Loughborough University Spontaneous Expression Database and Baseline Results for Automatic Emotion Recognition

by

Segun AINA

A doctoral thesis submitted in partial fulfilment of the requirements
for the award of Doctor of Philosophy (PhD), from
Loughborough University

July 2015

Advanced Signal Processing Group,
School of Electronic, Electrical and Systems Engineering,
Loughborough University, Loughborough
Leicestershire, UK, LE11 3TU

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

……. (Signed)

……. (candidate)

I dedicate this thesis to my loving wife, parents, siblings

and all those in the pursuit of knowledge.

Abstract

The study of facial expressions in humans dates back to the 19th century and the study of the emotions that these facial expressions portray dates back even further. It is a natural part of non-verbal communication for humans to pass across messages using facial expressions either consciously or subconsciously, it is also routine for other humans to recognize these facial expressions and understand or deduce the underlying emotions which they represent.

Over two decades ago and following technological advances, particularly in the area of image processing, research began into the use of machines for the recognition of facial expressions from images with the aim of inferring the corresponding emotion. Given a previously unknown test sample, the supervised learning problem is to accurately determine the facial expression class to which the test sample belongs using the knowledge of the known class memberships of each image from a set of training images. The solution to this problem – building an effective classifier to recognize the facial expression is hinged on the availability of representative training data.

To date, much of the research in the area of Facial Expression Recognition (FER) is still based on posed (acted) facial expression databases, which are often exaggerated and therefore not representative of real life affective displays, as such there is a need for more publically accessible spontaneous databases that are well labelled. This thesis therefore reports on the development of the newly collected Loughborough University Spontaneous Expression Database (LUSED); designed to bolster the development of new recognition systems and to provide a benchmark for researchers to compare results with more natural expression classes than most existing databases. To collect the database, an experiment was set up where volunteers were discretely videotaped while they watched a selection of emotion inducing video clips.

The utility of the new LUSED dataset is validated using both traditional and more recent pattern recognition techniques; (1) baseline results are presented using the combination of Principal Component Analysis (PCA), Fisher Linear Discriminant Analysis (FLDA) and their kernel variants Kernel Principal Component Analysis (KPCA), Kernel Fisher Discriminant Analysis (KFDA) with a Nearest Neighbour-based classifier. These results are compared to the performance of an existing natural

expression database – Natural Visible and Infrared Expression (NVIE) database. A scheme for the recognition of encrypted facial expression images is also presented. (2) Benchmark results are presented by combining PCA, FLDA, KPCA and KFDA with a Sparse Representation-based Classifier (SRC). A maximum accuracy of 68% was obtained recognizing five expression classes, which is comparatively better than the known maximum for a natural database; around 70% (from recognizing only three classes) obtained from NVIE.

Table of Contents

Statement of Originality

The contributions of this thesis are mainly on the development and validation of the Loughborough University Spontaneous Expression Database (LUSED), a naturalistic facial expression dataset for the development and application of expression recognition systems.

Chapter Three contains a detailed and up-to-date review of existing natural (spontaneous) databases and catalogues the process of the collection of the new database; from the elicitation of the expressions through the subsequent processing to the ground truth labelling. The following peer reviewed publications support the novelty and contributions of this chapter:

- S. Aina, R. C.-W. Phan, S. M. Naqvi and J. A. Chambers, "Spontaneous Facial Expression Recognition: A New Database and its Application," submitted to IEEE Transactions on Affective Computing, 2015.

- S. Aina, M. Zhou, J. A. Chambers, and R. C.-W. Phan, "A New Spontaneous Expression Database and a Study of Classification-based Expression Analysis Methods," *Proc. 22nd Eur. Signal Process. Conf.*, pp. 2505–2509, 2014.

Chapter Four includes a review of historic methods for FER followed by the evaluation of the new dataset to establish baseline results. Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Fisher's Linear Discriminant Analysis (FLDA) and Kernel Fisher's Discriminant Analysis (KFDA) are used for feature extraction prior to applying a Nearest Neighbour classifier. The results are compared to the performance of an existing database. Additionally, a novel model for performing natural FER in the Encrypted Domain (ED) based on FLDA is also proposed. The following peer reviewed publications support the novelty and contributions of this chapter:

- S. Aina, R. C.-W. Phan, S. M. Naqvi and J. A. Chambers, "Spontaneous Facial Expression Recognition: A New Database and its Application," submitted to IEEE Transactions on Affective Computing, 2015.

