a specific label in the dataset that has been trained independently. Subsequently, the use of a hierarchical structure benefits the whole system from the advantages of making some features more discriminative for specific classes. Consequently, the models were trained and tested on the widely used benchmarking datasets CKPLUS and KDEF as they comprise labels of different expressions, where annotations have been provided by Affective Computing experts or acted to depict particular expression, in contrast of using the collected dataset of this PhD described in Chapter 3 for training and validation. The comparisons have been conducted on the benchmark datasets using the commonly known evaluation metrics including *Accuracy*, *Precision* and *Recall* associated with each label from the *Confusion Matrix* of the tested models. Furthermore, the best model yielded was subsequently used to

analyse the percentages of classified emotional states across different tasks from the collected dataset presented in Chapter 3. Additionally, two Deep Learning approaches have been investigated for the purpose of facial expression analysis as they represent the state-of-the-art in image classification: Convolutional Neural Networks for facial expression classification, and Long-Short-Term-Memory Recurrent Neural Networks for facial expression activation. Additionally, the existence of a relationship between the captured affective states and the self-reported scores has been investigated through Pearson correlation analysis to check the credibility of using the self-reporting tools in the context of HCI.

4.3 Facial Expression Analysis

Facial expression analysis typically involves a pipeline of four stages, as shown in Figure 4.1 (Mäkinen, 2008). The first stage is to detect faces in a given image and estimate their location in the image (Yang et al., 2002). The second stage is normalisation, comprising geometric normalisation to reduce alignment effects, and lighting normalisation to reduce illumination effects, which aims to enhance the quality of the overall analysis process. The third stage is feature extraction and representation, including locating facial landmark points, which fits the subsequent classification method. The fourth stage is to classify and categorise extracted features using a prediction model that is trained on a set of images.

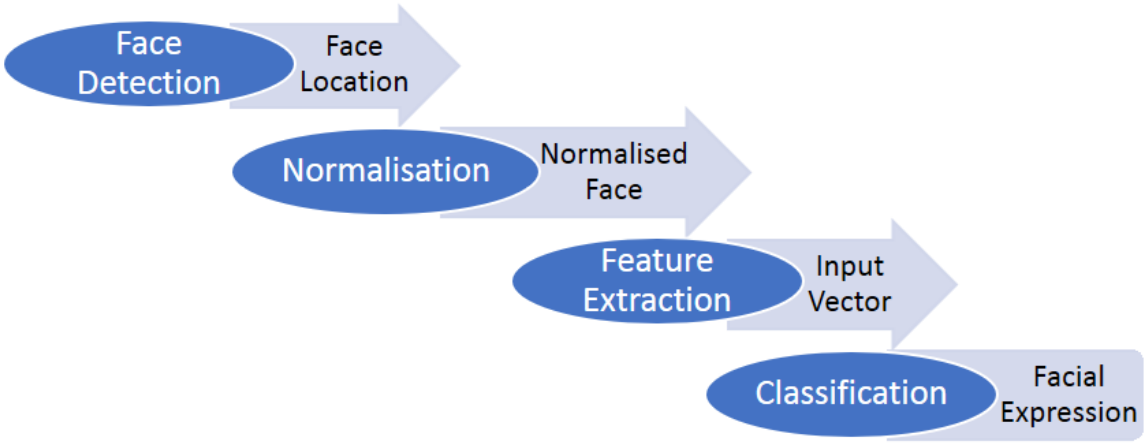
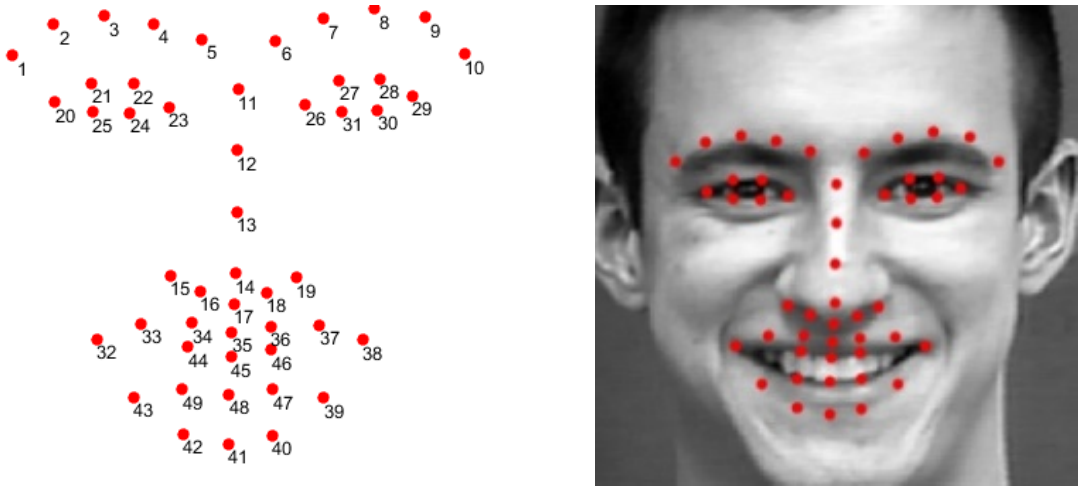


Fig. 4.1 Facial expression analysis pipeline (Mäkinen, 2008).

4.3.1 Feature Extraction and Representation

Feature extraction from images is a fundamental basis for the facial expression analysis pipeline. Different factors that are related to feature extraction techniques that substantially



(a) The 49 facial landmark points detected. (b) Detected points on a sample face.

Fig. 4.2 The 49 facial landmark points detected on a sample face.

make an impact on the overall process such as feature type and representation, length of the produced feature vector and the quality of the image. Subsequently, the adopted techniques that have been used to extract features from facial images will now be discussed.

Geometric-Based Facial Feature Detection

Geometric-based techniques extract features from local regions of interest and then generate a representation based on facial geometrical properties (Jiang et al., 2011; Shan et al., 2009). In other words, geometric-based techniques for facial expression analysis are based on locating the facial landmark points (referred to as *fiducial points*) and determining the location and the shape of associated facial components including the *eyebrows, eyes, nose, lips* and *mouth*.

Asthana and Zafeiriou (Asthana et al., 2014) developed a tool that serves as a geometric-based technique, entitled *"Chehra"*, which is a facial landmark detector based on discriminative facial deformable models, trained using a cascade of linear regressions. As shown in Figure 4.2, the *Chehra* detector locates 49 facial landmark points as follows: [1-10] *eyebrows*, [11-19] *nose*, [20-31] *eyes*, [32-43] *mouth outer lips* and [44-49] *mouth inner lips*.

Distance-Based versus Point-Based Facial Feature Descriptors

Facial point location can be directly used as Cartesian coordinates for machine learning. Using 49 coordinates of facial landmark points where each coordinate is a XY tuple, and the captured facial landmark points posteriorly converted into a sequence of tuples that produces a feature vector of length 98-dimensional features. Subsequently, the generated

feature vectors can be deployed and used by the classification model for training and testing. However, this approach is not robust enough to permit recognition of facial expressions in images that were not provided in the training data set. In other words, the generalisability of system is not very efficient to recognise the facial expressions of different individuals (Tian et al., 2001; Whitehill et al., 2013). This is due to the fact that the constellation of these points varies among the myriad of facial shapes that comprise different facial morphologies (Salah et al., 2009). Therefore, researchers have attempted to find alternative descriptors. Martinez proposed a shape-based model, which defines *configural features* that represent intra-facial component distances, in particular, the vertical distances between eyebrows and mouth (Martinez, 2011). Additionally, (Romera-Paredes et al., 2012) have used distances between facial points for the recognition on their dataset. Consequently, in this PhD a novel method is proposed that uses a complementary representation based on finding the separation between all facial landmark points. Subsequently, the *Euclidean distance* metric for two points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ is given as:

$$d_{1,2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

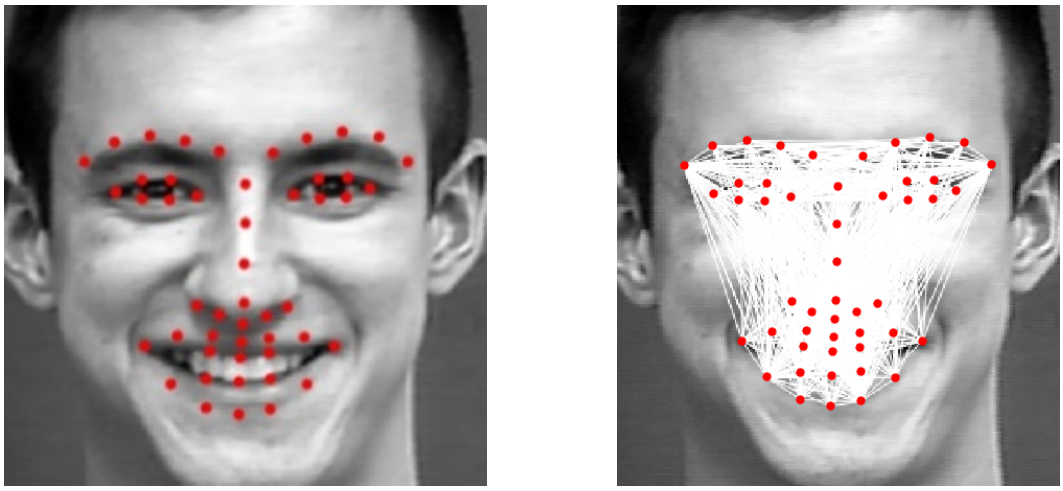
Consequently, to facilitate training the classifier using distances among facial points requires the production of a feature vector that represents the Euclidean distances between all points, as illustrated in Figure 4.3. As a result, combinations of 49 coordinates C_2^{49} will produce a vector of 1176-dimensional features. This novel approach in the feature representation increases the number of yielded features in the vector, which may potentially improve the classification accuracy as well as potentially enhancing the generalisability of the facial expression recognition model.

4.3.2 Facial Expression Classification

A a number of classification approaches has been investigated: using single classifiers, hierarchical classification, and deep learning.

Single Classifier

Extracted facial features should be exploited with machine learning techniques that support pattern recognition. Subsequently, machine learning is explored and exploited in order to build an effective predictive model, which is an approximation of the affective state of a user derived from facial expression analysis. Within this study WEKA, a suite of machine



(a) The 49 facial landmark points detected. (b) Lines between facial points combinations.

Fig. 4.3 Distance-based feature representation generated from fiducial points (i.e. facial point coordinates).

learning software, was used for the purpose of exploration and to ascertain the accuracy of various machine learning methods. Specifically, the following techniques were used:

1. *SimpleLogistic* classifier, which is implemented by Landwehr et al. (Landwehr et al., 2005), and which uses a stage-wise fitting process to build logistic regression models.

2. *Logistic Regression Tree (LMT)* classifier, which comprises classification trees with logistic regression functions at the leaves.

3. *Multi-Layer Perceptron*, which uses back-propagation to train a neural network consisting of multiple layers of nodes using the sigmoid logistic function.

4. *Support Vector Machine (SVM)*, which is widely used in data analysis and binary/multiclass classification. SVM tries to find an optimal hyperplane that separates labelled training data categories with the maximum margin. Two variations of SVM have been used in this study:

   (a) *Sequential Minimal Optimization (SMO)*, proposed by Platt (Platt, 1998).

   (b) *C-Support Vector Classification (C-SVC)* with linear kernel, which is available in the LIBSVM library developed at National Taiwan University that can be deployed in WEKA (Chang and Lin, 2001; EL-Manzalawy, 2005). For simplicity, we hereafter refer to C-SVC as SVM.

The datasets used throughout the experiments are the KDEF and CKPLUS. However, as CKPLUS dataset includes 7 classes of affective state it named as CKPLUS-7. Moreover, as sequences in the CKPLUS dataset are captured from a neutral state as the start frame is followed by sequences until the peak frame, therefore, an additional 112 images were annotated as *neutral*. This modification makes it a slightly different dataset from the CKPLUS dataset, hence it will be referred to as CKPLUS-8, as it contains 8 classes of affective state. Subsequently, extracted facial features from the set of images in the datasets were input to the set of prescribed machine learning algorithms in WEKA, with the aim of comparing different classifiers in terms of classification performance. 10-fold cross validation was used in all experiments. Additionally, a 95% confidence interval of classification results has been used in order to show the lower and upper limits.

Results

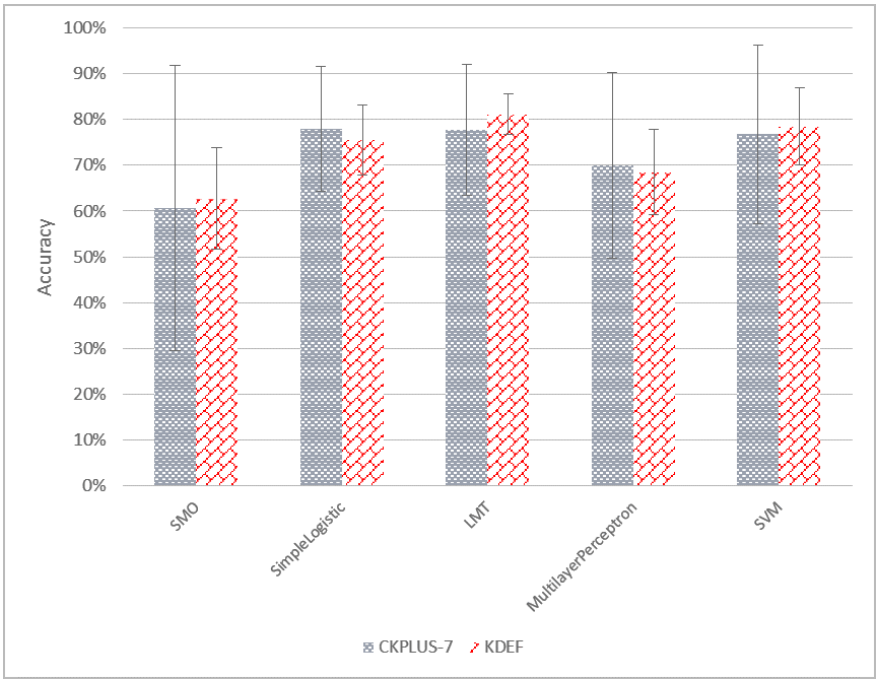


Fig. 4.4 Automatic facial expression classification accuracy (using 10-fold cross validation), of different machine learning techniques on CKPLUS-7 and KDEF datasets, together with lower and upper bounds using a 95% confidence interval.

The classification accuracies obtained, along with the corresponding lower and upper bounds, are depicted in Figure 4.4. An interesting observation from the results in Figure 4.4 is that some techniques outperform others, depending on the dataset, yet SVM provides comparable performance on both the CKPLUS-7 and KDEF datasets. For instance, classification

Table 4.1 Snapshot of the intermediate dataset generated by binary classifiers during the first stage, which is subsequently used to train the final aggregator classifier

Afraid/ Others	Angry/ Others	Disgust/ Others	Happy/ Others	Sadness/ Others	Surprise/ Others	Final State
afraid	others	others	others	sadness	others	afraid
others	others	disgust	others	others	surprise	surprise
others	others	others	happy	others	others	happy

accuracies for CKPLUS dataset are 60.59%, 77.9%, 77.68%, 69.93% and 76.77% using SMO, Simple Logistic, LMT, Multilayer Perceptron and SVM respectively. Similarly, using the same set of classifiers the accuracies achived for KDEF dataset are 62.76%, 75.41%, 81.12%, 68.47% and 78.47%.

Hierarchical Parallelised Binary Support Vector Machines

In this part of the study an approach that combines a set of SVM classifiers was implemented in order to improve the overall performance of the system. Initially, binary labelled datasets (equal to number of labels i.e. classes) were produced from the original datasets (i.e. CKPLUS and KDEF datasets). Where each dataset has a pair of labels: either one of the labels of the original dataset (which is an emotional state), or others. Afterwards, each binary dataset produced was used to train a binary SVM classifier. Subsequently, several binary classifiers are trained on differently labelled binary datasets, which represent the first stage of the hierarchical classification process. The second stage is a multi-class classifier that gives the final result.

Thus, the resulting classification framework, Hierarchical Parallelised Binary Support Vector Machines (HPBSVM), operates in two stages as given in Figure 4.5. During the first stage, binary SVM models are constructed from annotated data, with one model employed for each emotional state within the dataset. Furthermore, during the second stage, a multiclass SVM model is constructed to predict the state based on the combination of the decisions given from the binary SVM models of the first stage.

Table 4.1 shows a snapshot of the intermediate dataset, which has been generated by executing the aforementioned parallel classifiers. Subsequently, the length of the intermediate vectors equals the number of binary classifiers utilised in the first stage. The output from the first stage, which is a multiple components vector produced by the set of binary classifiers, represents the intermediate feature vector that is used to train the second stage multi-class classifier, entitled the Aggregation Classifier, which gives the final decision on the emotional state.

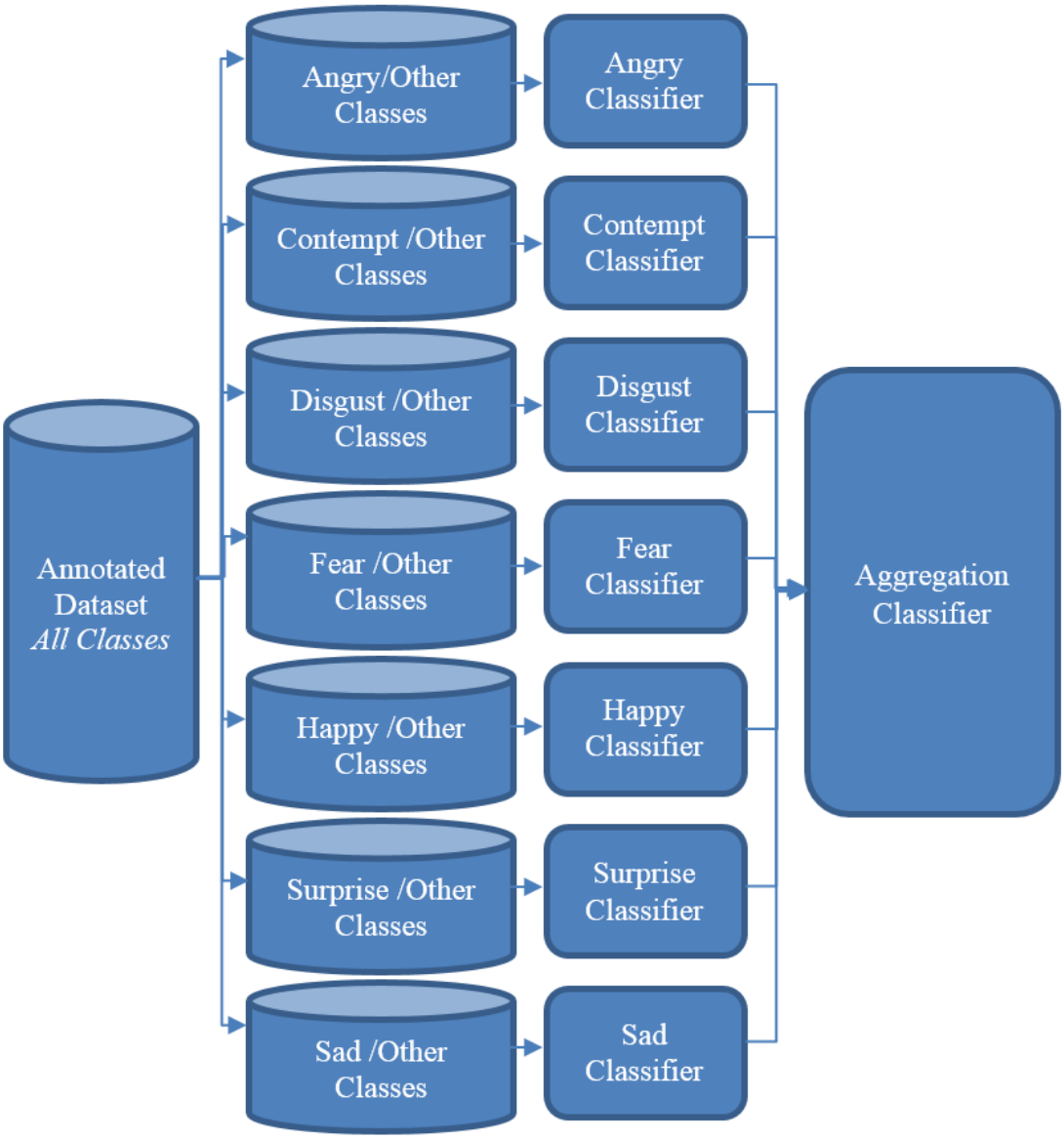


Fig. 4.5 Hierarchical parallelised binary support vector machines (HPBSVM) for facial expression classification.

Table 4.2 Classification accuracy of different datasets using Point-coordinates/Distance-based with SVM/HPBSVM classification models. Distance-based feature with HPBSVM outperforms Point-coordinates with SVM with statistically significant improvement (P<0.001).

Dataset	Point-coordinates & SVM	Distance-based & SVM	Point-coordinates & HPBSVM	Distance-based & HPBSVM
CK-7	80.75% $\pm$ 4.28	82.86% $\pm$ 4.09	96.02% $\pm$ 2.24	96.94% $\pm$ 2.01
CK-8	76.76% $\pm$ 3.95	78.36% $\pm$ 3.85	95.22% $\pm$ 2.06	95.67% $\pm$ 1.98
KDEF	78.47% $\pm$ 2.57	81.84% $\pm$ 2.42	85.71% $\pm$ 2.20	96.22% $\pm$ 1.22

This approach facilitates more efficient classification of the whole system, and achieved a higher classification accuracy as presented in the following section. The use of a hierarchical structure benefits the whole system from the advantages of making some features more discriminative for specific classes. In other words, this framework decomposes the overall problem into smaller micro-decisions that are made by specialised classifiers, which are trained independently.

Results

The HPBSVM was evaluated against the normal, single classifier scheme (i.e. SVM), using both the distance-based feature descriptor and a feature vector comprising Cartesian point coordinates. Subsequently, the experimental results presented in Table 4.2 detail the four possible combinations resulting from the two feature representations, point-based and feature-based, and the two classification methods, single classifiers and the HPBSVM approach. From the results obtained, it can be observed that the distance-based feature descriptor with HPBSVM outperforms point coordinates with SVM. Yet the improvement caused by the HPBSVM is substantial, with comparison to the improvement resulting from using distance-based descriptor. Therefore, further experiments were carried out on the three datasets (CKPLUS-7, CKPLUS-8, and KDEF) using both single SVM-based classifier and HPBSVM models.

The results obtained are illustrated in Figure 4.6, which shows that the hierarchical model achieves 96.9% while the normal model achieves 82.9% for CKPLUS-7. Moreover, the hierarchical model achieves 95.7%, while the normal model achieves 78.4% for CKPLUS-8. Lastly, the hierarchical model achieves 96.2% for KDEF, whereas the single SVM-based model achieves 81.8%.

Furthermore, Table 4.2 shows the classification accuracy of the four possible combinations of feature representations (point-coordinate and distance-based) and machine learning models (SVM and HPBSVM), where the use of HPBSVM technique considerably improves the

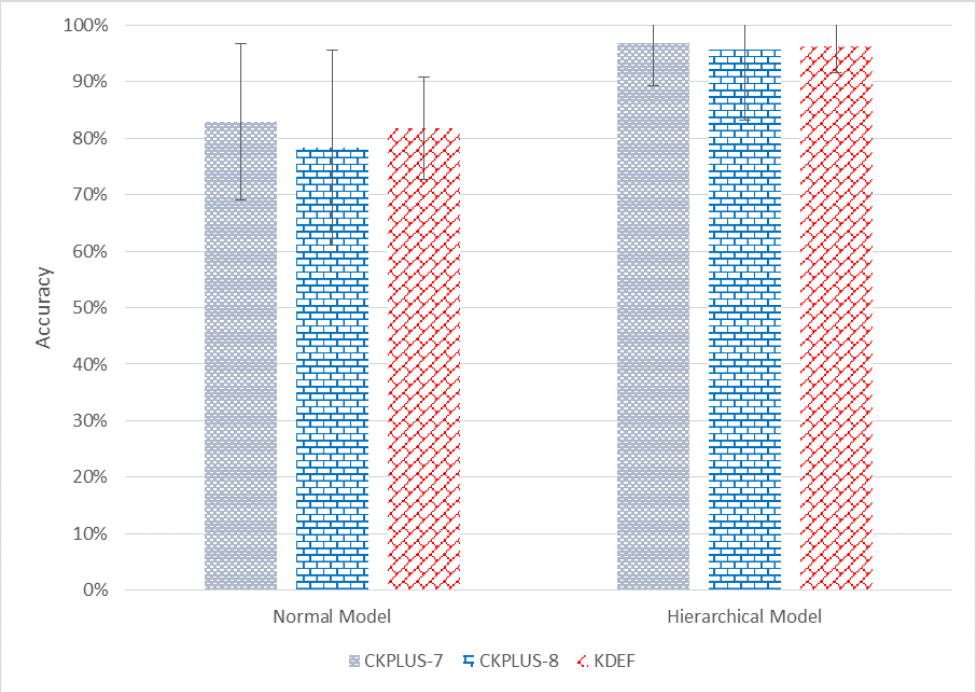


Fig. 4.6 Facial expression classification accuracy, using normal and hierarchical models, on CKPLUS-7, CKPLUS-8 and KDEF datasets; together with lower and upper bounds using a 95% confidence interval.

classification accuracy from 82.86% to 96.94% for CKPLUS-7 dataset. Likewise, for CKPLUS-8 dataset improved from 78.36% to 95.67%. Similarly, from 81.84% to 96.22% for KDEF dataset. Table 4.3 and Table 4.4 show the confusion matrices from the classification experiment which uses distance-based features with single SVM, and the classification experiment which uses distance-based features with HPBSVM respectively.

HPBSVM with distance-based descriptor achieves higher accuracy rates across all three datasets. Moreover, the achieved results are comparable to the work presented in the literature, particularly on the CKPLUS-7 dataset such as the 91.2% (Liew and Yairi, 2015), 97.35% (Ghimire and Lee, 2013) and 95.1% (Shan et al., 2009). Consequently, this technique assures a credible result to be used with other data such as the one collected in the Data Collection Study presented in Chapter 3, and DEAP dataset mentioned in Chapter 2. Therefore, this scheme has been adopted in the experiments that endeavour to categorise facial expressions from videos recorded during several different user interface contexts.

Table 4.3 Confusion matrix, precision, recall and F-measure of CKPLUS-7 classification using distance-based feature representation with single SVM.

	Anger	Contempt	Happy	Sadness	Surprise	Disgust	Fear
Anger	**29**	4	1	7	0	4	0
Contempt	2	**14**	0	1	1	0	0
Happy	2	0	**65**	0	0	0	2
Sadness	5	1	0	**17**	1	1	3
Surprise	0	2	0	2	**79**	0	0
Disgust	6	1	3	0	1	**48**	0
Fear	0	1	2	1	2	0	**19**
Precision	0.66	0.61	0.92	0.61	0.94	0.91	0.79
Recall	0.64	0.78	0.94	0.61	0.95	0.81	0.76
F-Measure	0.65	0.68	0.93	0.61	0.95	0.86	0.78

Table 4.4 Confusion matrix, precision, recall and F-measure of CKPLUS-7 classification using distance-based feature representation with HPBSVM.

	Anger	Contempt	Happy	Sadness	Surprise	Disgust	Fear
Anger	**45**	0	0	0	0	0	0
Contempt	0	**17**	0	1	0	0	0
Happy	1	0	**61**	6	0	1	0
Sadness	0	0	0	**28**	0	0	0
Surprise	0	0	0	1	**82**	0	0
Disgust	0	0	0	0	0	**59**	0
Fear	0	0	0	0	0	0	**25**
Precision	0.98	1	1	0.78	1	0.98	1
Recall	1	0.94	0.88	1	0.99	1	1
F-Measure	0.99	0.97	0.94	0.88	0.99	0.99	1

Deep Learning for Facial Expression Classification

Convolutional Neural Networks Transfer Learning

Further to the aforementioned investigations for facial expression analysis methods, the research also examined the use of state-of-the-art machine learning techniques for facial expression analysis. Convolutional Neural Networks (CNN), a method for Deep Learning, stands in the state-of the-art for image classification in general. On one hand, CNN training needs strong computation resources because CNN deals directly with image files and performs automatic feature extraction. On the other hand, *Transfer Learning* is a practical and efficient approach that is aimed at achieving high classification accuracy with small effort. It trains an already trained deep learning model only to tune the weights and parameters of the final layer using a customised but usually small dataset. In this study, a CNN computer vision model was trained for *TensorFlow* called *MobileNets* (Howard et al., 2017), on the same benchmark datasets (CKPLUS7, CKPLUS8 and KDEF) separately. *MobileNets* models are optimised models, so can be re-trained on small and limited resources quickly, achieving high accuracy without the need for large computation resources.

Variation *MobileNets* depends on two hyper-parameters that influence the training time and the model accuracy. The first hyper-parameter is Resolution Multiplier, which is responsible for the resolution of the input image as well as the internal representations of each layer in the network. Current version of *MobileNets* have four different possibilities for *resolution multiplier*, which are 128, 160, 192 and 224, and due to similarity of the results only *resolution multiplier* using 128 and 224 were reported, which represent the lowest and the highest. The second hyper-parameter is *Depthwise Separable Convolution*, which enforces filters on each input channel that factorise a standard convolution into a depthwise convolution, which is subsequently combined with pointwise convolution (i.e. 1 x 1 convolution), to eventually generate a feature representation (Sifre, 2014). There are four options for the depthwise separable convolution parameter that have been utilised within the MobileNets-based experiments, which are 0.25, 0.50, 0.75 and 1.0.

Results

In the experiments models were trained using different values of the two hyper-parameters, where the four mentioned options of depthwise separable convolution were tested against 128 and 224 resolution-multiplier values, classification accuracy, along with the corresponding lower and upper bounds calculated using 95% confidence interval, are depicted in Figure 4.7 and Figure 4.8 respectively. As may be observed the classification accuracy achieved for CKPLUS7 ranges from 79% to 86% using 224 resolution multiplier, and were slightly

better using the smaller value of 128, which ranges from 86% to 90%. CKPLUS8 produced approximately the same classification accuracies in both case, where classification accuracy ranges from 79% to 88%. On the other hand, results on KDEF were as expected, where higher resolution produced higher accuracy ranging from 76% to 92% using resolution multiplier of 224, and ranging from 71% to 83% using multiplier of 128.

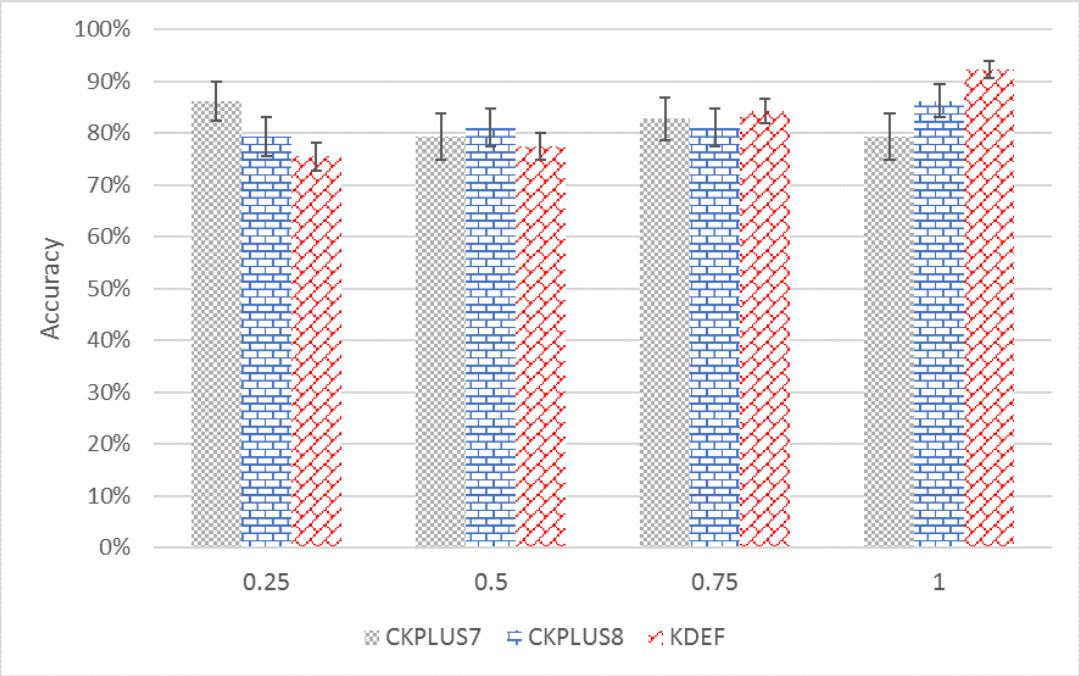


Fig. 4.7 Facial expression classification accuracy achieved by training transfer learning of MobileNet model using Resolution Multiplier of 224, with Depthwise Separable Convolution of (0.25, 0.5, 0.75 and 1.0) on CKPLUS-7, CKPLUS-8 and KDEF datasets.

Facial Expression Activation as a Time Sequence using Recurrent Neural Networks

Proceeding from the CNN method given in the previous section, a temporal-based facial expression analysis has been investigated as well, attempting to determine if a facial expression can be captured from a set of consecutive images that show the activation of the expression from a neutral state. In this study facial expression is considered as recorded in a sequence of images and recurrent neural networks employed for classification purposes. The instances used for training and testing are time series objects, i.e., sequences of images. The CKPLUS dataset was presented as a set of sequentially captured snapshots from a neutral state (referred to as Onset Frame) to the Peak Frame as illustrated in Figure 4.9. The last frame is presumed to represent the actual facial expression if it includes the action units that conform to the

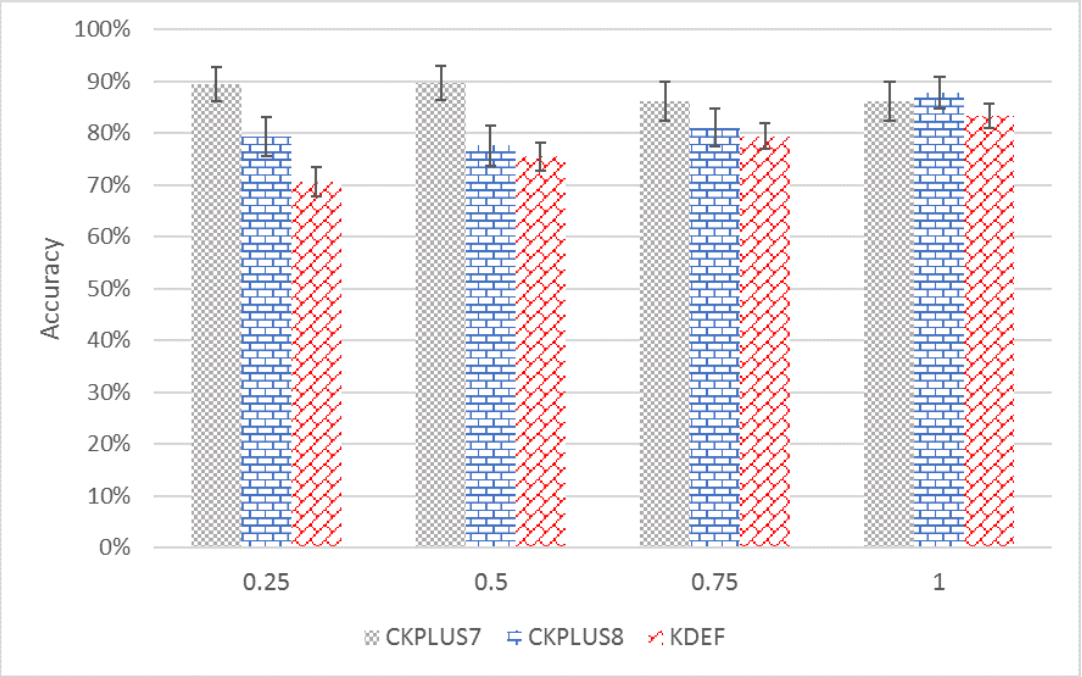


Fig. 4.8 Facial expression classification accuracy achieved by training transfer learning of MobileNet model using Resolution Multiplier of 128, with Depthwise Separable Convolution of (0.25, 0.5, 0.75 and 1.0) on CKPLUS-7, CKPLUS-8 and KDEF datasets.

facial expression according to FACS system. Therefore, a classification experiment was conducted on CKPLUS dataset (i.e. CKPLUS7 because all objects start from neutral state).

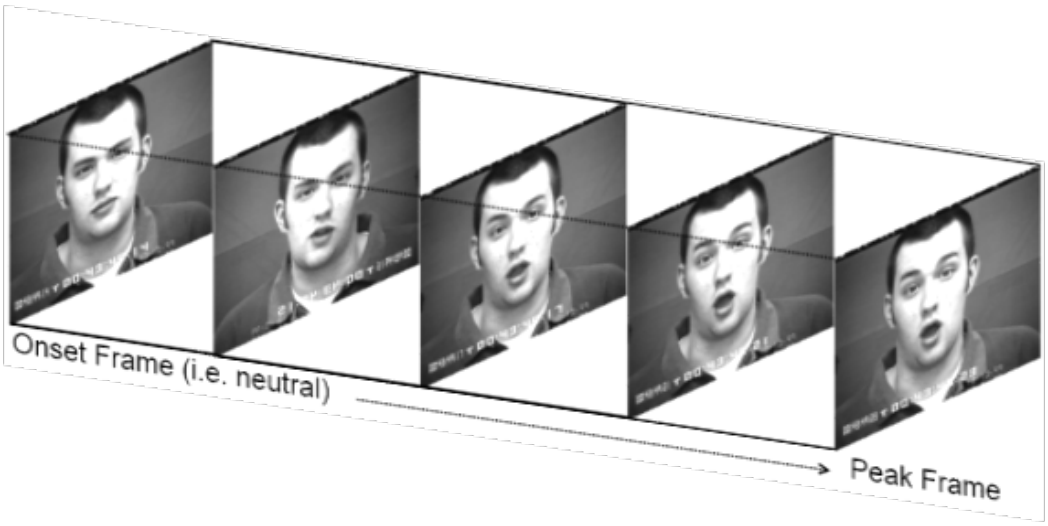


Fig. 4.9 Images sequence of a Surprise facial expression from CKPLUS dataset, which starts from neutral state (called onset) to the peak frame that manually coded and labelled to a facial expression according to FACS and facial action units.

Table 4.5 Parameters values used in training the LSTM-RNN of the time sequence objects extracted from CKPLUS dataset that represents the facial expression activation.

Parameter	Value
Model type	Multi-Layer Network
Layers	6
Input size	1176
Output size	7
Total parameters	13289707
Optimization algorithm	Stochastic Gradient Descent
Optimization algorithm iterations	10
Learning rate	0.005
Layers activation function	TANH

Following the same feature extraction procedure, for each frame, facial fiducial markers were extracted, then generated distance-based features. As shown in Figure 4.10, a multivariate time sequence object (composed of 1176 features) was generated, which corresponds to one label (i.e. facial expression). These instances were forwarded to train a *Long Short-Term Memory Recurrent Neural Network (LSTM-RNN)*.

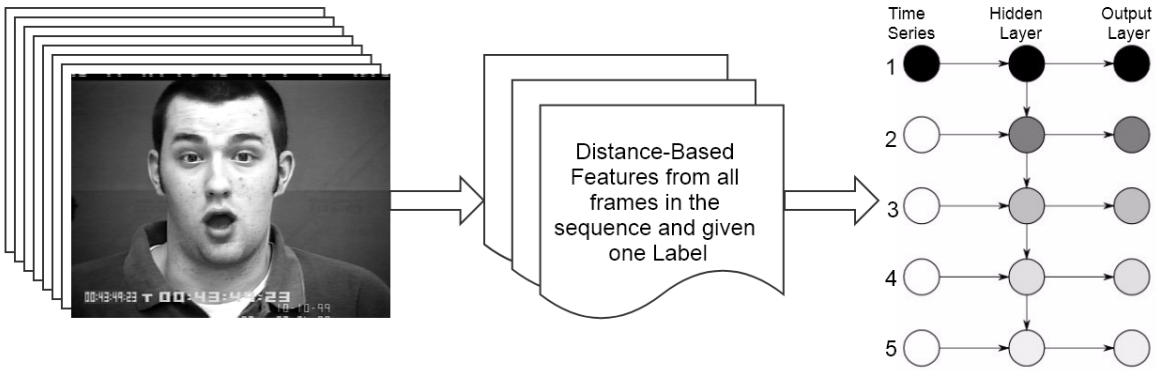


Fig. 4.10 An overview of the implementation of the facial expression recognition expression experiment carried out, which uses the sequences captured from onset frame till the peak frame as provided within CKPLUS dataset.

The network configuration is given in Figure 4.11, which comprises 6 layers (layer 0 to layer 5). The input size of the first layer (i.e. layer 0) should match the number of features extracted from the image sequence, which is 1176. The input size of the subsequent internal layers decreases gradually until the number of labels in the dataset is reached, which is 7 facial expressions. The training parameter values that were used to achieve reported results are given in Table 4.5.

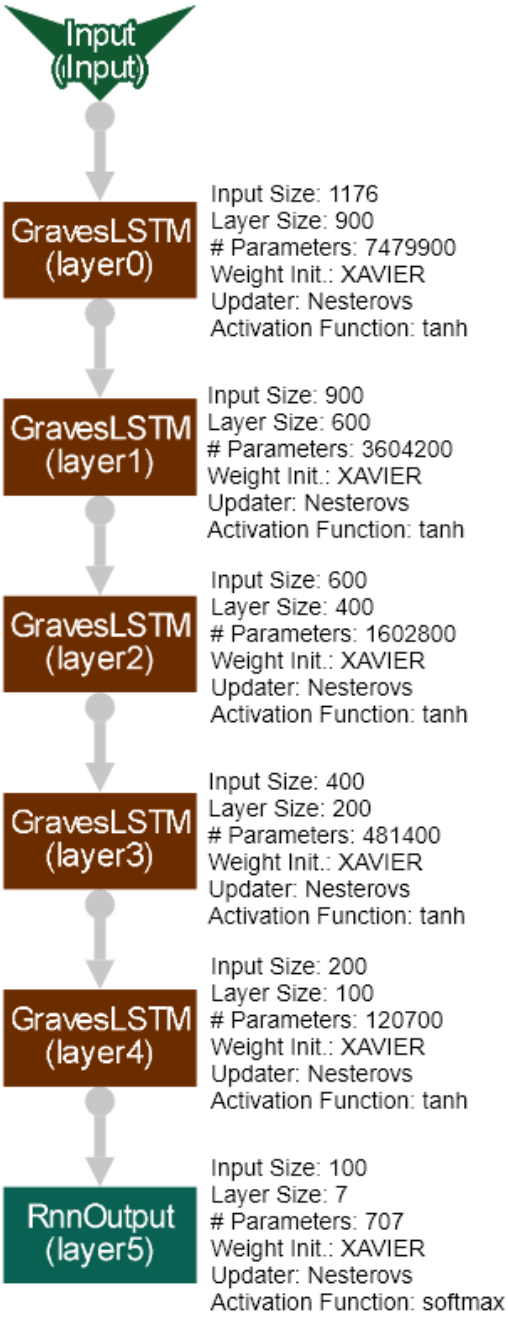


Fig. 4.11 Graph shows the details of our neural network structure with information of each layer. Input features are the 1176 distance-based features and the number of outputs are 7 that represents expressions in CKPLUS7 dataset.

Results

A number of designs of the network have been examined to train the generated dataset. Monitoring the convergence/divergence of the loss value indicates the potential of getting good results or not. Therefore, after a number of trials the best results of accuracy have been achieved, which ranges from 94% to 99% using different training/testing split of 0.65, 0.70 and 0.75.

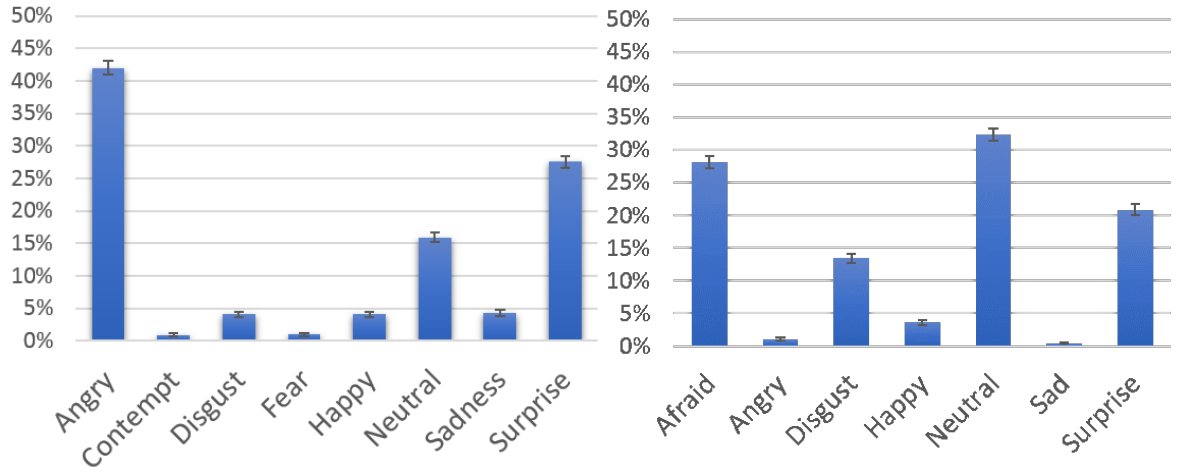
4.4 Facial Expression Classification during Human-Computer Interaction

Relating to the first research question, *what are the emotional states that can be detected and utilised for the purpose of modelling the user within an adaptive HCI context?* , the study presented below aims to explore the actual facial expressions manifested by different users whilst they are interacting with different interfaces. Subsequently, the current study uses the trained models mentioned in Section 4.2.2 to automatically classify emotions via facial expressions, by analysing video frames that were acquired whilst users attempted and interacted with each of the computer-based tasks previously described in the Chapter 3. Hence, the main objective of the classification is to model the affective states of users within a HCI context, and the association between the self-reporting and the facial expression. Subsequently, the tasks were categorised into two groups as *active* and *passive* interactions.

- *Active interaction*, which represents the situations where the user is working and undertaking an effort to carry out the task. These situations include: general operating systems tasks (tagged as **OS**), online shopping tasks (tagged as **Online**), and spreadsheet tasks (tagged as **Excel**), and entertainment task playing Pacman (tagged as **Pacman**).

- *Passive interaction*, which represents situations where the user does not exert effort, which refers to the recordings contained within the DEAP dataset (Koelstra et al., 2012), where the subjects passively watched videos on **Youtube** without explicit interaction.

HPBSVM with distance-based features were used to train two models using the validated datasets CKPLUS-8 and KDEF, so that each resulting model can automatically classify the video frames of the recordings obtained during the computer-based tasks at a frequency of 1 frame per second.



(a) (CKPLUS-8 model output

(b) KDEF model output

Fig. 4.12 Percentages of expressions on CK-8 and KDEF trained models applied on recordings of Online task context respectively. Results provide lower and upper bounds using a 95% confidence interval.

Results

Figure 4.12 shows the percentages of facial expressions using the earlier mentioned trained models, which is achieved by automatically classifying the video frames of the user recording while carrying out the Online task. Figure 4.12a presents the expression percentages using a classification model that is trained on CKPLUS-8 dataset. Similarly, Figure 4.12b presents the expression percentages using a trained classifier using KDEF dataset.

As depicted in Figure 4.12, different percentages of each expression have been found for both models. However, one can view these percentages differently by considering the fact that some expressions are much more precisely recognised than others. Generally, detecting states such as *happy* and *surprise* is comparably superior than detecting other states such as *contempt*, *neutral*, *fear*, *angry*, *sadness* and *disgust*, which is possibly due to the similarity in the geometric shape of these expressions. Moreover, the work presented in (Joho et al., 2009) underlined this type of grouping, by devising the expressiveness level (i.e. the strength) of the associated expressions, where these expressions belong to a *Low Pronounced* level.

Therefore, the expression labels *angry, contempt, disgust, fear* and *sadness* from the CKPLUS-8 dataset used in our analysis can be combined together to represent the *negative* state. Likewise, the labels *afraid, angry, disgust* and *sad* from the KDEF dataset can be combined together to represent the *negative* state as well. Additionally, from the Circumplex Model it may be observed that there is a common aspect among these expressions, in that

such negative labels occur on the negative side of the pleasant-unpleasant continuum, e.g. the *valence* axis, as previously shown in Figure 2.2.

Consequently, the negative states grouping (to be within the *negative* state) applied to CKPLUS-8 results as well as KDEF results. After that, the resultant percentages obtained by averaging the output of the two trained models. Accordingly, the results depicted in Figure 4.13, show percentages of each expression obtained from the videos recorded during each task using the average percentages across both the CKPLUS-8 and KDEF trained models.

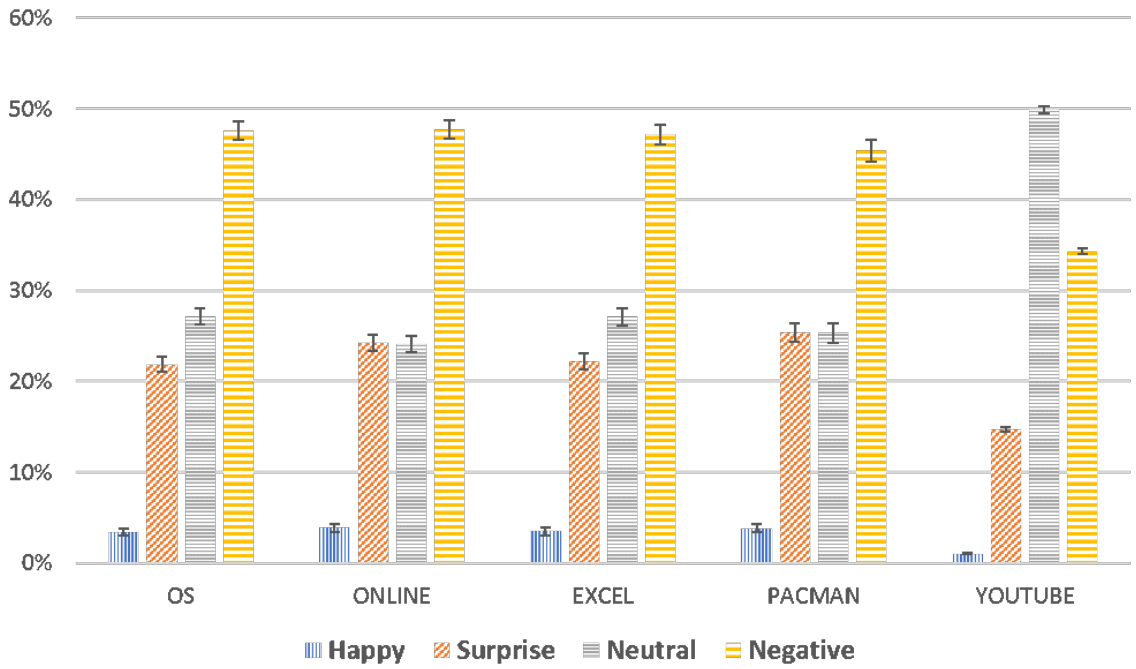


Fig. 4.13 Percentages of facial expressions across tasks by averaging outputs from the two trained models on CKPLUS-8 and KDEF datasets.

Additionally, it can be observed from Figure 4.13 that *neutral* and *negative* expressions occupy the highest percentages across the different tasks, with a greater number of *negative* expressions being shown during the tasks involving *active* interaction by participants. By contrast, in the *passive* interaction i.e. Youtube task, a greater level of *neutral* expression can be observed. Moreover, although a small percentage of *happy* expression may be observed in all tasks, within the Youtube task, the lowest percentage of *happy* expression is found. Therefore, it may potentially be surmised that, during *active* interaction tasks more variation occurs within the facial expressions of participants, than occurs within passive interaction tasks. However, such expressions might not reflect the actual feelings of the participants.

4.5 Facial Expressions versus Reported Valence and Arousal

The hypothesis of the following study, that there is an association between the detected facial expression from the facial image of the user and the self-reported scores for *Arousal* and *Valence*, subsequently relates to the overarching aim of the research concerning the affective states of the users in a HCI context. As previously described, each participant carried out a self-reported assessment after performing each task using the SAM scale; this is the case in both the dataset obtained from the Data Collection study and the DEAP dataset. Thus, each recorded video from each task is associated with *valence* and *arousal* scores.

Consequently, the relationship between facial expression percentages and the self-reported *valence* and *arousal* scores given by the participants for each task were further investigated as described herein. However, during analysis, rather than using the actual SAM ranges, i.e. [1.0-9.0], a mapping of the reported values into three ranges (*Low, Medium* and *High*) was used. Correspondingly, classification performance improves during supervised learning when the number of target classes is reduced (Aha, 1992), especially when the combined classes have common properties and similarity, as is the case with the collected dataset. Therefore, a transformation was applied to the rounded values, as illustrated in Figure 4.14, whereby values within the range [1.0-3.0] were labelled as *Low*, values within the range [4.0-6.0] were labelled as *Medium*, and values within the range [7.0-9.0] were labelled as *High*.

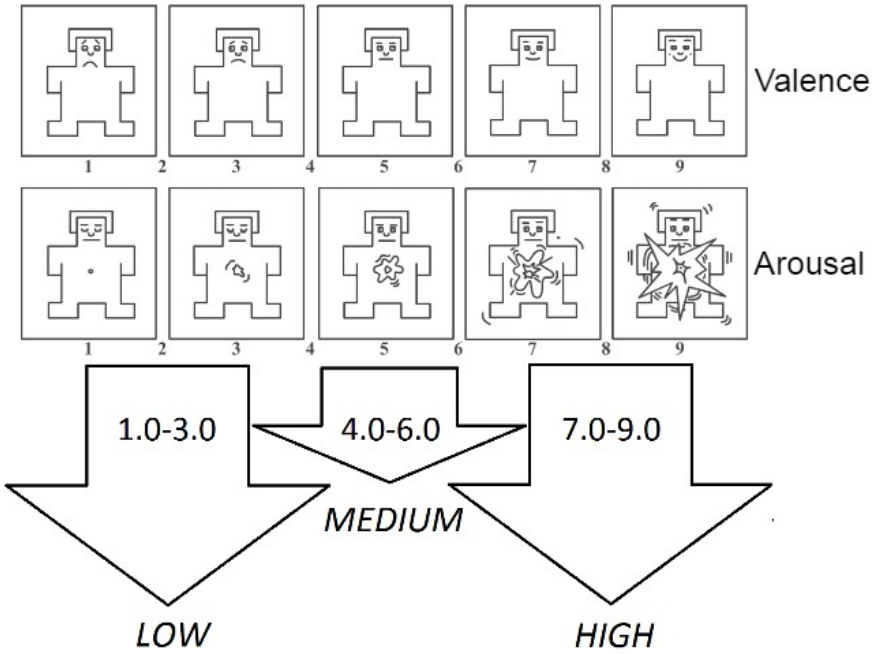


Fig. 4.14 Mapping from SAM scale value ranges into three labels (scores from [1.0-3.0] mapped to Low, [4.0-6.0] mapped to Medium, and [7.0-9.0] mapped to High.

Table 4.6 Facial expression percentages obtained based on classification of video frames using the average of the two trained models versus self-reported values of *Valence*(Val) and *Arousal*(Aro).

Stimuli	Score	Happy(%)		Surprise(%)		Neutral(%)		Negative(%)	
		Val	Aro	Val	Aro	Val	Aro	Val	Aro
OS	Low	2.64	8.31	28.33	14.54	11.65	39.99	57.38	37.17
	Med	3.72	2.53	20.07	15.65	33.84	32.56	42.38	49.26
	High	3.38	3.76	18.12	31.91	17.27	17.23	61.23	47.1
Online	Low	4.58	2.96	23.83	17.39	13.71	27.19	57.88	52.47
	Med	4.08	4.51	21.91	23.7	25.83	25.93	48.18	45.85
	High	2.53	3.08	32.94	26.85	27.62	20.42	36.91	49.65
Excel	Low	0.81	3.02	27.69	25.75	26.49	34.44	45.01	36.79
	Med	4.72	5.1	19.19	17.55	27.38	28.87	48.71	48.49
	High	2.08	1.03	26.31	28.82	26.82	21.86	44.79	48.29
Pacman	Low	2.75	4.67	18.44	32.45	23.82	31.25	54.99	31.63
	Med	1.59	3.63	28.17	15.22	26.3	30.66	43.94	50.49
	High	5.1	3.83	25.82	28.3	25.28	22.24	43.8	45.63
Youtube	Low	0.5	0.97	16.84	15.05	47.62	54.83	35.03	29.15
	Med	0.65	1.04	13.08	12.99	53.35	50.38	32.92	35.59
	High	1.91	1.12	15.38	17.61	47.2	45.28	35.51	35.99

Therefore, the facial expression percentages with the ratings that were self-reported by subjects themselves were compared for both *valence* and *arousal*. Table 4.6 gives the facial expression percentages obtained from averaging the results of the two trained models, across all the tasks (both *active* and *passive*), using the aforementioned mapping for the self-reported scores of *valence* and *arousal* respectively. In addition to this mapping, the labels corresponding to the self-reported values of *valence* and *arousal* have been used to represent the four quadrants of the Circumplex Model, as illustrated in Figure 4.15, during the analysis carried out.

Results

Subsequently, a combination of these scores, as depicted in Figure 4.15, has been aggregated versus facial expression percentages as given in Table 4.7. From the results given in table 4.7, it is apparent that the lowest percentage of frames, across all tasks, show the *happy* expression. However, somewhat surprisingly, tasks where participants self-reported high *valence* values correspond to the facial expression percentages where the *happy* expression is lowest too.

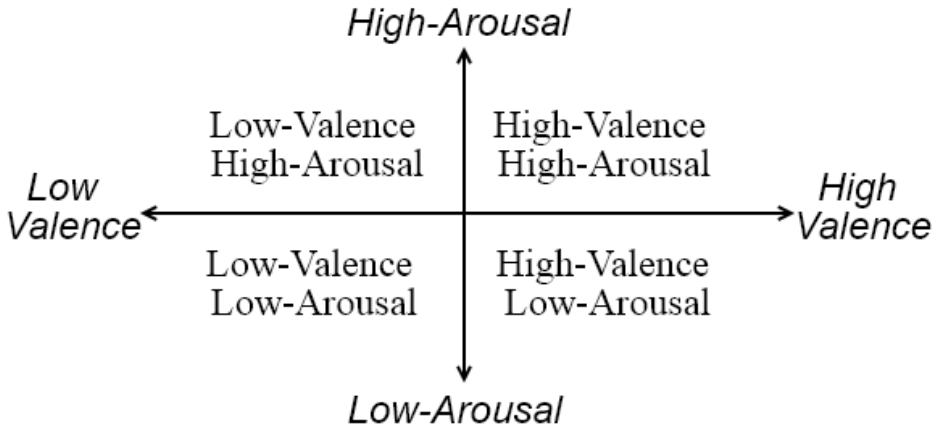


Fig. 4.15 Representation of Circumplex Model quadrants using combinations of Valence and Arousal score mappings, which utilises only the mappings corresponding to the High and Low labels.

Table 4.7 Facial expression percentages versus combination of subject self-reported values of *Valence* and *Arousal* together.

Stimuli	Valence/Arousal	Happy(%)	Surprise(%)	Neutral(%)	Negative(%)
OS	Low/Low	0	0	0	0
	Low/High	0.13	41.59	3.25	55.03
	High/Low	0	0	0	0
	High/High	2.43	20.56	27.25	49.76
Online	Low/Low	0	0	0	0
	Low/High	6.84	18.9	10.04	64.22
	High/Low	6.67	30	31.4	31.93
	High/High	1.28	30.36	25.84	42.51
Excel	Low/Low	1.95	15.61	51.46	30.98
	Low/High	0.81	26.74	25.58	46.86
	High/Low	6.65	18.97	42.36	32.02
	High/High	1.59	32.34	21.55	44.52
Pacman	Low/Low	2.45	46.57	12.25	38.73
	Low/High	2.7	13.41	24.3	59.58
	High/Low	6.49	32.15	39.23	22.12
	High/High	4.66	30.15	22.39	42.79
Youtube	Low/Low	0.61	16.5	56.9	26
	Low/High	0.77	18.2	40.48	40.55
	High/Low	1.47	13.68	52.02	32.83
	High/High	1.48	18.65	46.21	33.67

4.5.1 Discussion

By merging *Low Pronounced* facial expressions such as *contempt, neutral, fear, angry, sadness* and *disgust* as a single facial expression label, *negative*, makes the analysis of the relationships between the facial expressions and affective states of users performing computer-based tasks much more persuasive. Especially, distinguishing between these facial expressions automatically is a non-robust task (Joho et al., 2009). Referring to the overarching aim of modelling user affective states across different UIs, the hypotheses of the association between the actual and automatically detected emotional states along with the possibility of capturing the emotional state via facial expression analysis have been examined. Therefore, a number of different analysis approaches have been explored in pursuit of understanding the potential relationships between self-reported affect and the corresponding recorded facial expressions over a number of tasks.

Firstly, it was anticipated that there is a difference in the facial expression percentages between *passive* and *active* interaction contexts, due to the significant difference between the nature of the contexts. Although a general inference is that both *active* and *passive* contexts are similar with regard to the relationships between the self-reported measures and the observed facial expressions, one difference found was that the facial expression alternation that occurs within an *active* context is marginally increased over that found within a passive context where an expression mostly remains as it is.

In addition, there is a weak correlation, and inconsistency in some cases, between the individual and combined reported values of *valence* and *arousal* with facial expression found in the recordings. Subsequently, the hypothesis that facial expressions reflect the actual feelings of users within a HCI context does not meet on the collected dataset. On the other hand, it may potentially indicate the fact that when participants attempted to assess and self-report their actual feelings, they were unable to accurately distinguish and report on different emotions and feelings. This point of view certainly agrees with the argument given in (Picard, 2003) that humans often do not know how to articulate their actual feelings and affective states due to ambiguity and mixed mental activities. In addition to the significant differences in *valence* and *arousal* values that were reported for the various tasks, simultaneously the same facial expressions were still portrayed. Consequently, this concludes that the hypothesis of detecting the actual affective state via facial expression analysis in HCI context is not false.

Moreover, the results presented reveal that the accuracy of automatically detecting affective states using facial expression analysis, particularly within a HCI context, is not comparable to the accuracy achieved by facial expression analysis when acted and performed deliberately as commonly found in most work within the Affective Computing literature.

Consequently, this could be due to the nature of the relationship between humans and computers. Ultimately, humans do not (currently) exchange emotions and feelings with machines in the same manner as they do when interacting with each other.

For these reasons, researchers look to other technologies for the purpose of detecting human emotions and states (Jaimes and Sebe, 2007). Some of these technologies have very good recognition accuracy rates for certain states such as stress, which can be reflected through physiological responses such as heart rate and blood pulse volume, though they could be obtrusive and too noisy to be employed for generic HCI purposes. Additionally, some of the technologies may require an extra effect on the users and cause additional feelings for humans, rather than detecting actual feelings they are attempting to measure (Hernandez et al., 2014). Therefore, HCI approaches and designs begin, especially within the last decade, to shift the focus from the behaviours and procedures of UX, towards psychology and sociology concepts, which take into account human factors, emotions, cognitive aspects and individuals' behaviours (Harper et al., 2008; Jeon, 2017; Samara et al., 2017).

4.6 Summary

This chapter presented an investigation into user modelling and affective state detection via facial expression analysis within HCI context. Experiments presented in this chapter examined automatic affect recognition during common computer usage with results suggesting that facial expressions doubtfully indicate the actual feelings of users during interactions with computers. Subsequently, the outcome of the presented studies does not completely serve the aims that the research looks to achieve, which is capturing the actual affective states via facial expression analysis. Therefore, work is needed to determine much more appropriate and effective techniques that reason upon users' experiences during interaction with computers in order to facilitate the generation of intelligent and adaptive systems. While the work presented within this chapter investigates facial expression variation across different user interaction tasks, future work could undertake more in-depth experimentation and analysis of related cognitive load.

Although a lot of attention is paid towards deploying facial expression analysis in Affective Computing, the challenges encountered in these endeavours may not only be in terms of technical issues. On the contrary, one potential and significant challenge that may need to be addressed, for example, might be in humans' perception of computers; the perception that the computer is a machine that is a task oriented tool, which is inexpert to reason upon human feelings with the same intellect as that of another human. Thereupon,

work is progressing in different directions within various disciplines to reinforce the link between the human and the computer.

Chapter 5

Eye Tracking Data Analysis and Cognitive Load Measurement

5.1 Overview

Facilitating an adaptive form of HCI requires minimising the cognitive load of users whilst they interact with UIs in attempting to complete their tasks and achieve their goals. Therefore, it is a fundamental pillar for the intelligent machine to be able to evaluate and measure the amount of mental processing and cognitive load that a user is experiencing in order to facilitate the appropriate adaptation. Subsequently, computers and machines considered to be within adaptive HCI are presumed to recognise the cognitive state of a human user in order to interact intelligently.

Consequently, the motive that leads to the experiments presented within this chapter is related to the hypothesis that through automatic analysis of the eye gaze behaviour, machines would be able to measure the cognitive load of the users whilst they are interacting with machines. Therefore, the studies given in the current chapter reiterate examination of the hypothesis that has been tested in Chapter 4, which states can visual-based channels capture the actual user states. However, the focus in Chapter 4 was on the affective states, and the focus herein is on the cognitive load. Overall, these studies aim to find out the appropriate way to endow machine with intelligent perception capabilities and particularly handle the second research question, *"What is the best method that can capture the user state that would be useful for adaptive HCI?"*. In this regard, this chapter presents the techniques along with the results of analysing eye tracking data populated from the Data Collection Study that was previously presented in Chapter 3. Feature extraction approaches and data transformation

and mapping schemes are employed for tracking and analysis in order to recognise different user's states and measure the associated cognitive load.

5.2 Methodology

Based on the emphasis given in the literature regarding the relationship between eye-gaze tracking data and cognitive load, the study presented in this chapter attempts to address the possibility of using features extracted from eye-gaze behaviour as input for the computer to measure the cognitive load of users through pupil dilation whilst interacting with different UIs. Bivariate Correlation Analysis is being used to find the relationship between two input variables, which can demonstrate the presence or absence of a relationship between the targeted features. Subsequently, two types of data have been populated from the Data Collection Study presented in Chapter 3, which are features extracted from the eye tracker device, and features based on self-reported scores given by participants after finishing the attempted task, to validate the hypothesis that eye gaze behaviour analysis entails information about current cognitive load and mental processing being undertaken by the user. Consequently, a number of experiments have examined the amount of correlation between self-reported scores and extracted features using statistical functions including *Mean*, *Mode* and *Median* to aggregate the recorded session frames into a singular value that is associated with the self-reported scores. Additionally, as the dilation level is the core feature of interest that may help to validate the hypothesis, the pupil size values of the sequence of the captured frames of the eye-gaze tracking data have been mapped into three levels of dilation, i.e. *Low*, *Medium* and *High*. Subsequently, a number of transformation schemes have been used to make the mapping, where each one uses different splitting criteria, and the correlation coefficient of each mapping and the scores associated with a task have been compared across the different tasks.

5.3 Eye-Gaze Tracking Feature Extraction

Using an infra-red light emitting eye tracker, several measurements from the eye gaze tracking data can be populated and employed within a feature vector for modelling user's states. Features were also extracted using a range of statistical functions applied on the eye gaze tracking measurements acquired by the eye-tracker during each session of the Data Collection Study.

5.3.1 Filtering

There are some moments when the user's eyes wander out of view from the eye tracker. Also illumination effects can affect the quality of calibration. Therefore, eye gaze tracking data often times can become noisy with intermittent spikes. Hence, there needs to be preprocessing before analysis to remove such artifacts from the analysis as these moments represent outliers in the recorded data stream. Consequently, the approach adopted involved removing outlier samples based on using the upper and lower bounds that can be calculated using following equation, which uses the *mean* and the *standard deviation* of the set of acquired samples. Subsequently, any sample beyond the range of the upper and lower limits were deemed to be an outlier.

$$UpperLowerBounds = Mean \pm StandardDeviation$$

Although, *mean* plus one *standard deviation* looks very small range that covers only 0.68 of values of normally distributed samples (Leys et al., 2013). However, using this equation the spikes were removed together with the sharp edges that resulted from the signal interruption during the session by looking at the graph of pupil size throughout the overall session.

5.3.2 Eye-Gaze Tracking Measurements

Eye gaze tracking produces many measurements for each eye separately, which are used by statistical functions to extract features and patterns. Some of these features are extracted within a window of specific length as will be given later, and some of them are calculated across all the recorded session. Correspondingly, the eye tracking features utilised are based on the the work of Steichen et al. (Steichen et al., 2014):

1. **Fixations Number**: the number of eye fixations detected during a window.
2. **Fixations Rate**: the rate of eye fixation samples with regard to the total number of samples throughout a window.
3. **Pupil Size**: the size of the pupil given in arbitrary units.
4. **Saccade Length**: the distance between two fixations (depicted as d in Figure 5.1).
5. **Absolute Saccade Angle**: the angle between a saccade and the x-axis (depicted as X in Figure 5.1).

6. **Relative Saccade Angle**: the angle between two consecutive saccades (depicted Y in Figure 5.1).

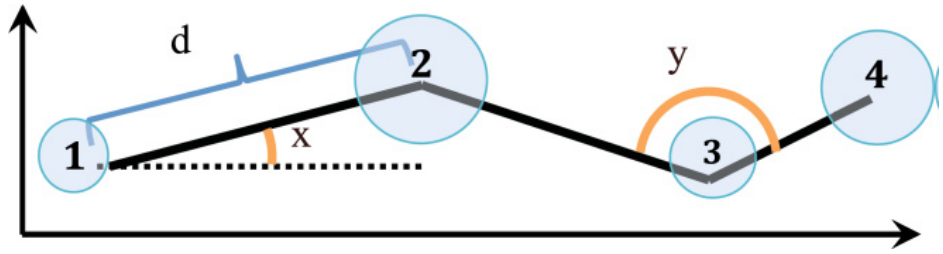


Fig. 5.1 Eye Gaze Measurements based on Fixations and Saccades (Steichen et al., 2014).

5.3.3 Statistical Metrics and Functions

The following statistical functions were applied to the measurements of eye gaze tracking data, subsequently used in the analysis:

1. **Mean** (μ): represents the average of the measurement values of the window.

$$\frac{1}{N}\sum_{n=1}^{N} X_n$$

2. **Median (median)**: represents the value that separates the higher half of the measurements values from the lower half.

3. **Mode (mode)**: represents the value that appears most often within all measurement values.

4. **Standard Deviation** (σ): represents the dispersion (i.e. variation) of the measurement values within the window.

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_n-\mu)^2}$$

5. **Summation (sum)**: where the measurements of acquired samples within a window (e.g. for 30 samples captured per second) are added sequentially.

$$\sum_{n=1}^{N} X_n$$

Table 5.1 Interpretation of the strength and direction of the relationship based on the calculated correlation coefficient (Lærd Statistics, 2017).

Strength of Association	Coefficient	
	Positive	Negative
Small	0.1 to 0.3	-0.1 to -0.3
Medium	0.3 to 0.5	-0.3 to -0.5
Large	0.5 to 1.0	-0.5 to -1.0

5.3.4 Bivariate Correlation Analysis

Correlation analysis is a measure of the association between two variables, in which the correlation coefficient ranges between [-1, 1], where the relationship between variables is weaker as the correlation coefficient value goes to 0. Furthermore, the sign of the correlation coefficient shows the direction of the relationship, whereas positive sign indicates directly proportional relationship between tested variables, while negative sign indicates inversely proportional relationship. Table 5.1 shows how the correlation coefficient can be interpreted to infer the strength of the the relationship between variables.

Correlation refers to the linear relationship between two variables. Linear correlation is used as an indicative of the relationship between tested variables. However, correlation does not imply causation. In other words, the strength does not mean always that the there is a relationship between tested variables (Mukaka, 2012). Therefore, the relationship between tested variables should be backed by the background knowledge and logical causal relationship.

Subsequently, the most common and widely used methods for correlation analysis are *Pearson's Product Moment* and *Spearman's Rank-Order* correlations. Accordingly, the selection of the appropriate correlation measure is according to the data type of the tested variables and other statistical properties such as normality, linearity and monotonicity.

Pearson's Product Moment Correlation

Pearson product moment coefficient measures the dependency strength and the direction of association that exists between two variables. Pearson in particular draws a line of the best fit through the data on the scatter plot of the two variables. The coefficient r shows how far these data from the line of the best fit, and can be calculated as follows for the two normally distributed variables x and y (Lee Rodgers and Alan Nice Wander, 1988):

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{(\sum(x-\bar{x}))^2(\sum(y-\bar{y})^2)}}$$

A number assumptions of the tested variables should be complied for Pearson, including that the variables are continuous, and have a linear relationship, which means that variables should have monotonic relationship, as well as homoscedasticity that scatter plot of the variables, that shows the data is normally distributed about the regression line. In addition that the two variable should be normally distributed. These restrictions are needed for Pearson to measure the association because it is very sensitive to outliers and false samples.

Spearman's Rank-Order Correlation

Spearman rank order correlation is a non-parametric measure shows the strength and direction of relationship between two variables, measures on at least an ordinal scale. Spearman correlation considers that the data type of tested variables either ordinal, ratio or interval. Furthermore, it is less sensitive to outliers than Pearson correlation. Moreover, it does not require normally distributed variables. Therefore, Spearman correlation had been used in many cases where the tested variables do not conform with the previously mentioned assumptions for Pearson. For x and y variables, the ranks are calculated between them, then using the following equation can be used to calculate Spearman correlation coefficient (Swinscow, 1976):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Consequently, analysing the relationship between two quantitative variables that are not normally distributed, needs to rank the values of each single variable separately starting from the smallest sample to be given rank 1. Then, after the ranks associated with each value of the tested variables are calculated, the squared difference d^2 between the ranks is computed and summed up to calculate the Spearman's correlation according to given equation for r_s.

Data Analysis Tool

There are several statistical analysis packages and tools. Within the work conducted, this PhD project used IBM SPSS Statistics Version-24, developed by IBM corporation. It is considered among the most common tools, because it is relatively easy to use in addition to its rich features such as comprising a lot of different statistical methods and functions, and flexible methods that facilitates data processing and generating different graphs and charts.

Moreover, it allows developers to generate and run scripts along with API interface to run SPSS scripts from external applications.

5.4 Analysis Approach

A thorough analysis was carried out of the correlations between subjective self-reports given by participants i.e. NASA-TLX scores and SAM scales, and features extracted from eye gaze data, which are descriptive statistical features for pupil size, i.e. *Mean, Media*n and *Mode*. The collected dataset described in Chapter 3 was used in the analysis. Additionally, several customised encoding schemes have been used to map values of pupil size into ordinal levels. Furthermore, self-reported scores have been normalised for each session. Subsequently, we have conducted different analysis experiments as will be given in the following subsections.

Spearman's Rank-Order Correlation has been used in the analysis to identify the dependency between extracted features and self-reported data, in order to observe and investigate the validity of the self-reporting techniques that were used during the sessions, when participants were interacting with typical UIs during tasks. Spearman's correlation has been used because the data is not normally distributed as well as it is less sensitive to outliers unlike Pearson's correlation.

Additionally, a rescaling has been applied to the self-reported scores, in particular Normalisation and Standardisation, in order to remove some of the potential bias and differences among different participants (Marquardt, 1980). Consequently, the following methods have been applied for the scores that had been used together with the original scores in the analysis:

1. **Normalisation**: which rescales the values of the variable into a range from [0,1], and it can be calculated using the following equation:

$$X_{normalised} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. **Standardisation**: which rescales the values of the variable into a range that has a zero *mean*, and *unit variance*, and can be calculated using the following equation exploiting only the *mean* (μ) and *standard deviation* (σ):

$$X_{standardised} = \frac{x - \mu}{\sigma}$$

5.5 Correlation Analysis between Pupil Size and NASA-TLX Scores

The following study examines the hypothesis that there is an association between pupil dilation tracked during a task and the NASA-TLX scores self-reported by the user. Therefore, a bivariate correlation analysis has been carried out between *Pupil Size* and *NASA-TLX* scores, where, in the following section, statistical features represent *Pupil Size* variable, and in the subsequent section percentages of *Pupil Size* levels using different mapping encoding schemes to group dilation levels.

5.5.1 Descriptive Statistics Features of Pupil Size versus NASA-TLX Scores

In this analysis, on one side a number of statistical features of the *Pupil Size* variable that were calculated from the recorded samples over the whole recorded session, which are:

- *Mean*: the average of all values of the captured pupil size throughout the whole session.
- *Median*: the middle value of the captured pupil size throughout the whole session.
- *Mode*: the most frequently occurred value of the captured pupil size throughout the whole session.

The other side of tested variables, are the self-reported scores for *NASA-TLX*, which are scales that range from 0 to 20 as given below (Hart, 1986):

- *Mental Demand*: refers to the amount of mental activity required to achieve the designated task.
- *Physical Demand*: refers to the amount of physical activity required for the task. However, physical demand in the context of HCI and interacting with user interfaces was a controversial issue as many participants found this question inapplicable as they did not exert a physical effort through interacting with requested computerised task. Physical demand reports are given even though many participants skipped it as it does not make sense from their point of view.
- *Temporal Demand*: refers to the amount of pressure that occurs because of the time given to complete the task.

Table 5.2 Bivariate Spearman's correlation coefficient between statistical measures of pupil size and NASA-TLX scores. Significant results at the 0.05 level are boldfaced.

		Right Eye			Left Eye		
		Mean	Median	Mode	Mean	Median	Mode
Mental Demand	Score	0.121	0.115	0.07	0.091	0.081	-0.009
	Normalised	0.069	0.067	0.038	0.03	0.011	-0.051
	Standardised	0.085	0.065	0.061	0.031	0.008	-0.045
Physical Demand	Score	0.093	0.097	0.138	0.087	0.094	0.136
	Normalised	0.13	0.133	0.143	0.125	0.127	0.14
	Standardised	0.025	0.004	0.047	0.011	0.009	0.059
Temporal Demand	Score	-0.026	-0.007	-0.026	0.016	0.028	-0.007
	Normalised	-0.069	-0.038	-0.012	-0.011	0.012	0.017
	Standardised	-0.13	-0.1	-0.041	-0.057	-0.031	-0.001
Performance	Score	-0.017	-0.009	-0.119	-0.024	-0.018	-0.063
	Normalised	-0.03	-0.022	-0.133	-0.029	-0.029	-0.062
	Standardised	-0.081	-0.076	**-.168***	-0.074	-0.072	-0.082
Effort	Score	**.191***	**.195***	0.141	**.175***	**.178***	0.114
	Normalised	**.158***	**.170***	0.126	0.127	0.126	0.081
	Standardised	**.173***	**.175***	**.156***	0.136	0.137	0.117
Frustration	Score	0.037	0.038	0.009	0.038	0.037	-0.014
	Normalised	-0.016	-0.009	-0.008	-0.012	-0.019	-0.022
	Standardised	-0.018	-0.018	0.005	-0.009	-0.018	-0.006

- *Performance*: refers to the amount of success that the participant believes that he/she had achieved. Subsequently, it reflects about participant's satisfaction with the accomplished goals of the requested task.
- *Effort*: refers to the amount of effort that had been exerted to complete the requested task.
- *Frustration*: refers to the amount of negative feelings that participants felt whilst working on the requested task, which shows how much annoyed, irritated and stressed that happened during the task.

Results

Subsequently, Table 5.2 shows the calculated correlation coefficients between each of the descriptive statistical features extracted from *Pupil Size* session samples with *NASA-TLX* self-reported scores during that session. Additionally, it shows *normalised* values together with *standardised* values of the self-reported scores.

It can be read from Table 5.2 that a statistically significant small correlation was found, however, the best results where observed for the *Performance* and *Effort* scores, which are 0.168 and 0.191 with *mode* and *mean* of the pupil size respectively. Moreover, *normalisation* and *standardisation* applied on the self-reported scores did not improve the correlation. Although *standardised-Performance* score makes better correlation, it appears to weaken the coefficient with *Effort* score.

5.5.2 Percentage of Pupil Dilation Levels versus NASA-TLX Scores

Alternative approach was conducted to investigate the relationship between *Pupil Size* and *NASA-TLX* scores that is based on percentages of pupil dilation levels throughout the session. Thus, each pupil size value of the acquired sample is mapped into a dilation degree, then the number of samples mapped into each level is used to calculate the percentage of samples for each level. Subsequently, the correlations between calculated percentages and self-reported scores were calculated as will be given in this subsection.

It is emphasised in the literature that pupil size is proportional with the cognitive load and mental processing. Nevertheless, there is no definitive way of separation to distinguish when the pupil is dilated and when it is not. Therefore, an empirical approach has been applied using several mapping scales to transform the pupil size value into one of three levels: *low, medium* and *high*.

The pupil size split was conducted using four different mapping schemes as depicted in Figure 5.2. Where starting from symmetrical width for all levels as in *Mapping1*, the ulterior mappings (*Mapping2, Mapping3* and *Mapping4*); the width of *medium* level is expanded gradually as can be seen in the Figure 5.2. Consequently, pupil size samples were normalised and transformed into a range from [0-100], then the value of the pupil size of each sampled frame was mapped into one of the three levels (i.e. *low, medium* and *high*). Following this, a percentage of the number of samples of each level was calculated, which corresponds to one of the level labels. Thus, the relationship between percentage of pupil size levels and the self-reporting scores provided by participants has been examined in this part.

Mapping1

The three pupil dilation levels (i.e. *low, med* and *high*) were split into a balanced range of equal length. Therefore, the normalised value of pupil size was transformed into one of the three equal intervals according the sample value using the equation of Mapping1.

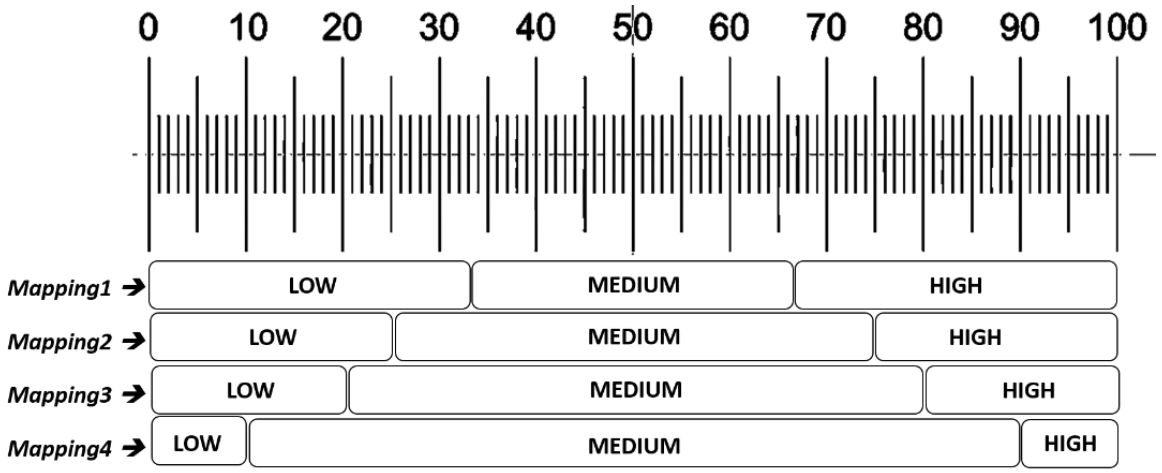


Fig. 5.2 Pupil size (normalised) mapping used within the analysis.

$$Mapping1 : x = \begin{cases} \text{low}, & \text{x} \leq 33.33\% \\ \text{med}, & 33.33\% \leq \text{x} \leq 66.66\% \\ \text{high}, & 66.66\% \leq \text{x} \end{cases}$$

Results given in Table 5.3 shows statistically significant correlation coefficients between percentages of pupil size dilation levels and NASA-TLX scores. Interestingly, moderate correlation coefficients achieved between pupil dilation levels and *Mental Demand* and *Effort* scores that are 0.336 and -0.302 respectively, which to a certain extent agrees with fact given in the literature that the pupil dilated with high cognitive load and mental processing. Additionally, for other scores small statistically significant correlations existed, apart from *Physical Demand* which was ignored by most participants, because there is no physical effort exerted in such Human-Computer Interaction.

Mapping2

The normalised value of pupil size was transformed into a dilation level, where the *medium* range expanded to be from [25-75] using Mapping2 equation.

$$Mapping2 : x = \begin{cases} \text{low}, & \text{x} \leq 25\% \\ \text{med}, & 25\% \leq \text{x} \leq 75\% \\ \text{high}, & 75\% \leq \text{x} \end{cases}$$

Table 5.4 shows the correlation coefficients between percentages of pupil dilation levels calculated using *Mapping2* scheme with NASA-TLX scores. A statistically significant correlation was observed that is marginally similar to the results achieved using *Mapping1*.

Table 5.3 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping1 scheme and NASA-TLX scores. Significant results at the 0.05 level are boldfaced.

Mapping1		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Mental Demand	Score	**.157***	**-.336****	**.203****	**.166***	**-.336****	**.164***
	Normalised	0.014	-0.148	0.1	-0.011	-0.119	0.08
	Standardised	-0.016	**-.154***	0.09	-0.029	-0.104	0.077
Physical Demand	Score	0.089	-0.054	0.027	0.114	-0.043	-0.007
	Normalised	-0.057	-0.011	0.061	0.02	0.032	-0.006
	Standardised	-0.073	0.086	-0.079	-0.033	0.14	-0.122
Temporal Demand	Score	**.180***	**-.169***	0.131	**.248****	**-.229****	0.108
	Normalised	0.116	0.017	0.027	0.116	0.033	0.02
	Standardised	0.074	0.07	0.008	0.112	0.071	0.009
Performance	Score	0.047	-.154*	0.135	0.07	**-.207****	0.146
	Normalised	-0.05	-0.055	0.078	-0.03	-0.088	0.091
	Standardised	-0.074	0.014	0.02	-0.049	-0.027	0.052
Effort	Score	**.223****	**-.276****	0.118	**.221****	**-.302****	0.074
	Normalised	0.139	-0.142	0.03	0.089	**-.161***	0.009
	Standardised	0.124	-0.135	0.031	0.086	-0.149	0.012
Frustration	Score	0.146	**-.197***	0.1	**.159***	**-.225****	0.062
	Normalised	0.004	0.035	-0.02	-0.016	0.031	-0.057
	Standardised	-0.027	0.068	-0.052	-0.043	0.072	-0.089

Table 5.4 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping2 scheme and NASA-TLX scores. Significant results at the 0.05 level are boldfaced.

Mapping2		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Mental Demand	Score	**.160***	**-.327****	**.255****	**0.079**	**-.325****	**.240****
	Normalised	0.037	-0.119	0.092	-0.035	-0.096	0.073
	Standardised	0.011	-0.126	0.09	-0.033	-0.095	0.076
Physical Demand	Score	0.121	-0.071	0.087	0.113	-0.086	0.098
	Normalised	-0.04	0.016	0.011	0.034	0.031	-0.011
	Standardised	-0.045	0.081	-0.094	0.071	0.092	-0.096
Temporal Demand	Score	**.180***	**-.194***	**.196***	**.190***	**-.254****	**.202****
	Normalised	0.071	0.02	0.036	0.097	-0.011	0.05
	Standardised	0.03	0.072	0.018	0.114	0.028	0.039
Performance	Score	0.029	**-.172***	**.186***	0.029	**-.222****	**.193***
	Normalised	-0.066	-0.059	0.087	-0.053	-0.089	0.094
	Standardised	-0.085	-0.004	0.037	-0.034	-0.059	0.064
Effort	Score	**.228****	**-.283****	**.194***	**.158***	**-.293****	**.167***
	Normalised	0.15	-0.143	0.08	0.066	-0.134	0.044
	Standardised	0.122	-0.129	0.075	0.071	-0.122	0.043
Frustration	Score	0.134	**-.218****	0.146	0.06	**-.196***	0.11
	Normalised	0.015	0.012	-0.049	-0.085	0.101	-0.109
	Standardised	-0.009	0.04	-0.081	-0.088	0.136	-0.144

Table 5.5 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping3 scheme and NASA-TLX scores. Significant results at the 0.05 level are boldfaced.

Mapping3		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Mental Demand	Score	0.107	**-.312****	**.267****	0.047	**-.301****	**.256****
	Normalised	0.001	-0.095	0.092	-0.057	-0.051	0.046
	Standardised	-0.023	-0.103	0.094	-0.063	-0.047	0.05
Physical Demand	Score	0.139	-0.069	0.083	0.121	-0.104	0.129
	Normalised	-0.031	0.043	-0.027	0.029	0.043	-0.02
	Standardised	-0.03	0.08	-0.09	0.045	0.096	-0.084
Temporal Demand	Score	**.162***	**-.218****	**.225****	**.189***	**-.257****	**.226****
	Normalised	0.074	-0.021	0.064	0.106	-0.019	0.052
	Standardised	0.038	0.029	0.043	0.125	0.006	0.048
Performance	Score	-0.003	**-.153***	**.178***	0.022	**-.244****	**.218****
	Normalised	-0.096	-0.028	0.066	-0.052	-0.113	0.106
	Standardised	-0.116	0.019	0.022	-0.041	-0.087	0.083
Effort	Score	**.224****	**-.287****	**.207****	0.151	**-.284****	**.199****
	Normalised	**.175***	**-.152***	0.086	0.055	-0.113	0.06
	Standardised	0.148	-0.137	0.076	0.049	-0.098	0.058
Frustration	Score	0.134	**-.218****	0.147	0.095	**-.202****	0.117
	Normalised	0.035	0.02	-0.065	-0.032	0.099	-0.134
	Standardised	0.019	0.045	-0.095	-0.032	0.128	**-.165***

Mapping3

Similarly, the normalised value of pupil size was transformed into a dilation level, where the *medium* range expanded more and more at the expense of *low* and *high* levels, using the relation:

$$Mapping3 : x = \begin{cases} \text{low}, & \text{x} \leq 20\% \\ \text{med}, & 20\% \leq \text{x} \leq 80\% \\ \text{high}, & 80\% \leq \text{x} \end{cases}$$

Results shown in Table 5.5 using *Mapping3* have approximately comparable correlation coefficient using *Mapping1* and *Mapping2*. However, the value of correlation coefficients slightly decreases as the ranges of dilation levels are changed. Thus, Table 5.5 has statistically significant correlations between percentages of pupil dilation levels that is calculated using *Mapping3* scheme with NASA-TLX scores.

Table 5.6 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping4 scheme and NASA-TLX scores. Significant results at the 0.05 level are boldfaced.

Mapping4		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Mental Demand	Score	0.09	**-.283****	**.274****	0.02	**-.272****	**.255****
	Normalised	-0.014	-0.052	0.059	-0.078	0.012	0.003
	Standardised	-0.041	-0.063	0.064	-0.06	0.012	0.012
Physical Demand	Score	0.074	-0.055	0.07	0.044	-0.127	**.156***
	Normalised	0.011	0.062	-0.057	-0.031	0.046	-0.034
	Standardised	0.014	0.037	-0.055	0.072	0.065	-0.066
Temporal Demand	Score	0.129	**-.231****	**.251****	0.135	**-.269****	**.221****
	Normalised	0.044	-0.031	0.066	0.056	-0.019	0.013
	Standardised	0.013	0	0.048	0.101	-0.015	0.023
Performance	Score	0.007	-0.15	**.161***	0.027	**-.266****	**.223****
	Normalised	-0.113	-0.019	0.042	-0.033	-0.131	0.117
	Standardised	-0.122	0.009	0.01	0.01	-0.122	0.1
Effort	Score	**.207****	**-.253****	**.216****	0.104	**-.266****	**.215****
	Normalised	**.161***	-0.107	0.078	-0.001	-0.085	0.079
	Standardised	0.134	-0.094	0.063	0.002	-0.065	0.064
Frustration	Score	0.123	**-.181***	0.149	0.049	**-.181***	0.122
	Normalised	0.058	0.054	-0.086	-0.06	0.135	**-.153***
	Standardised	0.052	0.064	-0.1	-0.038	0.149	**-.166***

Mapping4

The normalised value of pupil size was transformed into dilation levels that are very small range for *low* and *high* and wider range for *medium* level [10-90] as given in the following relation:

$$Mapping4 : x = \begin{cases} \text{low}, & \text{x} \leq 10\% \\ \text{med}, & 10\% \leq \text{x} \leq 90\% \\ \text{high}, & 90\% \leq \text{x} \end{cases}$$

Table 5.6 shows statistically significant correlation as well, however, it is obvious that *Mapping4* does not provide as good a correlation as observed from the other mapping schemes. In other words, *Mapping1* presented high correlation coefficients, and then by expanding medium level range on account of low and high levels (i.e. going from *Mapping1* to *Mapping4*), correlation coefficients between percentages of pupil dilation levels and NASA-TLX scores decrease.

Overall different mapping schemes, *Effort-Score* and *Mental-Demand* shows the highest correlation. Then, *Temporal-Demand* comes in the second place. Followed by *Performance* and *Frustration*, such that *Performance* was better in *Mapping2*, *Mapping3* and *Mapping4*, while *Frustration* achieved stronger correlation using *Mapping1* and degraded for the other mapping schemes.

Moreover, it can be observed that the correlation decreases as the transformation make the separation between the pupil dilation levels close to the edges, i.e *Mapping1* and *Mapping2* achieved better correlations than *Mapping3* and *Mapping4*, whereas *Mapping1* has stronger correlations than *Mapping2*. Consequently, this agrees with the fact presented in the Literature Review that the level of pupil dilation may entail information about the mental processing, effort and cognitive load.

5.6 Correlation Analysis between Pupil Size and Self-Assessment Manikin Scales

This study validates the relationship hypothesis between pupil dilation tracked during a task and the SAM scale scores self-reported by the user for *Valence*, *Arousal* and *Dominance*. Therefore, the linear relationship has been investigated between pupil size and self-reported SAM scales following the same approaches mentioned in Section 5.4.1.

5.6.1 Descriptive Statistical Features of Pupil Size versus SAM scales

Herein the bivariate correlation is examined between *Pupil Size* and self-reported SAM scale scores, with statistical features of *Pupil Size* on one side i.e. *Mean*, *Median* and *Mode*, and SAMs scores together with *normalised* and *standardised* on the other side.

Table 5.7 presents the calculated correlation coefficients between *Pupil Size* and self-reported *SAM scales*. From Table 5.7 it may be observed that *Arousal* scores have moderate positive statistically significant correlation with *Pupil Size* statistical features. Particularly, *Mode* feature of *Pupil Size* achieved the highest correlation with *Normalised-Arousal* scores. Additionally, *Dominance* scores have negative small significant correlation with *Pupil Size* features. On the contrary *Valence* scores did not show a relationship with *Pupil Size* features.

5.6.2 Percentage of Pupil Dilation Levels versus SAM Scale

Similar to the manner described previously with NASA-TLX, in this part the bivariate correlation has been investigated exhaustively between percentages of pupil dilation levels

Table 5.7 Bivariate Spearman's correlation coefficient between statistical measures of pupil size and SAM scale scores. Significant results at the 0.05 level are boldfaced.

		Right Eye			Left Eye		
		Mean	Median	Mode	Mean	Median	Mode
Valence	Score	0.012	0.008	0.021	0.033	0.023	0.042
	Normalised	0.026	0.018	-0.019	0.023	0.027	0.037
	Standardised	-0.064	-0.074	-0.088	-0.054	-0.054	-0.046
Arousal	Score	**.230****	**.256****	**.265****	**.228****	**.230****	**.255****
	Normalised	**.295****	**.320****	**.313****	**.299****	**.298****	**.325****
	Standardised	**.216****	**.240****	**.237****	**.233****	**.226****	**.231****
Dominance	Score	-0.115	-0.138	-0.094	-0.111	-0.121	-0.115
	Normalised	-0.143	**-.166***	**-.154***	**-.169***	**-.158***	**-.157***
	Standardised	**-.246****	**-.270****	**-.246****	**-.260****	**-.260****	**-.279****

Table 5.8 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping1 scheme and SAM scale scores. Significant results at the 0.05 level are boldfaced.

Mapping1		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Valence	Score	0.011	-0.099	0.045	0.005	-0.095	0.044
	Normalised	-0.018	-0.078	0.036	-0.037	-0.076	0.035
	Standardised	-0.037	-0.073	0.023	-0.087	-0.098	0.094
Arousal	Score	**.163***	**-.195***	**.168***	**.157***	-0.116	0.041
	Normalised	**.155***	-0.142	**.207****	**.164***	-0.074	0.017
	Standardised	**.152***	-0.121	**.179***	0.121	-0.064	0.032
Dominance	Score	-0.127	**.156***	**-.182***	-0.052	0.108	-0.129
	Normalised	**-.162***	**.237****	**-.193***	-0.074	**.205****	**-.161***
	Standardised	**-.157***	**.249****	**-.237****	-0.098	**.191***	-0.14

versus the self-reported SAM scales scores. Thus the same aforementioned empirical approach was used, which uses several mapping scales to transform the pupil size value into one of three levels: *low, medium* and *high*. The following subsection present results of percentages of *Pupil Size* levels using different mapping encoding schemes to make the split between dilation level groups.

Mapping1

In *Mapping1*, the *medium* level considers the value of pupil size that range from [33.33-66.66]. Table 5.8 shows the small statistically significant correlation for *Arousal* and *Dominance* scores.

Table 5.9 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping2 scheme and SAM scale scores. Significant results at the 0.05 level are boldfaced.

Mapping2		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Valence	Score	0.049	-0.124	0.095	0.071	-0.151	0.134
	Normalised	-0.028	-0.07	0.046	-0.068	-0.097	0.082
	Standardised	-0.062	-0.072	0.039	-0.099	-0.131	0.134
Arousal	Score	**.203****	**-.234****	**.254****	**.193***	**-.199****	**.165***
	Normalised	0.13	-0.139	**.220****	0.09	-0.064	0.084
	Standardised	0.113	-0.129	**.204****	0.094	-0.063	0.084
Dominance	Score	-0.024	0.115	**-.167***	0.054	0.089	-0.13
	Normalised	-0.086	**.247****	**-.268****	-0.023	**.243****	**-.274****
	Standardised	-0.105	**.253****	**-.290****	-0.024	**.226****	**-.257****

Mapping2

Similar to the analysis conducted for NASA-TLX, *Mapping2* considers *meduim* level for pupil size values that ranges from [25-75]. Accordingly, Table 5.9 shows slightly better results of the same significant correlation for both *Arousal* and *Dominance*.

Mapping3

The *medium* range is expanded more to include pupil size values ranges from [20-80]. The results produced given in Table 5.10 shows a marginal improvement with the correlation coefficients.

Mapping4

The *medium* range is the wider in this scheme, which comprises pupil size values ranges from [10-90]. Table 5.11 shows regression in the results, which yielded results that show lower correlation than the ones produces from *Mapping1,Mapping2* and *Mapping3*. Whereas the strongest correlation for *Arousal* dimension with the *high dilation level* using *Mapping3* which is 0.267.

Consequently, the experimental results presented are not evident enough to support the hypothesis strongly that the relationship between SAM scales scores with level of pupil dilation. However, *Arousal* and *Dominance* scores achieved stronger correlation than *Valence*, which might be due to the relevance between the mental processing and effort exerted on the task and pupil dilation reported through the *Arousal* scores in the SAM scale. Moreover, the

Table 5.10 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping3 scheme and SAM scale scores. Significant results at the 0.05 level are boldfaced.

Mapping3		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Valence	Score	0.061	-0.131	0.113	0.051	**-.177***	**.189***
	Normalised	-0.054	-0.053	0.042	-0.084	-0.117	0.11
	Standardised	-0.084	-0.062	0.042	-0.104	**-.156***	**.163***
Arousal	Score	**.220****	**-.257****	**.267****	**.196***	**-.209****	**.206****
	Normalised	**.152***	-0.148	**.192***	0.082	-0.055	0.094
	Standardised	0.131	-0.137	**.180***	0.089	-0.046	0.083
Dominance	Score	-0.002	0.103	-0.136	0.058	0.065	-0.102
	Normalised	-0.081	**.256****	**-.272****	0.003	**.251****	**-.310****
	Standardised	-0.112	**.255****	**-.275****	-0.006	**.236****	**-.292****

activation and response induced by attempting a task is theoretically related to pupil dilation. Therefore, the outcome of the given experiment may push towards the main objective of the research to exploit infra-red eye tracker to model the cognitive load of the user.

5.7 Relationship between Pupil Dilation Levels versus Tasks

Another endeavour of analysis that focuses on the association between computer tasks, i.e. different UIs, and pupil dilation levels, the study presented herein examines the hypothesis that there is a relationship between the percentages of pupil dilation levels during interaction and the UI type (i.e. the attempted task). Subsequently,Table 5.12 shows the percentages of pupil dilation levels for each of different task. Moreover, the four mapping schemes that have been previously mentioned in Section 5.4.2 were used for splitting between dilation levels.

One can read from Table 5.12 that the percentage of pupil size of high level of dilation within each mapping scheme was during Pacman task, except *Mapping4* that shows a deviated result. Hence, a gaming context such as Pacman stimulates and induces mental processing and cognitive load more than the other tasks.

Furthermore, an example of the pupil size variation within a game-based task context is shown Figure 5.3, whereby one can easily relate the engagement and the amount of mental processing during game play. Consequently, as an interesting finding that pupil size changes during the game-based task context, where the participant was playing Pacman is apparent during the situations encountered throughout the task.

As it can be seen in Figure 5.3, at the beginning of the game, the pupil size starts increasing while playing until a peak value occurs, then when the player is defeated in the

Table 5.11 Bivariate Spearman's correlation coefficient between percentages of pupil size dilation levels using Mapping4 scheme and SAM scale scores. Significant results at the 0.05 level are boldfaced.

Mapping4		Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Valence	Score	0.124	-0.136	0.127	0.061	**-.222****	**.234****
	Normalised	-0.012	-0.048	0.037	-0.082	-0.147	0.136
	Standardised	-0.045	-0.047	0.034	-0.087	**-.189***	**.184***
Arousal	Score	**.206****	**-.216****	**.237****	**.155***	**-.216****	**.229****
	Normalised	0.13	-0.104	0.136	0.076	-0.018	0.048
	Standardised	0.111	-0.089	0.126	0.074	0.001	0.028
Dominance	Score	0.038	0.016	-0.042	0.066	-0.019	-0.006
	Normalised	-0.073	**.200****	**-.232****	-0.002	**.242****	**-.292****
	Standardised	-0.124	**.190***	**-.213****	-0.011	**.215****	**-.253****

Table 5.12 Percentages of pupil dilation levels using the four splitting mapping schemes versus each individual task that acquired in recorded sessions in the Data Collection Study.

	Task	Right Eye			Left Eye		
		Low	Med	High	Low	Med	High
Mapping1	OS	10.19%	58.75%	31.06%	10.93%	58.55%	30.52%
	Online	11.96%	53.60%	34.44%	13.85%	52.59%	33.56%
	Excel	14.00%	51.61%	34.39%	17.21%	47.31%	35.48%
	Pacman	21.73%	39.65%	38.62%	20.75%	41.84%	37.41%
Mapping2	OS	5.05%	75.57%	19.38%	5.31%	75.95%	18.75%
	Online	5.06%	75.53%	19.41%	6.15%	74.57%	19.28%
	Excel	11.69%	68.39%	19.92%	12.10%	67.52%	20.38%
	Pacman	17.65%	56.23%	26.12%	14.06%	60.97%	24.97%
Mapping3	OS	4.06%	81.65%	14.29%	3.50%	82.63%	13.87%
	Online	3.39%	83.34%	13.27%	5.50%	80.82%	13.68%
	Excel	10.37%	75.07%	14.57%	10.87%	74.58%	14.55%
	Pacman	11.69%	69.57%	18.75%	10.16%	72.15%	17.69%
Mapping4	OS	3.23%	89.02%	7.76%	2.66%	89.86%	7.48%
	Online	2.63%	91.72%	5.65%	4.10%	89.64%	6.25%
	Excel	8.00%	83.82%	8.17%	8.94%	82.66%	8.40%
	Pacman	7.71%	85.50%	6.79%	6.39%	87.55%	6.07%

game, the pupil size starts decreasing. Consequently, this relates to the amount of cognitive load demanded by the playing context when the participant is engaged in the level of the game. Therefore, during game-based tasks, one can easily relate the engagement and the amount of cognitive mental processing required when playing the game.

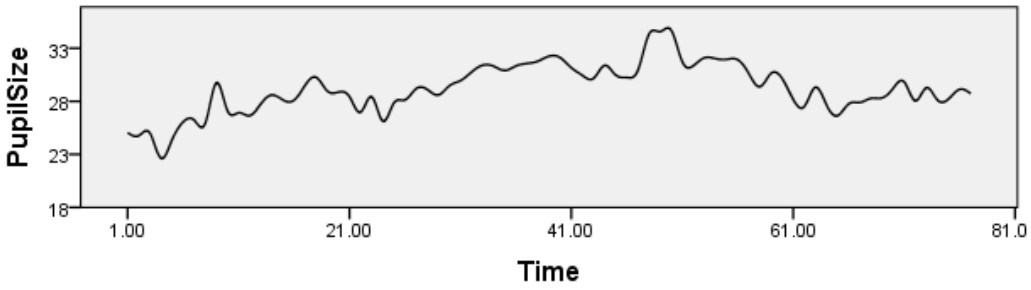


Fig. 5.3 Pupil size variation of the participant playing Pacman.

5.8 Eye Gaze Tracking Features for Task Classification

Finding a pattern for the eye gaze behaviour across different UIs possibly enhances the interaction quality and assists the achievement of an adaptive form of HCI. Therefore, a hypothesis that there is a consistent eye gaze behaviour against different tasks in HCI context, leads to the possibility of exploiting eye-tracking for task classification (UIs classification). In this section a number of classification experiments have been carried out using eye gaze features as input vectors, and the tasks that had been carried out by participants in the course of the Data Collection Study previously described in Chapter 3, as the classes (i.e. labels). Subsequently, descriptive statistical features extracted from eye gaze tracking data were employed for that purpose, which are *Sum, Mean* and *Standard Deviation (Std-Dev)*. However, these types of features work on a set of samples, thus it should be applied on a window that is composed of a number of samples. Moreover, because there is no ground truth for the proper window length; an empirical approach was adopted: eye gaze features were extracted using windows sizes of 1, 5, 10 and 20 seconds. For each window, 26 features were extracted as presented in Table 5.13.

Results

Extracted features were utilised to train a supervised learning model, namely an SVM. The classification experiment using cross validation of 10 folds, was applied for each individual participant separately. Table 5.14 presents the average, min, max, and standard deviation of

Table 5.13 Features that were extracted from eye gaze tracking data together the statistical functions that were used to aggregate the extracted feature from all windows of that session together.

Feature Name	Statistical Function	Number of features for both eyes
Fixations Number	Sum	2
Saccade Length	Sum	2
	Mean	2
	Std-Dev	2
Absolute Saccade Angle	Sum	2
	Mean	2
	Std-Dev	2
Relative Saccade Angle	Sum	2
	Mean	2
	Std-Dev	2
Pupil Size	Sum	2
	Mean	2
	Std-Dev	2

Table 5.14 Task classification that were extracted from eye gaze tracking data together the statistical functions that were used to aggregate the extracted feature from all windows of that session together.

Window Size (Sec)	Accuracy Average	Min Accuracy	Max Accuracy	Std-Dev
1	49.65% $\pm$ 0.55	28.68%	80.30%	10.76
5	50.57% $\pm$ 1.23	28.67%	84.29%	11.84
10	50.25% $\pm$ 1.73	28.57%	78.69%	11.64
20	48.60% $\pm$ 2.42	25%	74.19%	12.69

the classification results. The highest classification accuracy was achieved using a 5 second window, where the session with maximum accuracy obtained 84.29%, while the average of classification accuracy across all participants was 50.57%.

Moreover, the details of classification results have been reported herein for one individual participant to show *recall* and *precision* associated with each single task. This model achieves accuracy of 49.65% in which eye gaze tracking data was used to predict the task that has been undertaken. Subsequently, Table 5.15 shows the performance for the classification model using window size of 1 second length, whereby the rows show the actual task labels and the columns shown the predicted labels alongside the *recall* and *precision* per each label i.e. task.

Table 5.15 Confusion matrix using eye gaze tracking for task classification for one individual participant. The overall accuracy of this model is 49.65%

Confusion Matrix		Predicted				**Recall**
		OS	Online	Excel	Pacman	
Actual	OS	114	0	72	15	**0.567**
	Online	68	17	58	6	**0.114**
	Excel	42	7 1	80	6	**0.766**
	Pacman	66	2	23	49	**0.35**
Precision		**0.393**	**0.654**	**0.541**	**0.645**	

5.9 Summary

In this Chapter, different approaches and mapping schemes have been explored to employ extracted features from eye gaze tracking data, and to probe the relationship between pupil size and other eye gaze-based features, with self-reporting subjective tools, i.e. the NASA-TLX and SAM Scales, on different UIs for a range of computerised tasks.

NASA-TLX allowed participants to assess their performance whilst working on the given task and corresponding workload, and related efficiency aspects. Thus, NASA-TLX subscales have been studied and employed in this chapter, particularly the relationship between each individual dimension with the pupil size. Since it is widely reported that pupil size is strongly correlated with mental workload, i.e. the magnitude of pupil dilation is considered proportional to the mental effort required to process a task, which is known as the *Task-Evoked Pupillary Response* (TEPR) (Iqbal et al., 2004). Therefore, the study presented in this chapter went someway to validate this theory. However, despite the fact the results reported herein show small correlation coefficients, they show statistically significant relationships. Subsequently, this study can be used as indicative towards the TEPR.

On the other hand, SAM scales were used to let participants inform about their emotional experience whilst interacting with specified tasks. As shown in the results, the *Arousal-Dimension* of SAM scales achieved moderate statistically significant correlation coefficients with pupil size features. Providing that Arousal-Dimension scores refer to the amount of activation and agitation experienced during undertaking the task. This is relevant to a TEPR, which entails information about stress, effort, and performance from pupil size.

Additionally, several methods were deployed throughout the analysis such as normalising self-reporting as well as applying standardisation. Performing a rescaling on the self-reporting scores is also important in the analysis instead of sufficiency of using raw scores, in order to make scores per subject more consistent and aligned to normal distribution in addition to removing some of the potential bias among different participants. This perhaps posteriorly

enhances and improves the analysis, or may be negatively affecting the results. Consequently, the reported results showed that normalised and standardised scores did not improve the coefficients values in the correlation between pupil size features and NASA-TLX scores. In contrast, normalisation and standardisation produced marginally similar or even better correlation coefficients with SAM scales scores. Therefore, there is no conclusive rule for data adjustment and rescaling in statistical analysis, due to relative discrepancy among participants, besides the perplexity of human feelings that makes participants incompetent to assess different dimensions of the used self-reporting tools and questionnaire properly.

Furthermore, different mapping schemes were used for splitting and transforming pupil size values into pupil dilation levels. As such, an empirical approach is plausible since the divergence of characteristics between the tested UIs and the computerised tasks. Wherefore, trials of several splitting approaches with a view to use dilation levels instead of pupil size values, have been adopted to calculate the percentage of each dilation levels. The corresponding correlation coefficients between percentages of each dilation levels and the self-reporting scores were calculated and showed significant correlation as well. Percentages of pupil dilation levels that occurred during individual tasks were also computed and discussed using the four mapping schemes. Accordingly, pupil size that mapped to *high* level during gaming context was the uppermost percentage over other tasks. Providing that user's behaviour whilst playing games requires participants to be more engaged and needs larger amount of effort and mental processing. For this reason, game-based tasks are employed for measuring cognitive load and users engagement more than other contexts.

The studies and experimental results discussed in this Chapter investigated the hypotheses that are related to possible relationships between eye gaze tracking data and different scores of self-reports, which are given for the tasks and the affective and cognitive load levels, in addition to the hypothesis about the eye gaze behaviour across different UIs. Consequently, the validation of these hypotheses to be true, enriches the findings that help to answer the research question, *what is the best method that can be used as input channel to capture the user state that would be useful for adaptive HCI?*.

To conclude, there is a relationship between agitation and mental processing, and pupil dilation. Which becomes reasonably practicable through deploying techniques that track pupillary response behaviour using eye-trackers. Whereby endowing machines with the capability to perceive the user's cognitive load and the amount of mental effort, they become able to make the appropriate adaptation accordingly. Following chapter will elaborate in details the the benefits of evaluating the cognitive load and mental processing effort to facilitate the overarching objective to design and implement adaptive HCI as well as the

guideline of exploiting such techniques that promote the generation of a more intuitive and reliable relationship between users and computers.

Chapter 6

Facial-Based Features and Eye Gaze-Based Features

6.1 Overview

This following chapter further investigates the synchronised data (i.e. facial-based features and eye gaze-based features) from the two input modalities, which was acquired during the Data Collection Study previously described in Chapter 3. With a view that these input perception modalities have common properties. For instance they are deemed unobtrusive equipments, also they are slightly similar in terms of the nature that both of them are visual-based channels. In other words, they can operate and populate data about the user at a distance, and without the need for physical contact. Subsequently, there are different analysis approaches adopted within this PhD project using this data, that mainly involve classification, bivariate correlation and linear regression.

Firstly, the experiments and analysis given hereafter were performed to find out the relationship between the two visual-based perception input modalities via regression analysis. Therefore, an analysis approach was used that proposed to track the pupil size from facial-based features, which are features extracted from fiducial points of the boundary area of the eye, including eye and eyebrow. Subsequently, the work examines a hypothesis that there is a correlation among features extracted from different visual-based channels, particularly webcam and eye gaze tracking data. In short, a linear regression model was created, whereby making pupil size that was captured from the eye-tracker as dependent variable, and the independent variables were the facial-based features. Subsequently, this approach employs distance-based features that were extracted from facial point markers located around the eyebrow and the eye area.

Secondly, the other part of experiments that have used data from both channels, which are presented afterwards, examines the hypothesis that features populated from the visual-based channels, i.e. webcam and eye tracker, can be used for task classification. Therefore, experiments were performed to compare and investigate the patterns of the data populated from the two different input modalities, versus different UIs and designated computerised tasks, as previously described in Chapter 3. The analysis approach conducted for this employed supervised learning-based classification, whereby, given computerised tasks (i.e. different UIs) are the labels of the classifier, and the feature vector comprised from facial-based and eye gaze-based features together are the predictors.

A discussion has been provided in conclusion of each section, which attempts to show the benefits and the importance of these techniques.

6.2 Methodology

Facial-based and eye gaze-based features might be related as they are both captured via visual-based input modalities. Subsequently, two studies presented in this chapter involve exploiting visual-based input channels, namely facial-based features and eye gaze-based features, in an attempt to validate two hypotheses: the hypothesis that there is a correlation between facial-based and eye gaze-based features, and the hypothesis that tasks can be recognised from the pattern of features extracted from facial landmarks and eye-gaze behaviour features. The first study has been carried out through a linear regression analysis between pupil size dilation and distance-based features from facial landmark points captured from normal webcam. Linear regression has been used to explore the existing relationship in addition to using the generated linear regression model to predict future dependent variable values in the presence of independent variables, which means to approximate the pupil dilation from facial-based features rather than using a specialist infra-red eye-tracker. The second study has been conducted via a supervised classification approach that uses the tasks to be the target variable, while the predictor variables are either facial-based features, eye gaze-based features, or a combination between both. The classification experiments exposes the pattern of each set of features across the different tasks. Therefore, by endowing computers the ability to identify the pattern of facial-based and eye gaze-based features across different tasks and UIs through exploiting the visual-based input modalities, achieving the desired task-based adaptation of HCI may potentially become more possible.

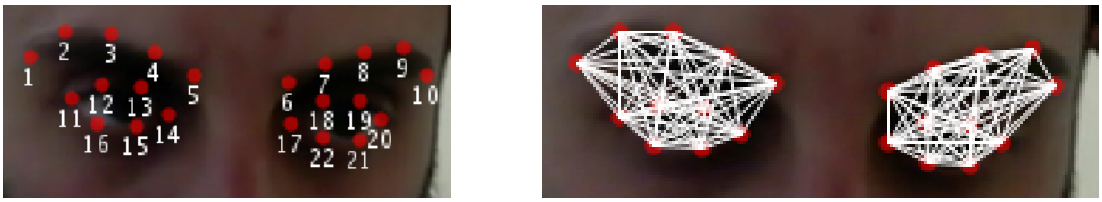
6.3 Tracking and Evaluation of Pupil Dilation via Facial Point Marker Analysis

Pupillary behaviour and dilation have been considered in the literature as an effective input for the measurement of cognitive workload and stress. However, using eye-trackers sometimes is disadvantageous. First, although new affordable trackers exist, still they are not cheap in comparison to webcams. Furthermore, occasionally eye tracking data becomes unreliable and invalid due to certain factors related to the user, such as cases when the pupil is too wide so it confuses the eye tracker, or may be occluded by the eyelids or big eyelashes (Schnipke and Todd, 2000). For these reasons, the work presented in these experiments is an attempt to find an alternative approach that can facilitate measuring (i.e. tracking) the pupil dilation using pervasive webcam technology, and employing distance-based features extracted from facial point markers of the eye bounding area. Subsequently, the hypothesis tested is that pupil dilation can be tracked by means of facial point markers captured via a normal webcam.

Generally, many of the currently available eye trackers do not give the absolute pupil size or diameter, rather they give a relative estimation of the pupil in arbitrary units. Moreover, because the variation of the pupil size is more informative than the absolute size, in the current work, the correlation between pupil dilation measurement using an infra-red eye-tracker and the features extracted by analysing frames captured using a typical webcam were investigated, specifically facial landmark points of the region containing the *eye* and the *eyebrow*. Therefore, in this work a new approach for pupil dilation evaluation using features extracted using normal webcams is proposed. Accordingly, the following sections will provide more details on the analysis process, comprising feature extraction, representation and linear model generation and results. Therefore, the following subsections present the methodology and the approach that has been undertaken together with the details of the experiment followed by a discussion.

6.3.1 Method and Experiments

The approach followed in these experiments involves the analysis of data from many participants who performed different activities using computer software. As presented in Chapter 3, the underlying experiment applied to typical human-computer interaction sessions carried out by different people, where each session involved the synchronised recording of two inputs: (1) webcam with resolution of 480x640 pixels; (2) eye-tracker (Eye-Tribe). The webcam recorded video at 30 frames/second, likewise the eye-tracker had a sample frequency of 30 frames/second.



(a) Facial point markers locating eyebrow and eye. (b) Distance-based features linking facial points.

Fig. 6.1 Facial-based features for eye area including eye and eyebrow.

For the purpose of investigating the correlation between the synchronised input data, each frame from the video recording is associated to the corresponding sample acquired by the eye-tracker.

Moreover, the eye-tracker generates the pupil size in an arbitrary unit, a relative metric that depends on the calibration and the distance between the tracker and the users' eyes. Consequently, there is no need for any pre-processing of the eye-tracking data, as the pupil sizes are automatically generated in real-time with each sample by the software driver that populates the data from the device. Besides, the video recording frames captured at a sample rate of 30 frames/second were associated synchronously with the eye gaze sample. If a corrupted video frame or eye gaze sample was observed, the corresponding time sample was excluded from the analysis.

A pre-processing step took place in order to extract features from the facial image frames of the recorded videos. Initially, the face region was located using the Viola-Jones algorithm for object detection (Viola and Jones, 2001). Following this, the *"Chehra"* facial landmark points detector was used to detect the location and shape of facial components, in particular the eyebrows and eyes (Asthana et al., 2014), whereby the *eyebrow* is depicted using five facial point markers and the *eye* using six facial point markers, as illustrated in Figure 6.1a.

Additionally, a distance-based feature representation was implemented by measuring the *Euclidean* distance among all generated points, as suggested in Chapter 4, resulting in 55 features produced from a combination of 11 points for each eye, as illustrated in Figure 6.1b.

Ultimately, each instance of the combined data represents one second of the session, composed of the eye-tracker measurements (primarily the pupil size), and was considered the *dependent variable* in the regression model, with the distance-based features extracted from the video recording image frame considered as *independent variables*.

Accordingly, as shown in Figure 6.2, the acquired recording of each session is pre-processed and populated to find the relation between the extracted features and the pupil size. Therefore, a regression model of the prepared data from the recording was generated for each session. Subsequently, the correlation between the actual values of the pupil size, and

the predicted values given using the generated linear model used as an objective function to calculate the fitness of the model. In other words, the two sets of data, which are distance-based features extracted from points captured by webcam and pupil size value extracted from the Infra-red eye-tracker. Furthermore, WEKA explorer was employed using 10-fold cross validation to measure the performance of the generated model (Witten et al., 2016). Moreover, for generating the linear regression model, the M5 attribute selection method is used, which is implemented within WEKA toolkit, which steps through the independent variables and eliminates the ones that are associated with small standardised coefficients until no improvement is observed in the estimate of the error based on the Akaike Information Criterion Measure.

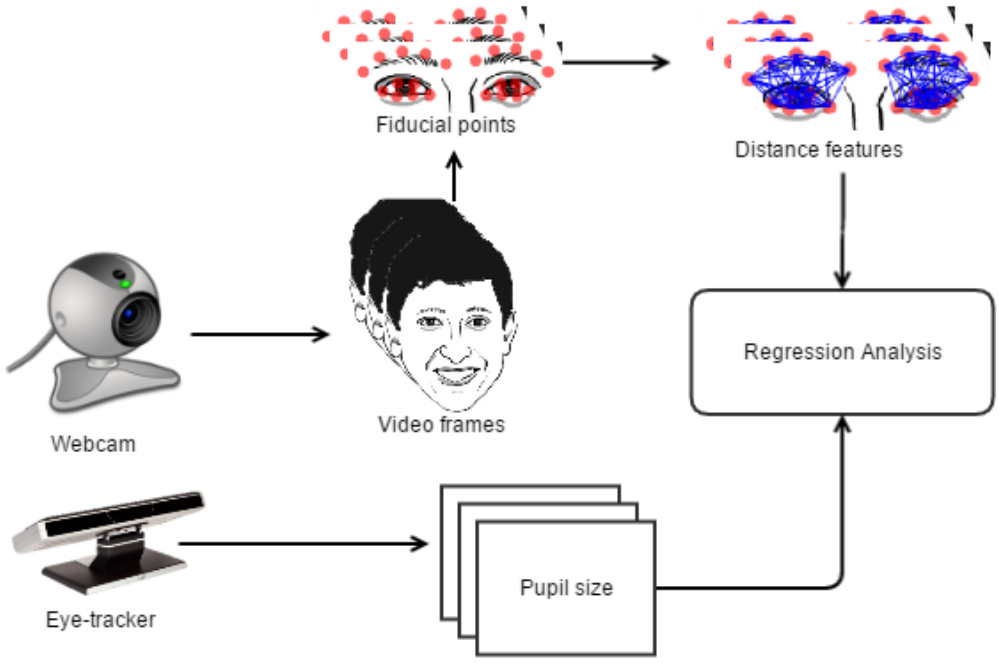


Fig. 6.2 Overview of the process carried out to extract features from the fiducial points of the eye and eyebrow until generation of the regression model; the data are prepared and combined to build the linear regression model, with the pupil size as the dependent variable, and the distances features extracted from the fiducial points as an independent variable.

A variety of measurements have been used within the experiments to present the goodness of fit, which are the following: *Pearson's Correlation Coefficient, Coefficient of Determination* (i.e. *Adjusted R-Square*), *Root Mean Squared Error* (*RMS-Error*), and the *P-value* is calculated to indicate statistical significance, where tests conducted herein were significant with $p < 0.001$.

Table 6.1 Summary of the recorded sessions with eye-gaze recording status.

Recorded Sessions	**Number of Sessions**	**Percentage with Respect to Overall**
Overall	168	100%
With Glasses	60	35.70%
Invalid	33	19.60%
Invalid with Glasses	20	11.90%
Invalid without Glasses	13	7.70%

The 42 participants participated were interacting with the predefined computer-based tasks including: (**T1**) a basic operating system task; (**T2**) an online shopping task; (**T3**) a spread-sheet manipulation task; (**T4**) video game-based task. Subsequently, from the Data Collection Study, 168 sessions were recorded from 42 participants, where each participant recorded four sessions, and each session corresponded to a single task.

6.3.2 Results

The correlation coefficient that is currently being investigated, was examined between the input channels for all participants, although recordings obtained from some sessions were considered invalid due to different reasons, as will be mentioned later. Table 6.1 summarises the information about the recorded sessions and the percentages of valid/invalid sessions, with corresponding information about whether or not the participants wore glasses during the sessions. Subsequently, 15 participants were wearing glasses during the experiments, who recorded 60 sessions, which represents 35.7% of the overall number of sessions. Although the eye-tracker can work over the glasses, it decreases the quality of the recorded eye-tracking data and the reliability. Furthermore, 33 sessions of data were considered as invalid as the number of invalid frames of the acquired eye-tracking data outnumbered the valid frames. Additionally, 20 of the invalid sessions were from participants who wear glasses, which represents 19.63% of the overall number of recorded sessions. Consequently, glasses influence the performance of infra-red eye-trackers, even though calibration for some participants was good enough to achieve strong correlation, as given in the regression model results. These causes and other environmental conditions affected the calibration process, which subsequently impacts on the overall system performance and the resulting accuracy level of the eye-tracker. Regarding the generated regression models, Table 6.2 presents a summary of the analysis from the linear regression models that were applied to each of the generated features from each single session, with the session models that achieved strong correlation are shown, i.e. $r > 0.5$.

Table 6.2 Linear regression models summary using 10-fold cross validation, with p-value <0.001.

Task	Correlation Coefficient	Adjusted R-Square	RMS-Error	Glasses
T3	0.858	0.737	0.953	No
T4	0.801	0.642	2.223	No
T4	0.786	0.618	1.963	No
T1	0.782	0.611	1.265	No
T2	0.739	0.546	1.396	No
T3	0.677	0.458	1.539	No
T3	0.671	0.45	1.601	No
T2	0.642	0.412	1.538	No
T3	0.64	0.41	1.414	Yes
T3	0.633	0.401	1.357	No
T4	0.623	0.388	2.526	No
T1	0.621	0.386	1.593	No
T2	0.609	0.37	1.516	No
T2	0.593	0.352	1.657	No
T4	0.587	0.344	3.261	No
T1	0.588	0.975	1.421	No
T2	0.588	1.201	1.802	Yes
T1	0.586	1.214	1.641	No
T4	0.582	1.772	2.368	No
T4	0.580	1.764	2.313	Yes
T4	0.578	1.605	2.330	No
T4	0.571	2.274	2.964	Yes
T1	0.568	0.958	1.275	Yes
T4	0.564	2.005	2.778	No
T4	0.546	2.683	3.358	No
T4	0.532	2.362	3.118	No
T1	0.528	1.129	1.531	No
T1	0.512	0.883	1.140	No
T3	0.510	1.611	2.068	No
T4	0.510	2.825	3.685	No
T1	0.508	1.574	1.962	No
T2	0.507	1.395	1.787	No

Table 6.3 Summary report of the regression model of the recorded session that achieved the highest correlation coefficient, which is in the first row of Table 6.2 . This model generated from a session where the participant conducting a spread-sheet task.

Regression statistics output	Value
Correlation coefficient	0.8584
R Square	0.736924313
Adjusted R Square	0.7368
Mean absolute error	0.7218
Root mean squared error	0.9535
Relative absolute error	44.9466%
Root relative squared error	51.2949%
Total Number of Instances	7825

From the results given in Table 6.2, it is concluded that pupil size correlates with distance-based features extracted from fiducial points of the eye area. Whereas data from 34 session recordings yielded strong correlations greater than 0.5, many other sessions produced moderate correlations. Additionally, all of them are statistically significant with p-value < 0.001. Accordingly, as it can be read from Table 6.2 that there is a statistically significant strong positive correlation between the distance-based features extracted from 11 points located around the eye area (i.e. including the eye and eyebrow), and the pupil size.

Moreover, the session data that achieved the highest correlation coefficient of the generated regression model is T3, as given in Table 6.3, which is a spread-sheet based task, with a correlation of 0.858, *Adjusted R-Square* of 0.737, and *RMS-Error* of 0.953. Furthermore, the homogeneity of variance is clearly visible, as shown in Figure 6.3, which assures that the variance around the linear regression line is approximately the same across different values of the independent variables that are the distance-based features.

Nevertheless, the discrepancy among the results obtained from different sessions is potentially due to a variety of reasons. The video frames were recorded using a typical in-built webcam, which captures images with a fair level of quality. Additionally, the illumination factor plays a very important role that potentially affected the quality of the recorded video and the eye-tracker data during some sessions. Furthermore, the eye-tracker recording during some sessions was invalid as the calibration setup had not been properly executed. During other sessions, the acquired eye-tracking data was intermittent due to the participant's head moving while interacting with the tasks, therefore the position of the eyes went beyond the range of the eye-tracker at times. It may be partially due to these reasons that low correlation coefficients were obtained from the data populated during these sessions.

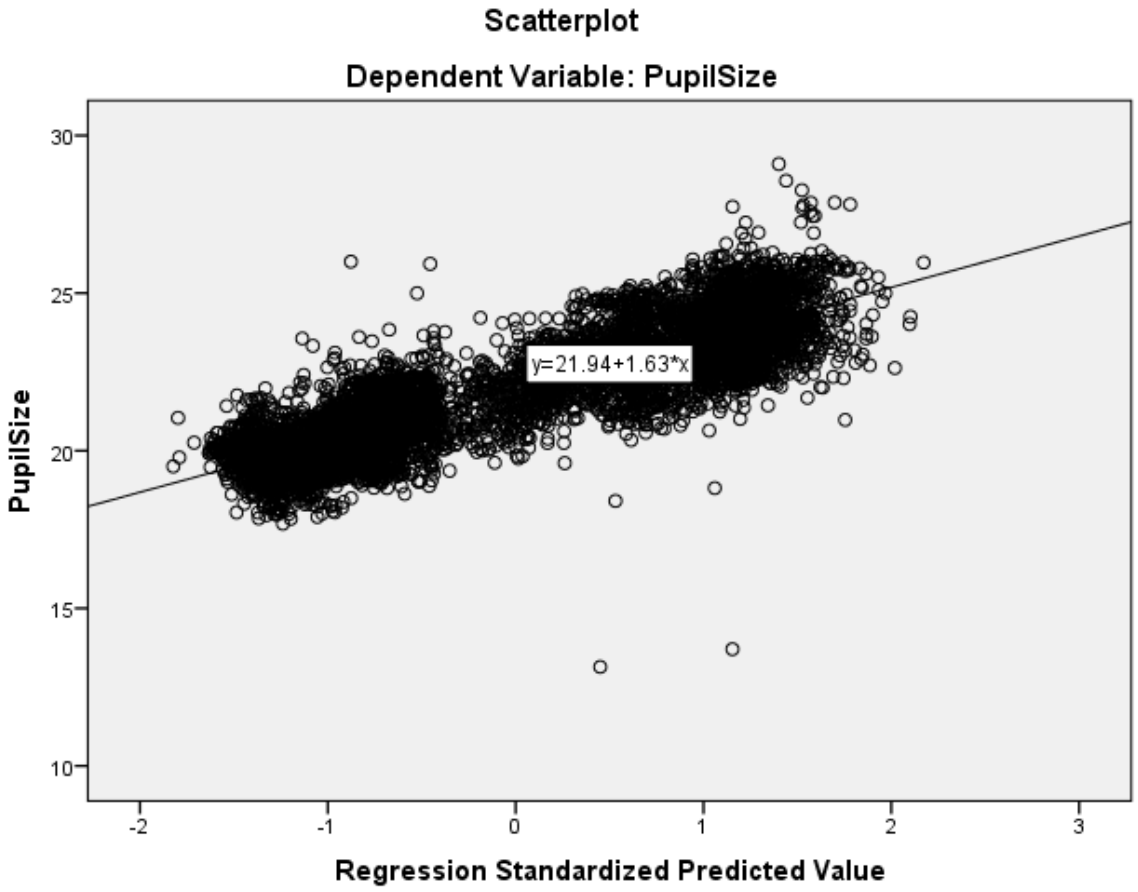


Fig. 6.3 Scatter plot shows the homoscedasticity of input features with approximately same variance around the regression line. This model generated from a session where the participant conducting a spread-sheet task, the generated linearly regression model achieved correlation coefficient of 0.858.

6.3.3 Discussion

Identifying and tracking user's cognitive state and mental workload by means of unobtrusive devices such as visual-based input modalities is a key for helping the user to achieve his/her goals and complete the task. It offers an opportunity and open new ways that facilitate automated intelligent for Adaptive style of Human-Computer Interaction. Regardless the cause of the cognitive load causes, it is important that to maintain the amount of cognitive load within the total cognitive capacity, in order to avoid cognitive ageing that resulted from working-memory capacity reduction due to irrelevant information, which consequently slows down mental processes, and thus affects the overall performance (Van Merriënboer and Sweller, 2005).

Therefore, tracking pupil dilation as an indicator of cognitive load is very important in this context. However, despite the fact that eye tracking is successful to be used to assess and track user's cognitive load and mental processing; until these day it is not considered pervasive and commonly used such as webcams. Subsequently, the proposed way of validation in this Chapter is a novel approach that evaluates and discloses the efficiency of using webcam captured frames, and correlates facial-based features with the measurements populated from synchronised eye-tracker, in a typical usage of user interacting with prevalent UIs.

The overarching aim of these experiments attempted to develop an alternative, cost effective technique for the representation of pupil dilation in order to track pupillary behaviour from images instead of employing specialised, high-cost eye-tracking devices, which typically require specialist expertise during setup and calibration. Consequently, these experiments given here examined the validity of the the hypothesis that eye gaze data con be used to model user's cognitive states, and respond to the second research question that eye gaze tracking data is a suitable channel for perceiving and modelling users states in HCI contexts. The approach was validated using the aforementioned set of video recordings acquired from real human-computer interaction sessions, which focused on a set of predefined, typical computer tasks. The facial and eye-gaze behaviours were recorded using a standard built-in webcam and an infra-red eye tracker respectively.

As a result, a strong correlation between the pupil size and distance-based facial features was observed for some of the sessions acquired from the Data Collection Study. Consequently, this suggests that one might be able to track the pupil behaviour using widely available webcams that come with almost all computers and smart phones. It is anticipated that there is great potential in the use of common webcams in determining pupil dilation using established computer vision techniques.

6.4 Task Classification via Visual-Based Input Modalities Combination

As previously mentioned, there is a need for another approach that empowers computers to perceive more about the task that is being conducted by users, so it can facilitate further adaptation that fosters task completion, resulting in a more reliable and efficient form of HCI. Subsequently, the aim of the work presented herein is to explore the relationship between the category of software application in use, and the physiological measurements, in particular visual-based channels including facial expressions and eye gaze metrics. This investigation examines the hypothesis that visual-based features can be used to classify different types of UIs, i.e. different HCI tasks. Therefore, this section describes the experiments that have been carried out, that explore user-driven task-based classification, whereby the classification model is capable of predicting the type of the interactive task that the user is currently undertaking. The classification algorithm used a feature vector that is composite from visual-based input modalities, i.e. facial expression via webcam, and eye-gaze via eye-tracker.

6.4.1 Method and Experiments

A number of supervised classification experiments have been carried out. Where each experiment manipulates data from two different input channels captured from the 42 participants who took part in the Data Collection Study, previously described in Chapter 3. Firstly, facial-based features have been extracted from fiducial landmarks generated from the users' facial expressions, as previously described in Chapter 4. Secondly, eye-gaze based measurements have been generated from the users' eye tracking data, which is described in Chapter 5. Thirdly, feature vectors have also been generated using a combination of the data from the two previously mentioned input channels, which will be given in this section along with details of the method of feature representation that have been used as well as the results obtained. The quantitative analysis adopted for the present experiments uses classification accuracy, obtained from a supervised machine learning approach as a measure of the relationship between features extracted from visual-based inputs and the computer task, i.e. UI, that is being performed.

As depicted in Figure 6.4, the combination-based feature vector resulted from concatenating the facial-based vector, composed of 1176 features as described in Section 4.3, with eye-based vector that comprises 26 features as described earlier in Section 5.8. Subsequently, the resultant vector represents the fused combination-based vector containing 1202 features.

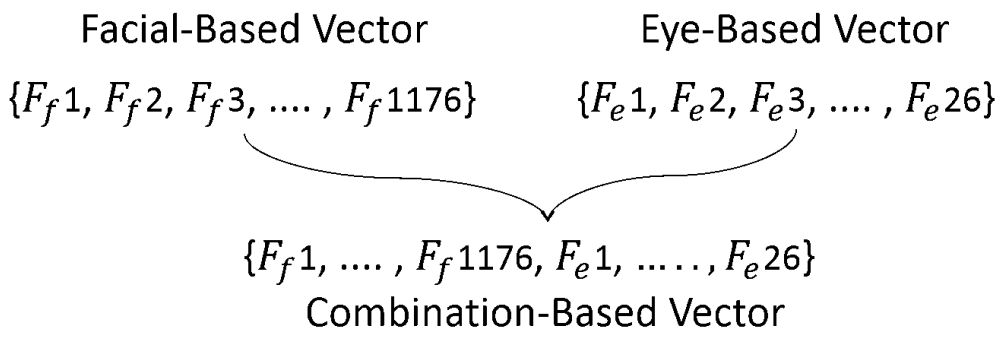


Fig. 6.4 The combination-based vector that composed of facial-based and eye-based features.

Table 6.4 Classification accuracy average across all subjects using the three types of features: Facial-based, Eye gaze-based, and Combination-based.

Feature Vector Type	**Classification Accuracy**	
	Average	Standard Deviation
Facial-Based	85.52% $\pm$ 0.38	6.57
Eye Gaze-Based	49.65% $\pm$ 0.55	10.76
Combination-Based	87.63% $\pm$ 0.38	6.61

For both input channels, after the features have been extracted and represented, as previously described, they were used as predictors, and the corresponding task label used as a target class for the supervised machine learning classifier. Accordingly, the extracted features were classified using a prediction model trained using 10-fold cross validation. The number of instances varies across different participants based on the time needed for that subject to complete the task or until the recording stopped because the time limit. The average number of instances is 671.78 across all 42 subjects. Accordingly, the average number of instances in the testing subset is 67 instances for each fold, where the model trained using the rest of 604 instances on average. In the investigations, the SVM variation C-Support Vector with linear kernel, as discussed in Chapter 4 and Chapter 5, was employed.

6.4.2 Results

As mentioned, evaluation of the classifier used during the investigations carried out, employed classification accuracy as the primary metric. Table 6.4 reports the average of the classification accuracy of all the 42 participants data using the facial-based features given in Chapter 4, the eye gaze-based features given in Chapter 5, and a combination of both features. As shown, the combination of facial and eye gaze-based features achieved the highest classification accuracy of 87.63%.

Moreover, Figure 6.5 shows the classification accuracy for each individual participant using the different predictor vectors. As may be observed in Figure 6.5, the feature vector using facial-based data produced a higher classification accuracy across all participants with an overall average classification accuracy of 85.52%, than the feature vector comprised of eye-gaze based features, which produced an average classification accuracy of 49.65% as given in Table 6.4. However, the highest average classification accuracy, 87.63%, was obtained using a combination of both facial and eye gaze-based features within the feature vector. Although, it is obvious that combination of both features marginally produces an even greater classification accuracy, facial-based yielded similar and slightly better accuracy for a few participants.

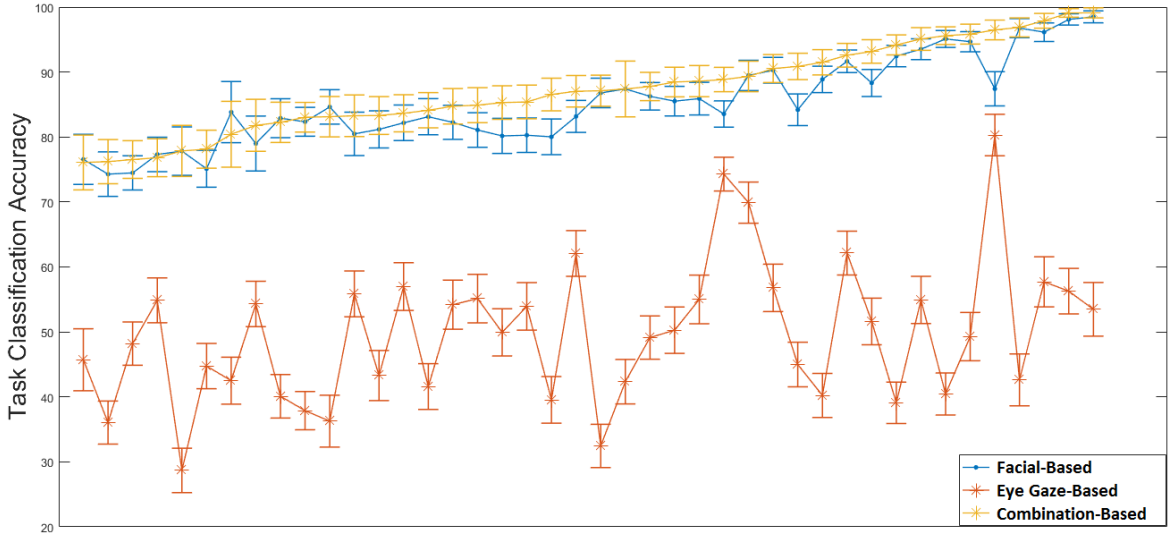


Fig. 6.5 Task classification accuracy of the 42 subjects, using facial-based features, eye gaze-based features, and a combination of both. Lower/upper bounds of the confidence interval are calculated using a 95% confidence level. The results sorted in ascending order of the combination feature accuracy.

6.4.3 Discussion

Within the investigations presented, classification accuracy was used as an evaluation metric. The results obtained showed that facial-based features achieve higher classification accuracy than the eye-gaze based features for most of the involved participants. Subsequently, the obtained results indicate that visual-based input channels may provide information about the UI and the nature of the computerised task being attempted.

Correspondingly, using a SVM classifier, the average classification accuracy achieved across the 42 subjects is 85.52% when utilising facial-based features as an input feature vector,

and an average accuracy of 49.65% when using eye-gaze based features as an input feature vector. Furthermore, using a combination of both types of features achieved an average classification accuracy of 87.63%. Consequently, although the combination of both features marginally produces better accuracy, facial-based features permit a task classification model with a higher degree of accuracy than that obtained when using eye-gaze based features, for the purpose of task classification of dissimilar UI-based software applications.

Accordingly, this would suggest that facial-based features potentially provide a more systematic pattern than eye gaze-based features when different user interfaces are used, which may make it a suitable input vector for categorising computer tasks based on user interaction. Moreover, the results show that combining facial with eye-based features improves the task classification process, whereas the classification accuracy of the combined-based features achieved the highest overall classification accuracy.

On the other hand, eye gaze data is commonly used to identify usability and visualisation issues such as *Area of Interest*, as well as examining perceptual and cognitive abilities such as visual and verbal working memory, which are more closely related to the complexity of the tasks and the level of difficulty of the same task kind. Despite that, eye gaze-based features in classification investigations conducted across different UIs of various designated tasks, did not show a pattern that makes it possible to classify the task (i.e. UI) using eye gaze-based data. Consequently, the hypothesis of task classification through exploiting features populated from visual-based input channels is true, although eye-based data alone does not appear to be useful for task classification, it slightly enhanced the classification accuracy when combined with facial-based data.

Additionally, it is important to recall that although facial-based features showed good classification results across different UIs (i.e. tasks), indeed this does not necessarily mean that there is a relationship between a particular facial expression with specific UI or task. The difference is that the interchange between user's facial expressions is noticeable whilst the user is interacting with a specific UI, attempting a task, more than other tasks, as mentioned in Chapter 4. For this reason there is a pattern for facial-based features across different UIs. In this perspective, one can reconcile the results achieved from different studies carried out and presented in the thesis.

6.5 Summary

The aim of this chapter is to investigate the relationship between the various data types, which were populated from different channels, primarily visual-based input perception modalities such as facial-based data that represents video recordings of user's face, and eye gaze-based

data, which represents the behaviour of user's eye that was captured and recorded using specialised eye-tracker device.

In the experiments presented in Section 6.2, the correlation was explored between pupil dilation, which is tracked and recorded using infra-red eye-tracker, and features extracted from low quality video frames that have been captured using a normal webcam during a set of computer-based tasks, populated from the Data Collection Study previously described in Chapter 3. Subsequently, the hypothesis that there is a correlation between features of different visual-based channels has been validated. Furthermore, Section 6.3 talked about the approach and results of the conducted investigations, for the purpose of computer-based task classification, through exploiting acquired facial and eye gaze data. Facial-based, eye gaze-based, along with a combination of these features were used as predictors, with the task categories used as the target classes, for a SVM classifier that attempted to model the participants task interaction. Therefore, this assures the validity of the hypothesis that facial-based features would be used for task classification, which subsequently contributes towards the achievement of intelligent HCI systems.

It is anticipated that expanding the work in these investigations, in order to exploit the same perception input channels for the purpose of modelling the user's affective and cognitive states, along with the prediction of the task being conducted, thereby embodying real-time information about the task conjointly with the user's affective and cognitive states. Such improvements would potentially act as enablers for the construction of intelligent computer interfaces within HCI contexts that are analogous to human intelligence within natural Human-Human interaction contexts. As a result, a more intuitive form of interaction may be engendered that promotes effective and efficient task completion, in contrast to the case of conventional non-adaptive HCI. Subsequently, complementing the user's affective and cognitive state with information about the task context will reinforce the intelligence of the machines, so a further step will be achieved towards adaptive and intelligent HCI.

Chapter 7

Conclusion and Future Work

7.1 Overview

This work is motivated by the goal to model user states whilst interacting with a UI in general. The investigations aimed to pave the road-map of the elementary components that facilitate adaptive HCI. In particular, endowing computers the capabilities to perceive user's *affective* and *cognitive* states through analysing non-verbal visual cues, which comprise facial expression and eye gaze behaviour. Therefore, different approaches and techniques have been proposed, tested and validated, in terms of the exploited input channels, methods of feature extraction and representation, as well as machine learning techniques for classification and statistical methods for making other inferences from the data. Furthermore, a thorough exploration has been conducted to investigate user emotional states in relation to the self-reported scores within a context of a human interacting with most common used UIs. Subsequently, the current chapter will recapitulate how the work parts previously presented addressed the research questions together with the outcomes and contributions achieved.

7.2 User Model and Adaptive Interaction

The idea of enhancing the interaction style between humans and computers has evolved in many ways and in different directions. With the appearance of Affective Computing, there has been research more focused toward enhancing HCI by imitating human-human interaction, which is characterised to be empathetic whilst supporting effective communication via conversation space (Harper et al., 2008; Picard, 2000b). In addition, this work becomes focused toward user engagement and production of a more attractive user experience out of

intelligent computer software that is to interpret actual affective and cognitive states (Karray et al., 2008; O'Brien and Toms, 2008).

Two case scenarios are given in Figure 7.1 and Figure 7.2. These demonstrate two different cases of affect-aware systems, where the systems infer and utilise the user's feelings, and attempt to foster the completion of the desired task that they are trying to accomplish. The two systems are completely different in the services they provide, and the functionality they afford and serve. However, the pivotal component that should exist in any intelligent system should perceive non-verbal cues and understand affective states of the user within that contextual scenario, which provides a key aspect of adaptive HCI or indeed adaptive UX.

Joe Bloggs is a 56 years old business man, who works for ACME LTD as sales executive. He decided to travel for leisure to Hawaii. His friend John helped him and prepared the booking for the hotel and the flight ticket. Joe arrived at the airport before 1 hour to his flight departure, and he wanted to check-in for his luggage, he loaded the luggage inside the cart, and put his ticket under the reader that uses infrared technology, the screen of this machine showed that the check-in process cannot be completed and beneath this message a line with red font "*click here for details*". However, Joe has no experience in dealing with such interfaces; he misunderstood what was going on. He unloaded the cart and put them back again, and he could not proceed. Joe started getting frustrated and angry. Fortunately, this ticket reader machine was supported with eye tracker and camera technology and the machine predicted Joe's feelings of perplexity and pinged the airport office about a problem, and announced: "*We are sorry for disturbing you, Please wait a moment. Our help desk advisor on her way to you".*

Fig. 7.1 Use case scenario of an Intelligent check-in machine.

Scenario presented in Figure 7.1 involves an old man who is digitally illiterate and has no experience in using computer interfaces. He needs to interact with a computer interface in order to self check-in and he could not avoid this check-in process. Also, he was confused and no staff was nearby to help. Therefore, an affect-aware machine would be very helpful in this scenario. Even though this system does not update or apply any kind of adaptation in the interface itself; it still recognises the difficulty that the user had faced so it prompted the help desk to assist the user in completing the task.

George Burdell is a student who officially was enrolled in a suit of two e-learning training modules in management skills and leadership integrity at Georgia Tech. The modules are interactive wizards, which give a piece of information about a single concept, then make simple review using multiple choice and drag-and-drop questions. George began the reading part, and he intended to proceed to the next concept, but he got the review part on the screen. In this moment, George was not aware what he should do. Fortunately, an eye tracker was fixed above the monitor; and it detected his pupil gaze movement in different directions turning upside-down, in a way that let the computer to predict that George was confused. Then the e-learning systems started showing tips and hints about this part, and it helped him in answering the question by in order to complete this part of the module to proceed.

Fig. 7.2 Use case scenario of an Adaptive e-learning system.

On the other hand, the scenario presented in Figure 7.2 involves a slightly different kind of an intelligent computer software that can predict user confusion with the input of an unobtrusive eye tracker fixed above the monitor. This sort of system is very helpful for users and increases the chance to enable them to complete their task efficiently and effectively. This is important since many users give up a task due to getting stuck in a counter-intuitive stage of the interaction.

Consequently, the main aspects of adaptive HCI and the User-Model role will be addressed and discussed. The subsequent section herein will discuss the framework design of a abstract adaptive HCI. Details of the main entities in the framework are also discussed. Further details are discussed regarding the exploitation of affective state and cognitive state perception for the purpose of designing an adaptive HCI. Components of this PhD project aim to provide a roadmap of the fundamental elements of an Adaptive HCI. Subsequently, the perspective regarding the operational pipeline of an Adaptive HCI can be illustrated in the flow chart depicted in Figure 7.3, where computer software tracks and infers the user's affective and cognitive states whilst they interact with the UI and progress through the task that they are attempting to complete.

Vision-based input perception channels including a video feed and eye-tracking are presumed to capture and populate features that can be utilised to generate a User-Model. These perception channels represent the crucial element of the overall process that will cause the adaptation to take place.

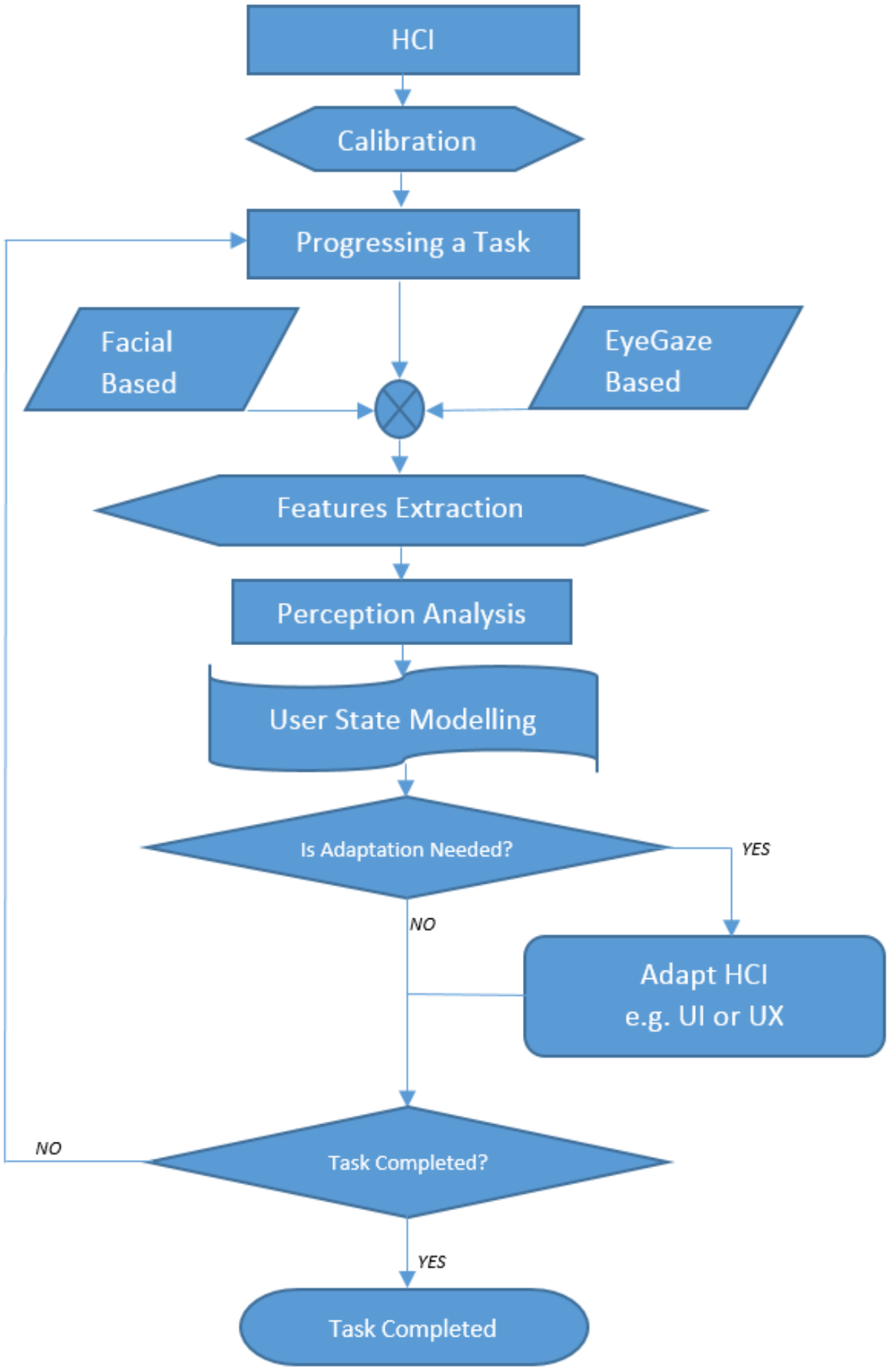


Fig. 7.3 Flowchart depicts the operation of an Adaptive HCI, which the system adapts the UI or the UX according to the generated User-Model in order to maximise the performance of the system and 'guarantee' task completion.

The generation of a User-Model will provide the computer with the ability to analyse, perceive and understand humans and adapt the HCI accordingly in real-time (Sullivan and Tyler, 1991). Thus, adaptive HCI enable systems to be self-aware of the relevant aspects that influence or prohibit a user in attempting and completing a computer task (Saffer, 2009). Therefore, the User-Model in the context of adaptive HCI provides a representation of the inferred affective and cognitive states of the user at a particular moment of time. This represents knowledge about the user that the system should use for the personalised benefit of that user. In accordance with modelled user states and the other application constraints; the system should be able to make the required adaptation either to the overall UX or to the UI only, to help the user to achieve the desired goals and complete the task.

Moreover, the whole process of an intelligent adaptive HCI should operate in real-time and with a high frequency of user tracking until the task finishes and is completed (as the flow chart illustrates in Figure 7.3). This is important since the affective-cognitive status of the user is not fixed and changes over different stages throughout the designated task yielding manifold user responses. In this way, the interaction between the human and the computer becomes more reliable and efficient fostering task completion and maximising the performance of the user providing more intuitive and convenient HCI. This would largely be due to the machine being more intelligent in understanding user states enabling relevant automated changes to the HCI until the task accomplished.

7.2.1 Framework for Adaptive Human-Computer Interaction

The primary challenge in designing Adaptive HCI is to provide computers with the ability to analyse and understand the user's affective-cognitive states from different interaction modalities in real-time, which requires a flexible, customisable and accurate User-Model that enables a more pleasurable experience through an adaptive and intelligent system that guarantees task completion, support and better error handling when compared to conventional non-adaptive systems (Karray et al., 2008; Oviatt, 2003; Sebe, 2009).

A conceptual framework for Adaptive HCI has been developed and presented in this chapter, which describes an abstraction of a system that continuously attempts to interpret and infer the user's affective-cognitive states whilst they are interacting with the UI. The idea is to generate a User-Model that leads to an appropriate adaptation style. Consequently, the generated User-Model should represent the actual states of user that are regularly inferred in real-time from various input modalities whilst he/she is interacting with the UI. As depicted in Figure 7.4, the Adaptive HCI framework is basically composed of two components:

1. **Perception Component**. The Perception Component is responsible for handling the input modalities to model user states. Subsequently, the role of the Perception Component is to model the affective and cognitive aspects of a user whilst interacting with a user interface, using features acquired from the input perception modalities.

   (a) *Affective perception* means recognising a user's feelings and emotions (Picard, 2003). Therefore, considering user's emotions may produce a novel and intuitive interaction, which potentially improve the user's performance and fosters effective task completion.

   (b) *Cognitive perception* will be used to assess the user's mental and cognitive workload (Iqbal et al., 2004). Consequently, evaluating the cognitive aspects of the user while interacting with an interface is significantly important for sustaining user engagement, resulting in a better HCI quality (Burleson and Picard, 2004; Chen and Epps, 2014). Ultimately, the output of the *Perception Component* is the User-Model, which subsequently will prompt the *Adaptation Component* to provide the suitable changes according to inferred user's states .

2. **Adaptation Component**. The Adaptation Component uses different approaches to adapt the UI or the UX based on the application. Subsequently, the role of the Adaptation Component is to adapt the UI or the UX according to the generated User-Model and predicted state of the user in a way that supports the user and facilitates optimal interaction (Picard, 2000b). Different approaches may be carried out to achieve the adaptation. Therefore, there may be a wide spectrum of adaptation techniques utilised in order to enable a suitable modification of the UX (Duric et al., 2002), such as:

   (a) Giving 'help' adjuncts.

   (b) Changing tasks organisation.

   (c) Modifying the interface.

While the proposed framework aspires to the eventual creation of adaptive HCI, the overarching focus of the research presented within this thesis has been on the *Perception Component* since applying the proper adaptation methods is controlled by the application constraints. Accordingly, it is anticipated that by integrating both the *Perception Component* and the *Adaptation Component* to work together would potentially achieve the fundamental objectives of an intelligent and adaptive kind of HCI. Nonetheless, the *Perception Component* could be considered a standalone portable component that can be used and integrated within

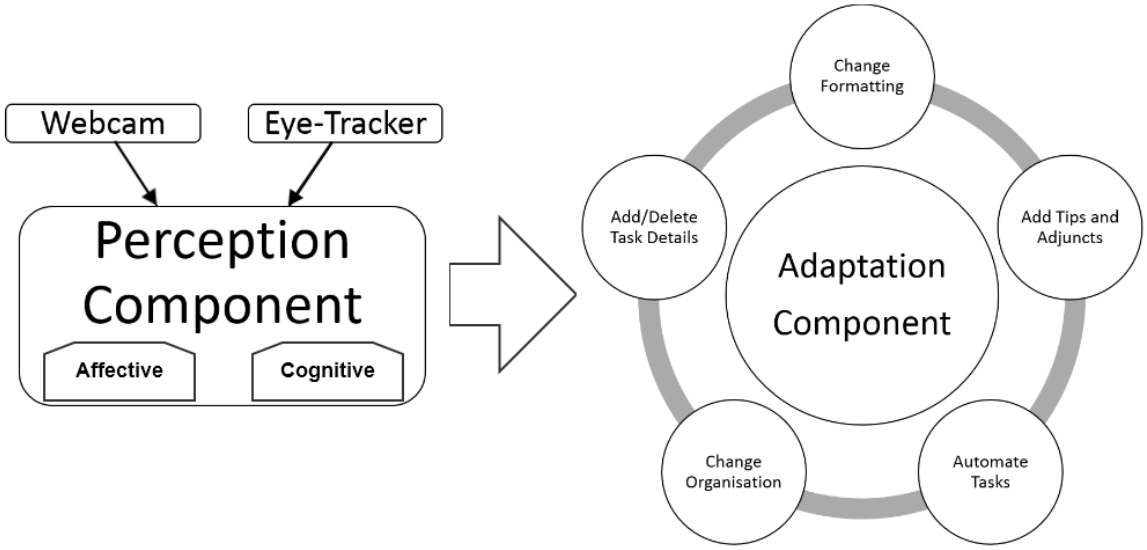


Fig. 7.4 Framework for Adaptive Human-Computer Interaction. a) The Perception Component is responsible for handling the input modalities to model user states and generate the User-Model. b) The Adaptation Component uses different approaches to adapt the user interface based on the application according to the generated User-Model.

different HCI contexts. Thus, the aforementioned figure represents a generic framework of adaptive HCI.

Affective State Perception

With regard to the adaptive HCI context, detecting the *Affective State* and information about the confusion of a user, or any emotional state that is directly related to the interactive task being undertaken, which predicts an adverse impact on task completion, is considered the most relevant information for the system in order to apply the appropriate adaptation before the user abandons the task. Accordingly, a novel terminology suggested from the current PhD project that is adaptive HCI should be able to detect the *User's Perplexity State*. Perplexity literally defined in Oxford dictionary as *"Inability to deal with or understand something"* (Oxford English Dictionary, 2016). Moreover, perplexity is defined in Britannica Encyclopedia as *"the state of being very confused because something is difficult to understand"* (Encyclopedia Britannica, 2016). Furthermore, its definition in Cambridge Dictionary is *"a state of confusion or a complicated and difficult situation or thing"* (Cambridge Dictionary, 2016). Thus, inferring user perplexity in the HCI context is sensible given that confusion in *Affective State* detection is considerably important since it represents the actual entry point for artificial intelligence to facilitate adaptive HCI (Graesser et al., 2008).

For this reason, the role of the *Perception Component* of an adaptive HCI is to capture the episodes when the user is perplexed whilst they are interacting with software. However, recognising user perplexity and confusion is still non-trivial and a challenging objective particularly in the context of HCI. Therefore, experiments conducted in previous chapters investigated the use of vision-based channels for that purpose. Thus, multiple recorded sessions of a number of people who conducted HCI tasks were analysed using validated facial-expression detection models, and examined the existing relationship between facial expressions that refer to the common basic emotional states and the actual feelings they experienced and reported using subjective self-reporting tools. As mentioned in Chapter 4, within the HCI context, humans do not express themselves using facial expressions whilst interacting with a machine or UI. Consequently, the complexity of human emotions and the strangeness of the relationship between humans and machines provides a challenge in Affective Computing and relevant research themes.

Cognitive State Perception

Regardless the cause of the cognitive load, it is important to maintain the amount of cognitive load within the total cognitive capacity in order to avoid cognitive ageing, which results from working-memory capacity reduction due to irrelevant information. This consequently slows down mental processes, and thus affects overall performance (Van Merriënboer and Sweller, 2005). Therefore, measuring the amount of cognitive load in the HCI context would be a practical and beneficial pillar for the *Perception Component* and the achievement of Adaptive form of HCI.

As seen in Chapter 5, eye gaze-based data, specifically the pupil size, shows stronger correlation with activation and agitation of users. This provides greater insight into the cognitive load and mental processing when compared to the correlation between the subjectively experienced affective states and facial expressions. Also, one of the studies presented in Chapter 5 showed that the pupil features can be used, quantified and mapped into a dilation level. Consequently, the pupil size can be used as an index for the user's cognition, which is suitable to be exploited and employed as an indicator for the amount of cognitive load and mental effort and processing being exerted. Pupil size could be used as an index for the amount of cognitive load as well as mental processing gauge. Thus, the HCI-Viewer tool, which has been detailed in Chapter 3 and Appendix D, includes a meter that shows the dilation level of the pupil size in the current moment of the playback together with the video of the participant's face and the screen of the UI where the interaction is taking place. The meter presumed to represent the amount of mental effort which relates to cognitive load.

Fig. 7.5 Cognitive load meter that uses normalised pupil size of the current moment. The meter shows coloured ranges of three pupil dilation levels. Each colour represents a dilation level of pupil size as follows: green [0%-33.33%], yellow [33.33%-66.66%], and [66.66% -100%].

As shown in Figure 7.5, the dilation levels of the pupil are calculated based on the encoding scheme that uses equally separated levels that are in turn based on the normalised value of the pupil size during that session. For illustration purposes, a graph will be addressed herein for the pupil size changes throughout a whole session of one of the recordings from the Data Collection Study presented in Chapter 3.

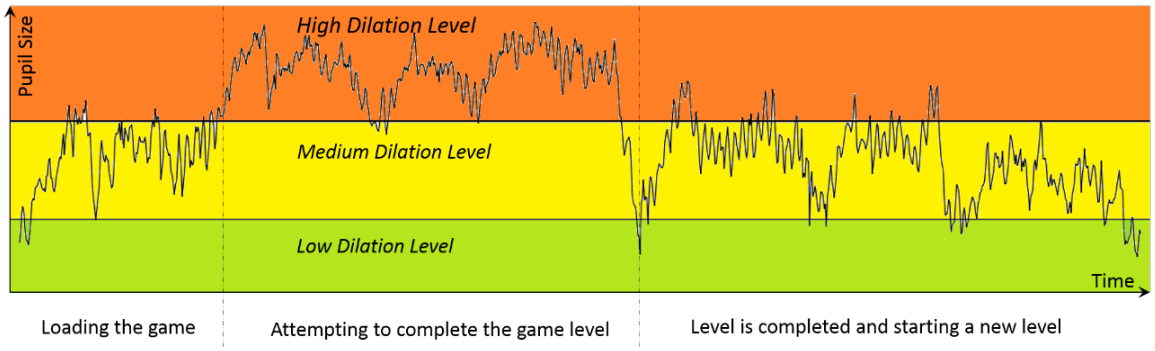


Fig. 7.6 Illustration of pupil dilation changes through a gaming context session. Red coloured area represents the zone of high-dilation level, Yellow coloured area represents the medium-dilation level, and Green coloured area represents the low-dilation level of the pupil size.

Figure 7.6 shows the pupil size variation during the session where the participant was playing Pacman. At the beginning, the participant launched the game and the game took some time to load. Apparently, in that time the pupil size values ranged within the yellow zone (*medium-dilation level*). Subsequently, during this period of time, the pupil size indicated a

normal cognitive load and a 'normal' level of mental effort. In the next stage the participant started the actual game by navigating the Pacman figure throughout the maze trying to collect the dots and avoiding the ghosts that roam around trying to terminate the Pacman figure. During this stage the participant was very focused and the pupil size obviously was dilated to the maximum level in the red zone, i.e. *high-dilation level*. Eventually, once the participant successfully completed the first level, the pupil size started to shrink to the *low-dilation zone* for a short time but then settled within the *medium-dilation level* as depicted in Figure 7.6. After that in the second level of the game, the participant reported feeling more relaxed as the required task was already achieved (i.e. completing the first level of the game). Thus, although the participant proceeding playing, they were not exerting the same amount of mental effort as used during the first level of the game.

Consequently, pupil size captured using eye-trackers can be used as an index that relates to the cognitive load and current amount of mental effort exerted on a task in real-time. Interestingly, this finding of the pupil size behaviour and the mental workload assessment concords with the *Task-Evoked Pupillary Response* theory that had been presented in the Literature Review in Chapter 2. *Task-Evoked Pupillary Response* is the linear relationship between pupil dilation and cognitive load resulting from the mental processing and the working memory demands caused by a task (Beatty, 1982).

7.3 Research Summary

A multi-modal dataset in typical HCI contexts has been collected throughout the PhD project. The aim of the Data Collection Study, presented in Chapter 3, was to collect features from different input modalities, which is used to reason about users' affective states whilst interacting with common computer software and attempting to complete typical computer-based tasks. A total of 42 participants took part in the study with different levels of computing experience, ranging from novice to expert computer users. Furthermore, the material for the tasks used had been chosen according to a study of computer usage statistics (Beauvisage, 2009), which presented the highest percentages of time spent on computer usage for both households and individuals. The tasks can be classified into four main categories: (1) basic operating system tasks; (2) online shopping tasks; (3) Excel spreadsheet manipulation tasks; (4) game-based tasks. Subsequently, the Data Collection Study was composed from 168 recorded sessions from 42 participants, where each participant recorded four sessions, and each session corresponded to a single task. Therefore, the dataset was generated to investigate the first research question: *"What are the emotional states that can be detected and utilised for the purpose of modelling the user within an adaptive HCI context?"*, in addition to exploring

the hypotheses derived from the second research question: *"What is the best method that can be used as input channel to capture the user state that would be useful for adaptive HCI?"*. Moreover, this dataset can be utilised for further analysis and investigations considering various aspects related to non-verbal recordings, tasks and UIs and relevant usability issues, or aspects related to self-reporting tools and methods.

In Chapter 4, a range of methods were investigated for automatic facial expression recognition using different supervised machine learning techniques against benchmark labelled facial expression datasets, in order to identify the possible answers for the third research question: *"What are the most efficient machine learning and statistical approaches to modelling user states within an adaptive HCI context?*". The experimental results obtained indicated a higher level of classification accuracy and robustness is achievable using a feature descriptor based on Euclidean distance as a representation of facial expressions, in comparison to using standard Cartesian coordinates from the facial landmark points. Furthermore, the HPBSVM classification technique presented facilitates more efficient and accurate classification of the facial expressions. The HPBSVM framework decomposes the overall classification process into smaller micro-decisions that are made by specialised classifiers trained differently for each single emotional state. This novel classification method showed a better accuracy level by making some features more discriminative for specific classes.

Referring to the first research question again, the work in the current thesis investigated the relationship between the facial expressions, as defined by Paul Ekman, and the self-reported emotional states specified by users using Russell's Circumplex model, in relation to the actual feelings and affective states as well as the identification of users' states in typical HCI contexts. Subsequently, the aforementioned studies employed facial expression analysis for detecting human affective states in HCI contexts, where users interacted with different UIs when attempting different tasks. The main conclusion advises that facial expressions cannot precisely reveal the actual feelings of users whilst interacting with common computerised tasks.

Moreover, relating to the second research question again, additional studies were conducted within this PhD, presented in Chapter 6, exploited visual-based features for task classification and pupil size variation tracking. Firstly, investigations explored user-driven task-based classification, whereby the classification algorithm deployed features from visual-based input modalities, i.e. facial expression, and eye gaze behaviour. The results indicated that the pattern alternation of facial expressions across different UIs can entail information that possibly distinguish between different computer tasks, i.e. UIs. Secondly, correlation analysis that has been conducted between eye-tracking data and features populated from

facial expression analysis showed that there is a moderate significant positive correlation achieved through the use of a linear regression model, which employs fiducial point features as independent variables, and pupil size measured by an infra-red-based eye-tracker as the dependent variable.

The aspiration for this research project is to utilise visual-based input modalities to infer user's *affective* and *cognitive* states, in order to facilitate intelligent and adaptive interface. To this end, the initial sections of the current Chapter described activities involve designing an adaptive HCI framework that comprises a *Perception Component* and an *Adaptation Component*. Furthermore, details are provided about the main role of the *Perception Component* to detect and predict the current *affective* and *cognitive* states of the user in real-time. Additionally, the *User's Perplexity* abstraction has been defined, which represents the *affective* state that the intelligent system should be able to capture and recognise in order to make the appropriate adaptation or assistance that supports task completion and helps the user to achieve desired goals. Moreover, *cognitive* load index using pupil size dilation is discussed, where the mental effort can be inferred by the amount of dilation that occurs whilst undertaking a task.

7.4 Limitations and Future Work

Regardless of substantial advancements in current technologies, endowing computers with the same level of intelligence to be able to resemble humans capabilities in perceiving user's feelings and emotional states remains a very challenging task. Therefore, further research is needed within HCI and Affective Computing research themes. Work is needed to enhance and improve the machine learning techniques alongside feature extraction and representation methods, where future work could undertake more in-depth experimentation and analysis.

While the collected data within this PhD represents the core foundation of the subsequent studies that addressed the research questions, the presented data has some limitations that could be addressed through further endeavours and future work, for example, the recorded sessions were constrained with time limits and performed within a lab context, therefore it would be interesting to conduct similar experiments and investigations within a naturalistic environment without the imposed conditions and constrains of the data collection protocol. Furthermore, the dataset that has been collected throughout current PhD project can be utilised for further analysis and investigations considering various aspects related to non-verbal modelling and interaction styles between humans and computers.

Referring to the overarching aim of the research presented in current thesis, that is exploiting visual-based input channels to populate facial-based features and eye gaze-based

feature in order to detect and recognise the spontaneous user feelings and experiences with HCI context, is possible and achievable. However, more studies need to be conducted to investigate other input channels for recognising the user's *affective* and *cognitive* states, along with the embodiment of real-time information about the task being conducted, for the purpose of making a comprehensive form of self-aware systems to achieve adaptive HCI. An additional suggestion for future work that could be beneficial in this research area, is to measure user's acceptance of using an adaptive form of HCI, through means of design and development of a real-world HCI context scenario for a system that attempts to perceive user's state and apply a suitable adaptation.

References

a.E. Kaufman, Bandopadhay, a., and Shaviv, B. (1993). An eye tracking computer user interface. *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*, pages 120–121.

Afzal, S. and Robinson, P. (2009). Natural affect data - collection & annotation in a learning context. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, pages 1–7.

Afzal, S., Sezgin, T. M., Gao, Y., and Robinson, P. (2009). Perception of emotional expressions in different representations using facial feature points. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*.

Aha, D. W. (1992). Generalizing from case studies: A case study. In *Proc. of the 9th International Conference on Machine Learning*, pages 1–10.

Ahn, H. I. and Picard, R. W. (2014). Measuring affective-cognitive experience and predicting market success. *IEEE Transactions on Affective Computing*, 5(2):173–186.

Akakin, H. Ç. and Sankur, B. (2010). Spatiotemporal-boosted DCT Features for Head and Face Gesture Analysis. In *Proceedings of the First International Conference on Human Behavior Understanding*, HBU'10, pages 64–74, Berlin, Heidelberg. Springer-Verlag.

Amershi, S., Conati, C., and Maclaren, H. (2006). Using feature selection and unsupervised clustering to identify affective expressions in educational games. In *Proceedings of Workshop on Motivational and Affective Issues in 8th International Conference on Intelligent Tutoring Systems*, pages 21–28.

Anderson, J. R. (1993). Rules of the Mind. 1993. *LawrenceErlbaumAssociates, Hillsdale, New Jersey*, pages ix, 320 p.

Andersson, R., Nyström, M., and Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research*, 3(3).

Aran, O., Ari, I., Guvensan, A., Haberdar, H., Kurt, Z., Turkmen, I., Uyar, A., and Akarun, L. (2007). A Database of Non-Manual Signs in Turkish Sign Language. *2007 IEEE 15th Signal Processing and Communications Applications*, pages 1–4.

Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2014). Incremental face alignment in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1859–1866.

Atchison, D. A. (2017). Optics of the Human Eye. In *Reference Module in Materials Science and Materials Engineering*.

Awad, M. and Khanna, R. (2015). *Support Vector Machines for Classification*, pages 39–66. Apress, Berkeley, CA.

Bal, E., Harden, E., Lamb, D., Van Hecke, A. V., Denver, J. W., and Porges, S. W. (2010). Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*, 40(3):358–370.

Balint, L. (1995). Adaptive interfaces for human-computer interaction: a colorful\nspectrum of present and future options. *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 1.

Bannert, M. (2002). Managing cognitive load—recent trends in cognitive load theory. *Learning and Instruction*, 12(1):139–146.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–636.

Barakova, E. I., Gorbunov, R., and Rauterberg, M. (2015). Automatic Interpretation of Affective Facial Expressions in the Context of Interpersonal Interaction. *Human-Machine Systems, IEEE Transactions on*, 45(4):409–418.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292.

Beauvisage, T. (2009). Computer usage in daily life. *Proceedings of the 27th international conference on Human factors in computing systems CHI 09*, page 575.

Benyon, D. and Murray, D. (1993). Applying user modelling to human-computer interaction design. *AI Review*, 7:43–69.

Biswas, P. and Langdon, P. (2011). A new input system for disabled users involving eye gaze tracker and scanning interface. *Journal of Assistive Technologies*, 5(2):58–66.

Biswas, P. and Robinson, P. (2010). A brief survey on user modelling in HCI. *Intelligent Techniques for Speech Image and Language Processing SpringerVerlag*.

Biswas, P., Robinson, P., and Langdon, P. (2012). Designing Inclusive Interfaces Through User Modeling and Simulation. *International Journal of Human-Computer Interaction*, 28(1):1–33.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. (2014). State-of-the-Art of Visualization for Eye Tracking Data. *Eurographics Conference on Visualization (EuroVis)*, pages 1–20.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, COLT '92, pages 144–152, New York, NY, USA. ACM.

Bradley, M. and Lang, P. J. (1994). Measuring Emotion: The Self-Assessment Semantic Differential Manikin and the. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(I):49–59.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information.

Bretzner, L., Laptev, I., Lindeberg, T., Lenman, S., and Sundblad, Y. (2001). A prototype system for computer vision based human computer interaction. *Report ISRN KTH/NA/P-01/09-SE*.

Bruce, V. (1992). What the human face tells the human mind: some challenges for the robot-human interface. *[1992] Proceedings IEEE International Workshop on Robot and Human Communication*.

Bruneau, D., Sasse, M. A., and McCarthy, J. (2002). The eyes never lie: The use of eye tracking data in HCI research. *Proceedings of the CHI*, 2:25.

Buettner, R. (2013). Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8077 LNAI, pages 37–48.

Bulling, A., Ward, J. A., Gellersen, H., and Tröster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):741–753.

Burbidge, R. and Buxton, B. (2001). An introduction to support vector machines for data mining. *Keynote papers, young OR12*, pages 3–15.

Burleson, W. and Picard, R. W. (2004). Affective agents: Sustaining motivation to learn through failure and a state of stuck. In *Workshop on Social and Emotional Intelligence in Learning Environments*.

Busso, C., Deng, Z., Yildirim, S., and Bulut, M. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. . . . *on Multimodal* . . . , pages 205–211.

Cain, B. (2007). A Review of the Mental Workload Literature. *Defence research and development Toronto (Canada)*, (1998):4–1–4–34.

Cambridge Dictionary, C. (2016). Cambridge Dictionary. *Meaning*, (entry 124):138–138.

Caprica Software (2009). VLCJ.

Card, S. K., Moran, T. P., and Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Commun. ACM*, 23(7):396–410.

Card, S. K., Newell, A., and Moran, T. P. (1983). *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., and Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Federation for Information Processing*, volume 247, pages 375–388.

Chandler, P. and Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4):293–332.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM - A Library for Support Vector Machines.

Chapanis, A. (1965). *Man Machine Engineering*. Behavioral Science in Industry Series. Wadsworth Publishing Company, Incorporated, Belmont, Calif.

Chen, S. and Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, 29(4):390–413.

Cheng, B. H. C., Lemos, R. D., Giese, H., Inverardi, P., Magee, J., Andersson, J., Becker, B., Bencomo, N., Brun, Y., Cukic, B., Serugendo, G. D. M., Dustdar, S., Finkelstein, A., Gacek, C., Geihs, K., Grassi, V., Karsai, G., Kienle, H. M., Kramer, J., Litoiu, M., Malek, S., Mirandola, R., Müller, H. a., Park, S., Shaw, M., Tichy, M., Tivoli, M., Weyns, D., and Whittle, J. (2009). Software Engineering for Self-Adaptive Systems: A Research Roadmap. *Software Engineering for Self-Adaptive Systems*, pages 1–26.

Chin, S. W., Seng, K. P., and Ang, L.-M. (2012). Audio-visual speech processing for human computer interaction. In *Advances in Robotics and Virtual Reality*, pages 135–165. Springer.

Chong, R. S. and Laird, J. E. (1997). Identifying dual-task executive process knowledge using EPIC-Soar. In *Proceedings of the nineteenth annual conference of the cognitive science society*, pages 107–112.

Colligan, L., Potts, H. W. W., Finn, C. T., and Sinkin, R. A. (2015). Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics*, 84(7):469–76.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.*, 20(3):273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and other kernel based learning methods.

Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1):125–151.

Den Uyl, M. J. and Van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. *Proceedings of Measuring Behavior*, 30:589–590.

Deng, J., Dong, W., Socher, R., and Li, L. (2009). A large-scale hierarchical image database. *Proc. CVPR*, pages 2–9.

DeRemer, S. (2015). Pupils Respond To More Than Light - Discovery Eye Foundation.

Dix, A., Finlay, J., Abowd, G. D., and Beale, R. (2004). *Human-Computer Interaction*, volume Third.

D'Mello, S., Graesser, A., and Picard, R. W. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4):53–61.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Commun. ACM*, 55(10):78–87.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Dumas, B., Lalanne, D., and Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5440 LNCS, pages 3–26.

Duric, Z., Gray, W. D., Heishman, R., Rosenfeld, A., Schoelles, M. J., Schunn, C., and Wechsler, H. (2002). Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, 90:1272–1289.

Ehmke, C. and Wilson, S. (2007). Indentifying Web Usability Problems from Eye-Tracking Data. *Proceedings of the HCI'07 Conference on People and Computers XXI*, 1:119–128.

Ekman, P. (1999). Facial expressions. In *Handbook of Cognition and Emotion*, volume 53, pages 226–232.

Ekman, P. (2003). *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124–129.

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System - Investigator's Guide*.

EL-Manzalawy, Y. (2005). WLSVM.

Encyclopedia Britannica (2016). Britannica Online Encyclopedia. *Encyclopædia Britannica*, pages 1–73.

Fairclough, S. H., Gilleade, K., Ewing, K. C., and Roberts, J. (2013). Capturing User Engagement via Psychophysiology: Measures and Mechanisms for Biocybernetic Adaptation. *Int. J. Auton. Adapt. Commun. Syst.*, 6(1):63–79.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition*, 36:259–275.

Fehrenbacher, D. D. and Djamasbi, S. (2017). Information systems and task demand: An exploratory pupillometry study of computerized decision making. *Decision Support Systems*, 97:1–11.

Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, 11(1-2):65–86.

Fraden, J. (1998). Handbook of Modern Sensors: Physics, Designs, and Applications, 2nd ed.

Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural networks : the official journal of the International Neural Network Society*, 18(4):389–405.

Freed, M. A., Shafto, M. G., and Remington, R. W. (1999). Employing simulation to evaluate designs: The APEX approach. In *Engineering for Human-Computer Interaction*, pages 207–223. Springer.

Frijda, N. H. and Parrott, W. G. (2011). Basic emotions or ur-emotions? *Emotion Review*, 3(4):406–415.

Gajos, K. Z., Czerwinski, M., Tan, D. S., and Weld, D. S. (2006). Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the working conference on Advanced visual interfaces - AVI '06*, page 201.

Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M. (1998). The belief-desire-intention model of agency. *Intelligent Agents V: Agents Theories, Architectures, and Languages. 5th International Workshop, ATAL'98.*, pages 1–10.

Ghimire, D. and Lee, J. (2013). Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors (Basel, Switzerland)*, 13:7714–34.

Goldberg, J. and Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization*, 10(3):182–195.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. *Nature*, 521(7553):800.

Gore, B. F. (2011). Man-Machine Integration Design and Analysis System (MIDAS) v5: Augmentations, Motivations, and Directions for Aeronautics Applications. *Learning*.

Graesser, A. C., D'Mello, S. K., Craig, S. D., Witherspoon, A., Sullins, J., Mcdaniel, B., and Gholson, B. (2008). The relationship between affective states and dialog patterns during interactions with AutoTutor. *Journal of Interactive Learning Research*, 19(2):293–312.

Gunes, H. and Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–1345.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: a survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500.

Hansen, D. W. and Pece, A. E. C. (2005). Eye tracking in the wild. *Computer Vision and Image Understanding*, 98(1):155–181.

Harezlak, K., Kasprowski, P., and Stasch, M. (2014). Towards accurate eye tracker calibration -methods and procedures. In *Procedia Computer Science*, volume 35, pages 1073–1081.

Hariharan, A. and Philipp Adam, M. T. (2015). Blended Emotion Detection for Decision Support. *Human-Machine Systems, IEEE Transactions on*, 45(4):510–517.

Harmon-Jones, E., Gable, P. A., and Price, T. F. (2012). The influence of affective states varying in motivational intensity on cognitive scope. *Frontiers in Integrative Neuroscience*, 6.

Harper, R., Rodden, T., Rogers, Y., and Sellen, A. (2008). *Being Human - Human-Computer Interaction in the Year 2020*.

Hart, S. G. (1986). NASA Task load Index (TLX). Volume 1.0; Paper and pencil package.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183.

Hart, S. G. and Wickens, C. D. (1990). Workload assessment and prediction. In *Manprint*, pages 257–296. Springer.

Hernandez, J., Paredes, P., Roseway, A., and Czerwinski, M. (2014). Under pressure: sensing stress of computer users. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 51–60. ACM.

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., and Verplank, W. (1992). ACM SIGCHI Curricula for Human-Computer Interaction. Technical report, New York, NY, USA.

Hoeks, B. and Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, 25(l):16–26.

Hofmann, M. (2006). Support vector machines-kernels and the kernel trick. *An elaboration for the Hauptseminar Reading Club SVM*.

Hollender, N., Hofmann, C., Deneke, M., and Schmitz, B. (2010). Integrating cognitive load theory and concepts of human-computer interaction. *Computers in Human Behavior*, 26:1278–1288.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Hartwig, A. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv*, page 9.

Howes, A., Vera, A., Lewis, R. L., and McCurdy, M. (2001). Cognitive Constraint Modeling : A Formal Approach to Supporting Reasoning About Behavior. *Proceedings of the 16th International Conference on Cognitive Modeling*, (1999):595–600.

Imotions (2016). Eye Tracking - The Definitive Guide - iMotions. *iMotions*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456.

Iqbal, S. T. and Bailey, B. P. (2004). Using Eye Gaze Patterns to Identify User Tasks. *The Grace Hopper Celebration of Women in Computing*, page 6.

Iqbal, S. T., Zheng, X. S., and Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. *Extended abstracts of the 2004 conference on Human factors and computing systems CHI 04*, page 1477.

Isezaki, T. and Suzuki, K. (2011). Depth image based analysis of facial expressions and head orientation. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 2537–2542. IEEE.

Jaimes, A. and Liu, J. (2005). Hotspot components for gesture-based interaction. In *Human-Computer Interaction-INTERACT 2005*, pages 1062–1066. Springer.

Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108:116–134.

Jameson, A. (2003). Adaptive interfaces and agents. *Human-computer interaction handbook*, 305:330.

Jannsen, B. (2008). Support Vector Machines for Binary Classification and its Applications.

Jaques, P. A. and Vicari, R. M. (2007). A BDI approach to infer student's emotions in an intelligent learning environment. *Computers and Education*, 49(2):360–384.

Jeon, M. (2017). *Emotions and affect in human factors and human-computer interaction*. Academic Press.

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

Joachims, T. (1999). Making large scale SVM learning practical. In *Advances in kernel methods: support vector learning*, pages 169 – 184.

Johanssen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., and Wickens, C. (1979). Final report of experimental psychology group. In *Mental Workload*, pages 101–114. Springer.

John, B., Vera, A., Matessa, M., Freed, M., and Remington, R. (2002). Automating CPM-GOMS. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 147–154.

John, B. E. and Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: comparison and contrast. *ACM Transactions on Computer-Human Interaction*, 3(4):320–351.

Joho, H., Jose, J. M., Valenti, R., and Sebe, N. (2009). Exploiting facial expressions for affective video summarisation. *CIVR '09*, page 1.

Kanade, T. and Cohn, J. (2000). Comprehensive database for facial expression analysis. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53.

Kapoor, A., Burleson, W., and Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human Computer Studies*, 65(8):724–736.

Kapoor, A. and Picard, R. W. (2005). Multimodal affect recognition in learning environments. *Proceedings of the 13th annual ACM international conference on Multimedia MULTIMEDIA 05*, page 677.

Karray, F., Alemzadeh, M., Saleh, J. A., and Arab, M. N. (2008). Human-Computer Interaction: Overview on State of the Art. *INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS*, 1:137–159.

Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3:33–48.

Kieras, D. E. (1994). GOMS Modeling of User Interfaces using NGOMSL. In *Proceedings of ACM CHI94 Conference on Human Factors in Computing Systems*, volume 2, pages 371–372.

Kieras, D. E. and Meyer, D. E. (1994). The EPIC architecture for Modeling Human Information-Processing and Performance: A Brief Introduction. *Security*, 1(1).

Kim, K. H., Bang, S. W., and Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42:419–427.

Kirakowski, J. and Corbett, M. (1990). *Effective methodology for the study of HCI*. North-Holland, Amsterdam u.a.

Kirschbaum, A. R. (1998). Method and apparatus to measure pupil size and position.

Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33.

Kleinsmith, A., Bianchi-Berthouze, N., and Steed, A. (2011). Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038.

Kneer, J., Elson, M., and Knapp, F. (2016). Fight fire with rainbows: The effects of displayed violence, difficulty, and performance in digital games on affect, aggression, and physiological arousal. *Computers in Human Behavior*, 54:142–148.

Ko, B. C. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, 18(2).

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31.

Krizhevsky, A., Sutskever, I., and Geoffrey E., H. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 1–9.

Lærd Statistics (2017). Pearson Product-Moment Correlation.

Laird, J. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

Lanatà, A., Valenza, G., and Scilingo, E. P. (2013). Eye gaze patterns in emotional pictures. *Journal of Ambient Intelligence and Humanized Computing*, 4:705–715.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning. *International Conference in Machine Learning*, page 38115.

Lee Rodgers, J. and Alan Nice Wander, W. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, 42(1):59–66.

Legin, A., Rudnitskaya, A., Seleznev, B., and Vlasov, Y. (2005). Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie. *Analytica Chimica Acta*, 534:129–135.

Lemaire, P., Ardabilian, M., Chen, L., and Daoudi, M. (2013). Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE.

Lew, M., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2:1–19.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Li, D., Babcock, J., and Parkhurst, D. (2006). openEyes: a low-cost head-mounted eye-tracking solution. *Etra*, 1(March):27–29.

Liew, C. F. and Yairi, T. (2015). Facial Expression Recognition and Analysis: A Comparison Study of Feature Descriptors. *IPSJ Transactions on Computer Vision and Applications*, 7(0):104–120.

Litman, D. and Forbes, K. (2003). Recognizing emotions from student speech in tutoring dialogues. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*.

Liu, P., Han, S., Meng, Z., and Tong, Y. (2014). Facial Expression Recognition via a Boosted Deep Belief Network.

Lopes, A. T., de Aguiar, E., De Souza, A. F., and Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610–628.

Lowenstein, O. and Loewenfeld, I. E. (1962). The pupil. *The eye*, 3:231–267.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 94–101.

Luckin, R. and Others (2007). Towards predictive modelling of student affect from web-based interactions. *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, 158:169.

Lundqvist, D., Flykt, A., and Ohman, A. (1998). The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, pages 91–630.

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE.

Mahmoud, M., Baltrušaitis, T., Robinson, P., and Riek, L. D. (2011). 3D Corpus of spontaneous complex mental states. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6974 LNCS, pages 205–214.

Mäkinen, E. (2008). Introduction to Computer Vision from Automatic Face Analysis Viewpoint. *Department of Computer Sciences University of Tampere, Finland*.

Marquardt, D. W. (1980). You Should Standardize the Predictor Variables in Your Regression Models. *Journal of the American Statistical Association*, 75(369):87–91.

Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity.

Martinez, A. M. (2011). Deciphering the face. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 7–12. IEEE.

Mavrikis, M., Grawemeyer, B., Holmes, W., Loibl, K., Hansen, A., Mavrikis, M., and Gutierrez-Santos, S. (2015). Light-Bulb Moment?: Towards adaptive presentation of feedback based on students' affective state.

McDuff, D., El Kaliouby, R., Senechal, T., Amr, M., Cohn, J. F., and Picard, R. (2013). Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected" In-the-Wild". In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 881–888. IEEE.

McDuff, D., Gontarek, S., and Picard, R. (2014). Remote measurement of cognitive stress via heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2957–2960. IEEE.

Menezes, M., Samara, A., Galway, L., Sant'Anna, A., Verikas, A., Alonso-Fernandez, F., Wang, H., and Bond, R. (2017). Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset. *Personal and Ubiquitous Computing*.

Mikut, R. and Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):431–443.

Miller, S. (2001). Workload measures. *National Advanced Driving Simulator. Iowa City, United States*.

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Moran, T. (1981). The command language grammar: A representation for the user interface of interactive computer systems. *International journal of man-machine studies*, pages 3–50.

Morimoto, C. H. and Mimica, M. R. M. (2005). Eye gaze tracking techniques for interactive applications.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Neoh, S. C., Zhang, L., Mistry, K., Hossain, M. A., Lim, C. P., Aslam, N., and Kinghorn, P. (2015). Intelligent facial emotion recognition using a layered encoding cascade optimization model. *Applied Soft Computing*, 34:72–93.

Newell, A. (1992). Unified theories of cognition and the role of Soar. In *Soar A cognitive architecture in perspective A tribute to Allen Newell Studies in cognitive systems Vol 10*, pages 25–79 ST – Unified theories of cognition and the.

Norman, D. A. (2002). *The design of everyday things*. Basic books.

Nummenmaa, L., Hyönä, J., and Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. *Emotion (Washington, D.C.)*, 6(2):257–268.

Oatley, K. and Johnson-Laird, P. N. (1987). Towards a Cognitive Theory of Emotions. *Cognition and Emotion*, 1(1):29–50.

Obrenovic, Z. and Starcevic, D. (2004). Modeling multimodal human-computer interaction. *Computer*, 37:65–72.

O'Brien, H. L. and Toms, E. G. (2008). What is user enagement? A Conceptual Framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

Ohno, T., Mukawa, N., and Yoshikawa, A. (2002). FreeGaze: a gaze tracking system for everyday gaze interaction. *ETRA '02: Proceedings of the symposium on Eye tracking research*, pages 125–132.

Olah, C. (2014). Conv Nets: A Modular Perspective. *Colah's Blog*, page 13.

Oliver, N. (1997). LAFTER: a real-time face and lips tracker with facial expression recognition. In *CVPR*, volume 33, pages 1369–1382.

Oviatt, S. (2003). User-centered modeling and evaluation of multimodal interfaces. In *Proceedings of the IEEE*, volume 91, pages 1457–1468.

Oviatt, S. L., Darrell, T., and Flickner, M. (2004). Multimodal Interfaces that Flex, Adapt and Persist. *Communications of the ACM*, 47:1–4.

Oxford English Dictionary (2016). Oxford English Dictionary Online.

Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 22:1424–1445.

Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, volume 91, pages 1370–1390.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., and Hays, J. (2016). WebGazer : Scalable Webcam Eye Tracking Using User Interactions. In *IJCAI*, pages 3839–3845.

Pazzani, M. J. and Billsus, D. (2007). Content-Based Recommendation Systems. *The Adaptive Web*, 4321:325–341.

Picard, R. W. (2000a). *Affective Computing*. MIT Press.

Picard, R. W. (2000b). Affective perception. *Communication of the ACM*, 43(3):50–51.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1):55–64.

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

Platt, J. C. (1998). Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *Advances in kernel methods*, pages 185 – 208.

Polzin, T. S. and Waibel, A. (1998). Detecting emotions in speech. In *Proceedings of the CMC*, volume 16. Citeseer.

Pomplun, M. and Sunkara, S. (2003). Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction. In *Human-centered computing: cognitive, social and ergonomic aspects*, pages 542–546.

Poole, A. and Ball, L. J. (2005). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In *Encyclopedia of Human-Computer Interaction*, pages 211–219.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715–734.

Pratt, L. Y. (1993). Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*, pages 204–211.

Qureshi, N. A. and Perini, A. (2009). Engineering adaptive requirements. In *Software Engineering for Adaptive and Self-Managing Systems, 2009. SEAMS '09. ICSE Workshop on*, pages 126–131.

Rao, A. S. and Georgeff, M. P. (1995). BDI agents: From theory to practice. In *ICMAS*, volume 95, pages 312–319.

Recarte, M. A., Pérez, E., Conchillo, A., and Nunes, L. M. (2008). Mental workload and visual impairment: differences between pupil, blink, and subjective rating. *The Spanish journal of psychology*, 11(2):374–385.

Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, 52:185–218.

Reisner, P. (1981). Formal Grammar and Human Factors Design of an Interactive Graphics System. *IEEE Transactions on Software Engineering*, SE-7(2):229–240.

Robinson, D. A. (1963). Movement Using a Scieral Search in a Magnetic Field. *IEEE Transactions on Bio-Medical Electronics*, 10:137–145.

Robles-De-La-Torre, G. (2006). The Importance of the sense of touch in virtual and real environments. *IEEE Multimedia*, 13:24–30.

Romera-Paredes, B., Argyriou, A., Berthouze, N., and Pontil, M. (2012). Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959.

Roy, N. (2014). DeluxePacman2.

Rozado, D., El Shoghri, A., and Jurdak, R. (2015). Gaze dependant prefetching of web content to increase speed and comfort of web browsing. *International Journal of Human-Computer Studies*, 78:31–42.

Rubio, S., Díaz, E., Martín, J., and Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, 53(1):61–86.

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., Mutlu, B., and McDonnell, R. (2015). A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum*, TBD(epud ahead of print):n/a–n/a.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Russell, J. and Lemay, G. (2000). Emotion Concepts. Handbook of Emotion. MH-J. Lewis, M. New York.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Russell, J. A. (2003). Core affect annd the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Saffer, D. (2009). *Designing for Interaction: Creating Innovative Applications and Devices*. Voices that matter. New Riders.

Salah, a., Sebe, N., and Gevers, T. (2009). Communication and automatic interpretation of affect from facial expressions. *Affective Computing and Interaction: ...*, pages 157–183.

Samara, A., Galway, L., Bond, R., and Wang, H. (2017). Affective state detection via facial expression analysis within a human–computer interaction context. *Journal of Ambient Intelligence and Humanized Computing*.

Saneiro, M., Santos, O. C., Salmeron-Majadas, S., and Boticario, J. G. (2014). Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *Scientific World Journal*, 2014.

Sano, A. and Picard, R. W. (2013). Stress Recognition Using Wearable Sensors and Mobile Phones. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 671–676.

Sano, A. and Picard, R. W. (2014). Understanding Ambulatory and Wearable Data for Health and Wellness.

Sarrafzadeh, A., Fan, C. F. C., Dadgostar, F., Alexander, S., and Messom, C. (2004). Frown gives game away: affect sensitive systems for elementary mathematics. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 1.

Schiessl, M., Duda, S., Thölke, A., and Fischer, R. (2003). Eye tracking and its application in usability and media research. *MMI Interaktiv Journal*, (6):1–10.

Schnipke, S. K. and Todd, M. W. (2000). Trials and Tribulations of Using an Eye-tracking System. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, pages 273–274, New York, NY, USA. ACM.

Schrammel, J., Paletta, L., and Tscheligi, M. (2010). *Exploring the possibilities of body motion data for human computer interaction research*. Springer.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer.

Sebe, N. (2009). Multimodal interfaces: Challenges and perspectives. *Journal of Ambient Intelligence and Smart Environments*, 1:23–30.

Serrano, B., Baños, R. M., and Botella, C. (2016). Virtual reality and stimulation of touch and smell for inducing relaxation: A randomized controlled trial. *Computers in Human Behavior*, 55:1–8.

Shan, C., Gong, S., and McOwan, P. W. (2007). Beyond Facial Expressions: Learning Human Emotion from Body Gestures. *Procedings of the British Machine Vision Conference 2007*, pages 43.1–43.10.

Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816.

Sharma, C. and Dubey, S. K. (2014). Analysis of eye tracking techniques in usability and HCI perspective. In *2014 International Conference on Computing for Sustainable Global Development, INDIACom 2014*, pages 607–612.

Sharp, H., Rogers, Y., and Preece, J. (2011). *Interaction Design: Beyond Human-Computer Interaction*, volume 11.

Shi, Y., Park, T., Ruiz, N., Taib, R., Choi, E. H. C., and Chen, F. (2007). Galvanic Skin Response (GSR) as an Index of Cognitive Load. *CHI EA '07 CHI '07 Extended Abstracts on Human Factors in Computing Systems*, pages 2651–2656.

Sifre, L. (2014). *Rigid-Motion Scattering For Image Classification*. PhD thesis.

Singh, H. and Singh, J. (2012). Human Eye Tracking and Related Issues: A Review. *International Journal of Scientific and Research Publications*, 2(1):2250–3153.

Slaney, M., Rajan, R., Stolcke, A., and Parthasarathy, P. (2014). Gaze-enhanced speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3236–3240. IEEE.

Sobol-Shikler, T. (2009). Analysis of affective expression in speech. *month*.

Soleymani, M., Pantic, M., and Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3:211–223.

Sourina, O. and Liu, Y. (2013). EEG-enabled affective applications. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 707–708.

Stangor, C., Jhangiani, R., and Hammond, T. (2014). *Principles of social psychology*. BC Campus.

Steichen, B., Conati, C., and Carenini, G. (2014). Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *Tiis*, 4(2):11:1—-11:29.

Stephanidis, C. (2001). User interfaces for all: New perspectives into human-computer interaction. *User Interfaces for All-Concepts, Methods, and Tools*, 1:3–17.

Sullivan, J. W. and Tyler, S. W. (1991). *Intelligent User Interfaces*. ACM Press Series. ACM Press.

Sun, D., Paredes, P., and Canny, J. (2014). MouStress: detecting stress from mouse motion. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 61–70. ACM.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In *Cognition and MultiAgent Interaction*, pages 79–99.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Sweller, J., van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3):251–296.

Swinscow, T. (1976). "Statistics at square one": Correlation. *British medical journal*, 2(6042):680–681.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 1–9.

TATTERSALL, A. J. and FOORD, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5):740–748.

Tian, Y. L., Kanade, T., and Conn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115.

Tonn-Eichstädt, H. (2006). Measuring website usability for visually impaired people-a modified GOMS analysis.

Torrey, L. and Shavlik, J. (2009). Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends*, pages 242–264.

Tryon, W. W. (1975). Pupillometry: A Survey of Sources of Variation. *Psychophysiology*, 12(1):90–93.

Van Merriënboer, J. J. G. and Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, 24(6):774–780.

Vega, K., Arrieta, A., Esteves, F., and Fuks, H. (2014). FX e-Makeup for Muscle Based Interaction. In *Design, User Experience, and Usability. User Experience Design for Everyday Life Applications and Services*, pages 643–652. Springer.

Vertegaal, R. and Vertegaal, R. (2003). Attentive User Interfaces. *Commun. ACM*, 46:30–33.

Vicente, A. D. and Pain, H. (2002). *Informing the Detection of the Students' Motivational State: an Empirical Study*, volume 2363.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition (CVPR)*, 1:I—-511—-I—-518.

Vroomen, J., Collier, R., and Mozziconacci, S. (1993). Duration and intonation in emotional speech. In *Third European Conference on Speech Communication and Technology*, volume 1p, pages 577–580.

Wang, H., Chignell, M., and Ishizuka, M. (2006). Empathic tutoring software agents using real-time eye tracking. *Proceedings of the 2006 symposium on Eye tracking research & applications - ETRA '06*, pages 73–78.

Wang, J. (2011). Pupil dilation and eye tracking. *A handbook of process tracing methods for decision . . .* , pages 1–33.

Wang, J., Zhang, G., and Shi, J. (2015). Pupil and glint detection using wearable camera sensor and near-infrared LED array. *Sensors (Switzerland)*, 15(12):30126–30141.

Weiser, M. and Brown, J. S. (1996). The coming age of calm technolgy. In *Beyond Calculation*, pages 75–85.

Whitehill, J., Bartlett, M., and Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6.

Whitehill, J., Bartlett, M. S., and Movellan, J. R. (2013). Automatic facial expression recognition. *Social emotions in nature and artifact*, 88.

Winfield, D. and Parkhurst, D. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, 3:79–79.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Xu, H., Caramanis, C., and Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510.

Yang, M. H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

Yik, M., Russell, J. A., and Steiger, J. H. (2011). A 12-Point Circumplex Structure of Core Affect. *Emotion*, 11(4):705–731.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2):151–175.

Zakharov, K., Mitrovic, A., and Johnston, L. (2008). Towards emotionally-intelligent pedagogical agents. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5091 LNCS, pages 19–28.

Appendix A

Table A.1 presents a survey of different models that are related to Human-Computer Interaction research

Table A.1 Comparative study of HCI relevant models

Year	Publication	Model	Description	Comment
1980	The keystroke-level model for user performance time with interactive systems (Card et al., 1980)	keystroke-level model (KLM)	Focuses on predicting user's performance; in particular the time an expert needs to perform a certain task using quantitative analysis of keystrokes and other low-level operations along with system's responses.	Only applies for experts and does not address beginners or intermediate users.
1981	Formal Grammar and Human Factors Design of an Interactive Graphics System (Reisner, 1981)	Formal grammatical description in particular "action languages"	Predictive tool to compare alternative designs for ease of use such as the action language for analysing and comparing interactive graphics systems. This tool was proposed to identify design inconsistencies from human point of view in the design cycle.	This kind of models is intended for interaction techniques and figuring out user's knowledge and competence rather than measuring performance during interaction.
1981	The Command Language Grammar: a representation for the user interface of interactive computer systems (Moran, 1981)	Command Language Grammar (CLG)	It took a top-down approach to decompose an interaction task and give a conceptual view of the interface before its implementation. Each level description contains procedures for accomplishing the tasks addressed by the system in terms of the actions available at that Level. That is the system is described by progressive refinement.	It completely ignores the human aspect of the interaction and does not model the capabilities and limitations of users.

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
1983	The psychology of Human–Computer Interaction (Card et al., 1983)	Human Processor Model (HPM) GOMS	Considered the first HCI model that take into considerations user's human aspects. HPM looks to understand and measure human performance in HCI context. HPM embodies a set of memories and processors with a set of operations. GOMS model defines the usability approach for accomplishing a Goal using available Methods which are composed of Operators taking into account Selections rules to choose the proper method.	Although HPM embodies a set of memories and processors with a set of operations in terms of HCI, and it is useful in design stage; it does not fit building intelligent computers that simulate humans' intelligence in human-human interaction.
1987	Soar : an architecture for general intelligence (Laird, 1987)	SOAR	SOAR models human cognition as a rule based system that exploits a learning technique to convert the course of operations into a production which will be used in situations when the user is not able to complete a task due to insufficient knowledge that is an impasse state, therefore this production can be used in similar situations when happened again.	SOAR mechanisms focus on working memory relation with efficiency, and don't address psychological aspects in reasoning processes.

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
1987	Intention, Plans, and Practical Reason (Bratman, 1987)	The Belief Desire Intention theory (BDI)	Theory of human practical reasoning, and discusses each component of this architecture and the rationale behind it. It theoretically captures the main components of practical reasoning, which are: intention handling, execution, option generation and deliberation.	It describes a higher level of abstraction and therefore not useful for practical rational reasoning systems.
1993	Rules of the Mind. 1993 (Anderson, 1993)	Adaptive Control of Thought-Rational (ACT-R)	Helps to derive assumptions about human cognition. ACT-R describes how humans apply knowledge to solve problems and achieve their goals. Moreover, it shows that human memory is divided to declarative and procedural memories. On one hand, declarative memory stores the goals and related facts. On the other hand, procedural memory include procedures and related rules for operation.	ACT-R definition is not precise and lacks of specificity and consistency; and therefore not clear how to use. Moreover, it contains too many parameters which affect performance dangerously.

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
1994	The EPIC architecture for Modelling Human Information-Processing and Performance: A Brief Introduction (Kieras and Meyer, 1994)	Executive Process-Interactive Control (EPIC)	Can be used for modelling human cognition and multiple task performance, and shows how humans perform complex multi-modal tasks by carrying out a set of production rules. However, it does also represent human intelligence together with the constraints on human abilities while performing tasks. EPIC architecture represented as a set of separated explicit processors: perceptual processors, cognitive processor, and motor processors.	EPIC can be described as an engineering model in practical domains such as computer user interface design. But there is nothing about adaptation and intelligence. Also it does not problem solving procedures. Furthermore, modeller needs to provide more details about perceptual and motor processors.
1994	GOMS modelling of user interfaces using NGOMSL (Kieras, 1994)	Natural GOMS Language (NGOMSL)	A structured language notation to describe the goals structure, and the sequence of operations prediction along with prediction of time to learn and time to complete operations.	This model explicitly represents the goal structure just like the CMN-GOMS and can so represent high-level goals. WHAT??
1995	BDI agents: From theory to practice (Rao and Georgeff, 1995)	Belief desire intention software model	BDI software architecture represents an implementation of BDI agent that is a model of an intelligent agent in terms of Beliefs, Desires and Intentions.	BDI software model does not consider agent interaction with other agents and does not support agent learning to be integrated in multi agent system or to adapt behaviour.

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
1997	Identifying dual-task executive process knowledge using EPIC-Soar (Chong and Laird, 1997)	EPIC-Soar model	It is a task independent learning procedure by combining EPIC with SOAR. Basically, they exploited perceptual and motor mechanisms existed in EPIC model with chunking mechanism existed in SOAR for problem solving procedures.	Not well implemented and it is difficult to be used.
1999	Employing simulation to evaluate designs: The APEX approach (Freed et al., 1999)	APEX	It is a human operator model that copes with time pressure and uncertainty inherent in multi-task environments. In particular, APEX is intended for realistic simulations of human pilots in cockpits and air traffic controllers. Procedure is represented in APEX using Procedure Definition Language (PDL) and it represents an operator knowledge about how to perform the task.	Modelling language has intrusions from general-purpose simulation engine. Psychological mechanisms are limited to those in CPM-GOMS

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
2001	Cognitive Constraint Modelling : A Formal Approach to Supporting Reasoning About Behaviour (Howes et al., 2001)	Constraint-based Optimizing Reasoning Engine (CORE)	CORE is based on Cognitive Constraint Modelling (CCM). It defines constraints and links between tasks, mental processes and peripheral events from environment; and builds the reasoning based on these constraints to model cognition.	CORE models cognitions as a set of constraints and an objective function. And it has a prediction strategy by solving constraints satisfaction problem instead of tasks hierarchy and structure used in other models
2002	Automating CPM-GOMS (John et al., 2002)	CPM-GOMS	It combines GOMS analysis with human processor usage at the level of cognitive, perceptual, and motor operations to predict skilled behaviour. Assumes that operators of the cognitive processor, perceptual processor, and the motor processor can work in parallel to each other. The most important point of CPM-GOMS is the ability to predict skilled behaviour from its ability to model overlapping actions.	Different from other GOMS family models that is all of the aforementioned assume all of the operators occur in sequence and do not contain operators that are below the activity level.

Continued on next page

Table A.1 – *Continued from previous page*

Year	Publication	Model	Description	Comment
2006	Measuring website usability for visually impaired people-a modified GOMS analysis (Tonn-Eichstädt, 2006)	Modified GOMS model for impaired people	An interaction model of blind users' interaction strategies. It can be used to measure aspects of website usability for blind users. The model evolved from findings of user observations and field studies. It can be applied to specific layouts in order to find the 'best' alternative.	This model is designated for specific domain.
2006	The CLARION cognitive architecture: Extending cognitive modelling to social simulation (Sun, 2006)	CLARION	It comprises of dual separated structures: implicit and explicit representations for processes explored in all functional subsystems. Moreover, CLARION adopts bottom-up learning approach, where implicit knowledge acquired before explicit.	This model is useful to investigate learning mechanisms, skills acquisition, and problem solving in humans.
2011	Man-machine Integration Design and Analysis System (MIDAS) v5: Augmentations, Motivations, and Directions for Aeronautics (Gore, 2011)	Man-machine Integration Design and Analysis System (MIDAS)	MIDAS is basically a try to integrate cognitive and performance models with the early mentioned HPM model. MIDAS is one of the tools in NASA Ames Research Centre used to predict human performance across different domains.	It is well developed for aviation specific applications including an approach to model human error for NASA Aviation Safety Program, however it is overcomplicated.

Appendix B

Table B.1 presents a survey of user modelling in the literature

Table B.1 Comparative study of user modelling literature

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2016	Fight fire with rainbows: The effects of displayed violence, difficulty, and performance in digital games on affect, aggression, and physiological arousal (Kneer et al., 2016)	1) Psychophysiological arousal using inter-beat intervals (IBI) and electrodermal activity (EDA). 2) Cognition using lexical decision task similar. 3) Aggressive behaviour using Competitive Reaction Time Task. 4) Post-game emotions reporting	Digital games	1) ANOVAs for the lexical decision task response latencies, and mean volume and duration scores. 2) rANOVAS for inter-beat interval and electro-dermal activity scores. 3) MANOVAs for the post-game affect scores	Joy, Surprise, Interest, Fun, Contentment, Love, Dolefulness, Anger, Disgust, Contempt, Fear, Shame, Guilt, Fascination, Enchantment, Boredom
2016	Virtual reality and stimulation of touch and smell for inducing relaxation: A randomized controlled trial (Serrano et al., 2016)	Subject feedback using questionnaires and self-assessment reports	Mood induction procedures in a virtual reality environment	Statistical analysis	Relaxation, Joy, Affective Valence, Arousal, Anxiety , Sadness

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2015	Intelligent facial emotion recognition using a layered encoding cascade optimization model (Neoh et al., 2015)	Facial expressions	Frontal images extracted respectively from the CK+ and MMI databases and 90 side-view images extracted from MMI	1) Feature extraction using a modified overlap Local Gabor Binary Patterns LGBP. 2) Discriminative feature selection using both of the proposed direct similarity and Pareto-based evolutionary algorithms. 3) Expression recognition using a neural network and an adaptive ensemble classifier.	Disgust, Anger, Happiness, Surprise, Fear, Sadness, Neutral
2015	Automatic Interpretation of Affective Facial Expressions in the Context of Interpersonal Interaction (Barakova et al., 2015)	1) Behaviour in a cooperative computer game. 2) Self-assessment questionnaires. 3) Video records of facial expressions during game play.	Collaborative gaming	Genetic programming (GP)	Emotional behaviour: Happy and Sad

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2015	Light-Bulb Moment?: Towards adaptive presentation of feedback based on students' affective state (Mavrikis et al., 2015)	Student feedback	E-learning environment	Researchers categorise affective states	Enjoyment, Surprise, Confusion, Frustration, Boredom
2015	Blended Emotion Detection for Decision Support (Hariharan and Philipp Adam, 2015)	1) Self-report. 2) Physiological measurements of skin conductance response (SCR) and heart rate (HR)	Financial decision trading	1) C4.5. 2) CART. 3) Random forest algorithms	Rejoice, Regret
2015	Sentiment, emotion, purpose, and style in electoral tweets (Mohammad et al., 2015)	Sentiment analysis electoral tweets	Electoral tweets	1) Automatic SVM. 2) Automatic rule-based	Joy, Sadness, Anger, Fear, Surprise, Anticipation, Trust, Disgust
2014	Measuring Affective-Cognitive Experience and Predicting Market Success (Ahn and Picard, 2014)	Facial expressions	Tasting beverages and answering Questions on the computer	1) Matthews correlation coefficient (MCC). 2) Likelihood ratios (LR).	Affective wanting, Affective liking

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2014	Remote Measurement of Cognitive Stress via Heart Rate Variability (McDuff et al., 2014)	1) Contact measurements of the blood volume pulse (BVP). 2) Electro-dermal activity were collected using finger sensors (EDA). 3) Breathing measured using a chest strap. 4) A camera, placed 3m from the participant.	Set down relax for Rest, and perform a mental arithmetic task for Stress	1) Bayesian classifier. 2) SVM	Rest and Stress
2014	Understanding Ambulatory and Wearable Data for Health and Wellness (Sano and Picard, 2014)	EEG, EDA, ACC AND mobile phone usage	Signals over different days. (physiological, behavioural, environmental, and social) ambulatory data	Using 6 different classifiers	Stress level Mood and Sleep behaviours

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2014	Under Pressure: Sensing Stress of Computer Users (Hernandez et al., 2014)	Pressure-sensitive keyboard and capacitive mouse	Experimental Tasks: 1) Text Transcription. 2) Expressive Writing. 3) Mouse Clicking	Statistics analysis: 1) Wilcoxon signed-rank test. 2) Kruskal–Wallis test	1) Relaxed/Stressed. 2) Valence: (very-unpleasant to very-pleasant). 3) Arousal (low-energy to high-energy)
2013	Stress Recognition using Wearable Sensors and Mobile Phones (Sano and Picard, 2013)	1) A wrist sensor (accelerometer and skin conductance). 2) Mobile phone usage (call, short message service, location and screen on/off). 3) Surveys (stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake and electronics usage)	Signals over different days.	1) Correlation analysis. 2) Classification: SVM, KNN, PCA+SVM, PCA+KNN	Stress Recognition

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2012	Multimodal Emotion Recognition in Response to Videos (Soleymani et al., 2012)	EEG, pupillary response and gaze distance.	Emotional video clips	SVM with RBF kernel	1) The Arousal classes are Calm, Medium aroused, and Activated. 2) The Valence classes are Unpleasant, Neutral, and Pleasant.
2012	Eye gaze patterns in emotional pictures (Lanatà et al., 2013)	Eye gaze patterns and Pupil size	Pictures extracted from the International affective picture system	Recurrence quantification-analysis (RQA)	Neutral Arousal
2011	3D corpus of spontaneous complex mental states (Mahmoud et al., 2011)	Hand-over-face gestures	Two types of sessions: 1) Interaction with a computer program. 2) Interaction with another person	Fleiss' kappa statistical measure	Bored, Happy, Interested, Thinking, Unsure, Other

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2010	Emotion Recognition in Children with Autism Spectrum Disorders: Relations to Eye Gaze and Autonomic State (Bal et al., 2010)	Eye gaze, Heart Rate with Respiratory Sinus Arrhythmia (RSA) using ECG	The Dynamic Affect Recognition Evaluation (DARE; Porges et al. 2007), which is stimuli includes video files (i.e. still images)	Statistical analysis	Anger, Disgust, Fear, Happiness, Sadness, Surprise
2009	Exploiting Facial Expressions for Affective Video Summarisation (Joho et al., 2009)	Facial expressions	Video clips taken from the contents in different genres.	Bayesian classifier	Facial expressions of the viewers were grouped into three pronounced levels: 1) No (Neutral). 2) Low (Angry, Disgust, Fear, Sad). 3) High (Happy and Surprise).
2009	Perception of Emotional Expressions in Different Representations Using Facial Feature Points (Afzal et al., 2009)	Facial expressions	Simulated driving scenarios and a computer-based learning setting	Expert coders label the samples	Interested, Confused, Bored, Happy, Surprised

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2009	Natural Affect Data Collection and Annotation in a Learning Context	Facial expressions	Computer-based learning scenarios	Expert coders label the samples	Confused, Interested, Surprised, Bored, Happy, Annoyed, Neutral, Other
2008	Automatic Facial Expression Recognition for Intelligent Tutoring Systems (Whitehill et al., 2008)	Facial expressions	Lecture videos	1) Correlation analyses. 2) Regression models	Difficulty level and speed of content
2008	Towards emotionally-intelligent pedagogical agents (Zakharov et al., 2008)	Facial expressions	Pedagogical agent-based educational environment	Algorithm uses actions units listed in FACS	Positive and Negative valence
2007	Beyond Facial Expressions: Learning Human Emotion from Body Gestures (Shan et al., 2007)	Facial Expressions and Body Gestures	The Bimodal Face and Body Gesture Database (FABO) (Gunes, H.; Piccardi, M 2006)	1) SVM. 2) Canonical Correlation Analysis (CCA)	Anger, Anxiety, Boredom, Disgust, Joy, Puzzle, Surprise

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2007	Multimodal human–computer interaction: A survey (Jaimes and Sebe, 2007)	Facial expression recognition, Emotion in audio, and combination.	Review	1) Algorithm uses actions units listed in FACS. 2) Bayesian classifier. 3) HMM.	1) Joy, fear, love, surprise, sadness, etc., 2) Valence describes the pleasantness of the stimuli, with positive or pleasant (e.g. happiness) on one end, and negative or unpleasant. The other dimension is arousal or activation. 3) Stress, inattention, anger, boredom
2007	A BDI approach to infer students' emotions in an intelligent learning environment (Jaques and Vicari, 2007)	User's actions and interaction patterns	Pedagogical agent-based educational environment	OCC psychological model	Valence (positive or negative), reactions

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2007	A Towards An Affect-Sensitive Autotutor (D'Mello et al., 2007)	Posture, dialogue and task information	Dialogue based ITS-Auto Tutor	1) Bayesian classifier. 2) Neural networks. 3) Logistic regression. 4) Nearest neighbour. 5) C4.5 decision trees. 6) Additive logistic regression	Flow, Confusion, Boredom, Frustration, Neutral
2007	Automatic prediction of frustration (Kapoor et al., 2007)	Facial expressions, Posture, Mouse pressure, Skin conductance (SCR), Task state	Automated Learning Companion	1) Nearest neighbour. 2) SVM with RBF kernel. 3) Gaussian process.	Pre-frustration, Not Pre-frustration
2007	Towards Predictive Modelling of Student Affect from Web-Based Interactions (Luckin and Others, 2007)	Interaction logs and Situational factors	Interactive Learning Environment- WALLIS		Frustration, Confusion, Boredom, Confidence, Interest , Effort
2007	Multimodal emotion recognition from expressive faces, body gestures and speech (Caridakis et al., 2007)	Expressive faces, Body gestures and Speech	Act and perform specific gestures that exemplify each emotion	Bayesian classifier	Anger, Despair, Interest, Irritation, Joy, Pleasure, Pride, Sadness

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2006	Eye Movement Assessment of Selective Attentional Capture by Emotional Pictures (Nummenmaa et al., 2006)	Eye Movement	128 stimuli pictures chosen from the International Affective Picture System (IAPS; Lang, Bradley, and Cuthbert, 2005).	ANOVA	Unpleasant, Neutral, Pleasant
2006	Using Feature Selection and Un supervised Clustering to Identify Affective Expressions in Educational Games (Amershi et al., 2006)	Skin conductance, Heart rate, EMG	Educational game-Prime Climb	Bayesian classifier	Affective reactions to game events
2006	Bi-modal emotion recognition from expressive face and body gestures (Gunes and Piccardi, 2007)	Facial expressions and Upper body gestures	The Bimodal Face and Body Gesture Database (FABO) (Gunes, H.; Piccardi, M 2006)	Bayesian classifier	Anger, Disgust, Fear, Happiness, Uncertainty, Anxiety
2005	Multimodal Affect Recognition in Learning Environments (Kapoor and Picard, 2005)	Facial expressions, Posture patterns and Task state	Educational Puzzle	1) SVM. 2) Bayesian classifier	Interest, Disinterest

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2004	Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information (Busso et al., 2004)	Audio, Facial expression and Bimodal information	An actress who read 258 sentences expressing the emotions.	1) Kernel regression. 2) Bayesian classifier. 3) Nearest neighbour	Sadness, Anger, Happiness, Neutral
2004	Frown gives game away: Affect sensitive tutoring systems for elementary mathematics (Sarrafzadeh et al., 2004)	Facial expressions	Elementary Maths ITS	Constraint-based modelling (CBM)	Happiness/Success, Surprise/Happiness, Sadness/Disappointment, Confusion, Frustration/Anger
2003	Recognizing emotions from student speech in tutoring dialogues (Litman and Forbes, 2003)	Acoustic-prosodic cues, discourse markers	Physics Intelligent Tutoring Dialogue System–ITSPOKE	AdaBoost, J48, Nearest neighbour, ZeroR	Negative, Neutral, Positive

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2002	Integrating Perceptual and Cognitive Modelling for Adaptive and Intelligent Human–Computer Interaction (Duric et al., 2002)	Eye region: Eyelids and Eyebrows, Facial expressions, Mouse gestures (speed of movement, force of click, and directness of movement to the clicked object), Combination	ARGUS, a simulated task environment for radar operator tasks	1) HMM. 2) Bayesian classifier. 3) Learned vector quantization (LVQ)	Anger, Surprise, Happiness, Baseline, Easy, Difficult, Alertness, Fatigue, Confusion, Stress, Momentary lapses, Misunderstanding
2002	Informing the Detection of the Students' Motivational State: an Empirical Study (Vicente and Pain, 2002)	User actions and interaction patterns; Experience sampling	Japanese numbers ITSMOODS	Self-report	1) Traits: Control, Challenge, Independence, Fantasy. 2) States: Confidence, Sensory Interest, Cognitive Interest, Effort, Satisfaction.
2002	Probabilistic Assessment of User's Emotions in Educational Games (Conati, 2002)	Interaction patterns, personality, goals	Educational game-Prime Climb	1) Decision Networks. 3) Dynamic Decision Network.	Valence (positive or negative) reactions

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
2001	Toward Machine Emotional Intelligence: Analysis of Affective Physiological State (Picard et al., 2001)	Physiological signals: 1) A triode electromyogram measuring facial muscle tension along the masseter. 2) Photoplethysmyograph measuring blood volume pressure. 3) A skin conductance sensor measuring electro-dermal activity. 4) Effect respiration sensor placed around the diaphragm.	Signals over different days.	1) Nearest neighbour. 2) Maximum a posteriori probability (MAP).	Neutral, Anger, Hate, Grief, Platonic love, Romantic love, Joy, Reverence

Continued on next page

Table B.1 – *Continued from previous page*

Year	Publication	Techniques	Learning Context	Analysis	Aspects
1998	Detecting Emotions in Speech (Polzin and Waibel, 1998)	Prosodic information can be combined and integrated with acoustic information within a hidden Markov model (HMM) architecture	Actors read sentences to act different emotion	HMM	Happy, Afraid, Angry, Sad
1996	Acoustic Profiles in Vocal Emotion Expression (Banse and Scherer, 1996)	Vocal Cues	Actors read sentences to act different emotion	Discriminant Analysis (DFA)	Fear, Disgust, Joy, Sadness, Anger
1993	Duration and intonation in emotional speech (Vroomen et al., 1993)	Pitch, Intonation, and Duration of speech	Actors read sentences to act different emotion	1) PSOLA speech synthesis. 2) ANOVA.	Neutral, Joy, Boredom, Anger, Sadness, Fear, Indignation
1971	Constants across cultures in the face and emotion (Ekman and Friesen, 1971)	Facial movement	Subjects view motion picture films.	Subjective judgment	Surprise, Fear, Happiness, Sadness, Anger, Disgust

Appendix C

Forms of Data Collection

C.1 Data Collection Study Information Sheet, Consent Form and Researcher Sheet

User Interaction Model for Adaptive Human-Computer Interaction

Chief Investigator:
Dr. Leo Galway
T: +44 (0) 28 9036 6674
E: l.galway@ulster.ac.uk

Participation Information Sheet

Invitation Paragraph

You are being invited to take part in a research study. Before you decide whether or not to take part, it is important that you understand what the research is for and what you will be asked to do. Please read the following information and do not hesitate to ask any questions about anything that might not be clear to you. Make sure that you are happy before you decide what to do.
Thank you for taking the time to consider this invitation.

Study Purpose

This study aims to populate features from users while they interacting with computer software. The features will be extracted from audio that is detected via microphone, video that is detected via camera, eye gaze that is detected via eye tracker and physiological signal via EEG-headset and wrist sensor. As a participant you will set in front of equipped workstation and wears headset and wrist sensor. The collected data will be used in subsequent study for analysis in order to build a user model that show corresponding affective and cognitive aspects of the user in the context of Human Computer Interaction (HCI). This study focuses on your interaction responses and it doesn't focus in your performance while doing a task.

Data Collected

The data will be collected in this experiment are:

- Recorded video for your face during the tasks, i.e. face gestures.
- Recorded eye gaze behaviour while interacting i.e. gaze movement, blinking rate, and pupil dilation.
- Recorded brain activity using Emotive EPOC headset.
- Task completion time and state.
- You will be asked to rate each task before and after attempting the task using a self-report form.
- You will be asked to indicate your emotional state after each task from a number of snapshots recorded during each task.

Data Protection

All data gathered during this study will be stored on a password protected hard drive along with accordance to Data Protection University Guidelines. Furthermore, only those persons directly involved in the research project will have access to the data. The recorded data (which includes video and audio) in this study will be kept until relevant features been extracted after study. Therefore, once all features have been extracted the video and audio data will be destroyed.

Ulster University

User Interaction Model for Adaptive Human-Computer Interaction

Chief Investigator:
Dr. Leo Galway
T: +44 (0) 28 9036 6674
E: l.galway@ulster.ac.uk

Participation Consent Form

User Interaction Model for Adaptive Human-Computer Interaction

Chief Investigator: Dr Leo Galway

This experiment is part of the PhD project entitled as *"User Modelling for Adaptive Human Computer Interaction"* at the school of Computing and Mathematics in Ulster University. Please refer to *Subject Information Sheet* provided, read it carefully in order to sign the consent form.

- I confirm that I have been given and have read and understood the information sheet for the above study and have asked and received answers to any questions raised
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason and without my rights being affected in any way
- I understand that the researchers will hold all information and data collected securely and in confidence, and I give permission for the researchers to hold relevant personal data
- I consent that video for my face, eye tracking for my gaze, and brain activity will be recorded throughout my participation in this study
- I agree to take part in the above study

Name of Subject	*Signature*	*Date*
Name of person taking consent	*Signature*	*Date*
Anas Samara *Name of researcher*	*Signature*	*Date*

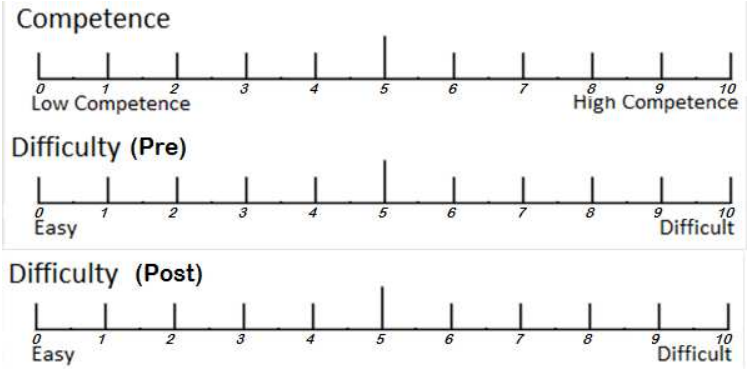
Ulster University

User Modelling for Adaptive Human Computer Interaction

Researcher Sheet

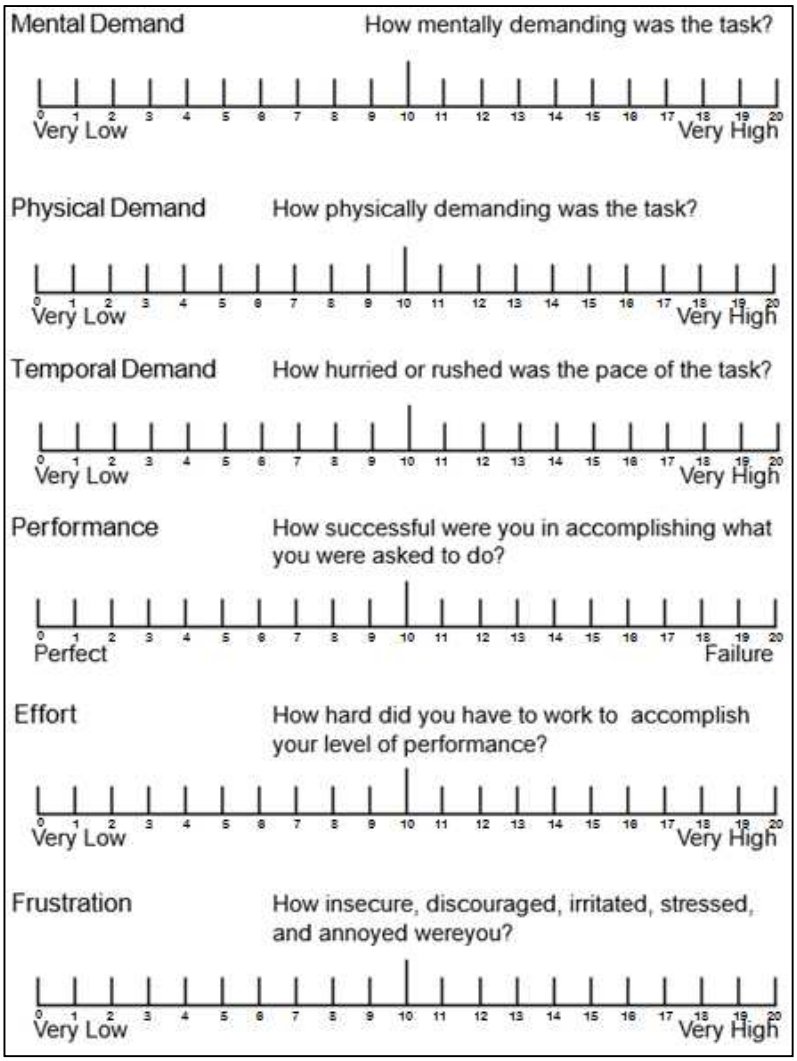
Reference

Task:



SAMS			
	Valence	Arousal	Dominance
Overall			
According to recorded video			
	Valence	Arousal	Dominance
Beginning			
Middle			
End			

Task Completion State	*Not Completed*	*Partially Completed*	*Completed*
Task Completion Time			



Appendix D

HCI-Viewer

D.1 Implementation Details of HCI-Viewer

This section will give technical details about the implementation of HCI-Viewer, the development kit, the exploited application packages, and the logic implemented within the tool.

D.1.1 Frame and Layouts

The Java programming language has been used to implement HCI-Viewer, due to several advantages, including platform independence, fostering rapid development iterations and the abundance of supported libraries and packages. Subsequently, the *Swing* toolkit API has been used to construct the main graphical user interface components. A *JDesktopPane* container was created as the virtual base that contains the various sub-views, where each sub-view is an instance of a *JInternalFrame*. Figure D.1 presents an actual screenshot of the current iteration of the tool.

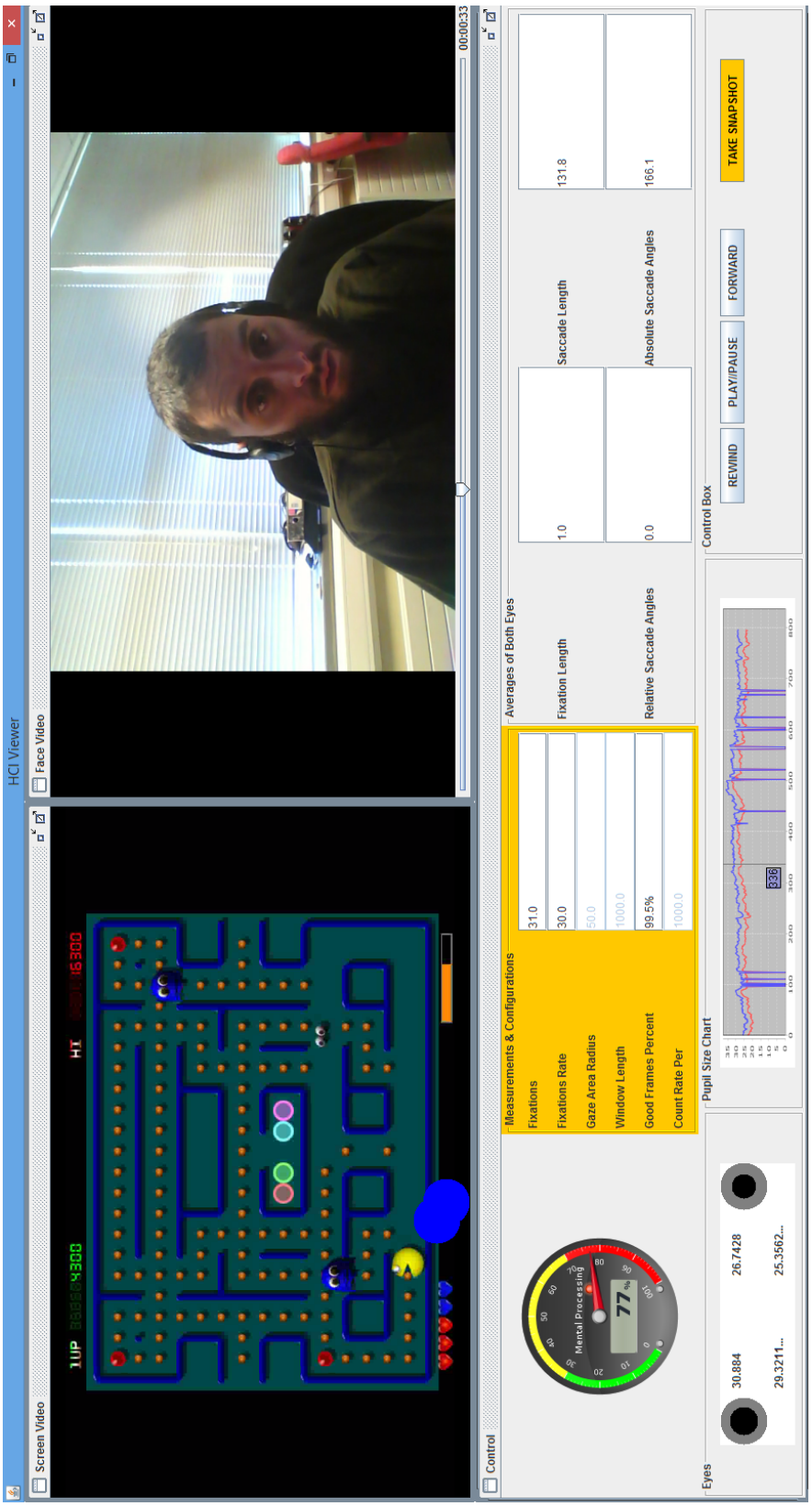


Fig. D.1 Screenshot of the current implemented iteration of the HCI-Viewer tool.

D.1.2 Video Player

Internally, the VLCJ framework has been employed for rendering the video frames of the recordings of both the user's face and the screen. VLCJ is the Java bindings of the popular VLC multimedia player, through which one can programmatically access VLC native libraries using Java code (Caprica Software, 2009). Therefore, it is assumed that VLC must be installed in order to use HCI-Viewer. Subsequently, the buttons in the control box such as *Play/Pause* invoke the corresponding VLC player functions

D.1.3 Data Preprocessing

Since the eye tracker usually represents data as a stream of serialised text, a pre-processing step is needed to make it compatible to be used within the tool. The common format of the data acquired by such devices are either JSON or XML, acquired at a sampling frequencies ranging from 25 - 2000 Hz (Andersson et al., 2010). For instance, we used the Eye-Tribe infra-red-light emitting eye tracker, which provides 30 samples per second as an array of JSON objects. Each JSON object represents a captured sample that contains measurements of both eyes including the pupil size, and the X, Y coordinates of the location of both gazes on the screen. Therefore, the pre-processing process transposes these samples into a map data structure, where the key represents the timestamp of a sample, and the value represents the sample object itself. Consequently, the eye-gaze data and associated measurements are updated and synchronised with the video player using the timestamps.

D.1.4 Fixations Overlay on the Screen Video

As previously discussed, the eye-gaze data is populated in a map data structure, thus, by accessing the corresponding eye-gaze sample object using the video player timestamp, the corresponding pupil size and coordinate information can be obtained. Subsequently, by embedding a *JWindow* container into the interface screen sub-view component, and using a *Graphics2D* object, a filled circle is painted at the specified coordinates, with the radius calculated based on the area of interest value provided in the configuration options sub-view. Furthermore, while the video is playing and updating the current timestamp, the corresponding eye-gaze samples are accessed from the map and the corresponding graphics painted.