- S. Aina, Y. Rahulamathavan, R. C.-W. Phan, and J. A. Chambers, "Spontaneous Expression Classification in the Encrypted Domain," *IMA 2012, Birmingham arXiv preprint arXiv:1403.3602*, 2012

Chapter Five contains a review of sparse representation based methods for FER followed by the evaluation of the new dataset to establish benchmark results. PCA, FLDA, KPCA and KFDA are applied together with a Sparse Representation-based Classifier (SRC). The following peer reviewed publications support the novelty and contributions of this chapter:

- S. Aina, R. C.-W. Phan, S. M. Naqvi and J. A. Chambers, "Spontaneous Facial Expression Recognition: A New Database and its Application," submitted to IEEE Transactions on Affective Computing, 2015.

- S. Aina, R. C.-W. Phan, and J. A. Chambers, "Robust Spontaneous Facial Expression Recognition using Sparse Representation," in *IET Intelligent Signal Processing Conference (ISP 2013)*, 2013, pp. 7.10–7.14

Acknowledgement

I owe a huge debt of gratitude to my supervisor Prof. J. A. Chambers for the time and effort committed to making my dream possible. I appreciate the patience taken to understand my strengths and my weaknesses and help me apply myself better. I attribute my progress to his encouragement when I was lacking motivation, his sternness when I was being lazy and his fatherly words of wisdom when I was lacking in knowledge. He taught me much more than he knows and has given me a level to aspire to, I will be forever grateful, God bless you. I will also like to thank my second supervisor Prof. R. Phan for the time and effort invested in me. His critical thinking always kept me on my toes and I am grateful he always challenged me to do better.

I am grateful to Loughborough University's School of Electronic, Electrical and Systems Engineering for the facilities to complete my research. I want to thank all the staff and students particularly those who participated in my experiments and those who served on the judging panels. A big thank you to my colleagues turned friends for the encouragement and support – Chinwe, Ozak, Peter and Funmi. To my friends in the HSN lab: Kostas, Fran, Abdul, Gines, and Will, thank you for your hospitality. Thank you to Mayowa Aina a brother and a friend indeed, also to Dr. Oluwaseun Ojerinde, a brother, friend and colleague who understands the struggle, your time here made it easier for me, God bless you both. To all my friends both here and at home that have monitored my progress and been there for me, may God be with you too.

I owe another huge debt of gratitude to my family, without whom I could not have come this far; my wife Motunrayo Aina – the pages of this thesis will not be enough to thank you for all your support, may God grant all the desires of your heart. My father, and now senior colleague Prof. Oluremi Aina – I cannot thank you enough for the intellectual, moral and financial support, I want to be like you when I grow up! A big thank you to my mother Mrs. Marion Ebun Aina, if your wishes could type, I would have completed this thesis in days! To my siblings, their spouses, my nephews and nieces; The Ainas II, The Fanirans, The Agenes, The Obomighies and Funke I wont be here without your invaluable support, I love you all, thank you!

Most of all, I give thanks to God almighty for the grace and mercy bestowed upon me, the motivation and knowledge to see this through.

List of Acronyms

Acronym	Definition
1D	One Dimension
2D	Two Dimensional
3D	Three Dimensional
4D	Four Dimensional
AAI	Adult Attachment Interview
AAM	Active Appearance Model
AM-FED	Affectiva Mit-Facial Expression Database
AN	Angry
ANN	Artificial Neural Network
ASPG	Advanced Signal Processing Group
AT&T	American Telephone & Telegraph
AU	Action Unit
BP4D	Binghamton-Pittsburg 4Dimension
CK	Cohn Kanade
CPU	Central Processing Unit
CS	Compressed Sensing
DI	Disgust
DISFA	Denver Intensity of Spontaneous Facial Actions
ECG	Electrocardiogram

Acronym	Definition
ED	Encrypted Domain
EOG	Electrooculogram
FACS	Facial Action Coding System
FE	Fear
FED	Facial Expression Database
FER	Facial Expression Recognition
FLDA	Fisher's Linear Discriminant Analysis
FM	Facial position and expression Mouse system
HA	Happiness
HCI	Human Computer Interaction
HD	High Definition
ICA	Independent Component Analysis
JAFFE	Japanese Female Facial Expression
KFDA	Kernel Fisher Discriminant Analysis
KNN	K – Nearest Neighbour
KPCA	Kernel Principal Component Analysis
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LOO	Leave One Out
LUSED	Loughborough University Spontaneous Expression Database
MDA	Multiple Discriminant Analysis

Acronym	Definition
MLE	Maximum Likelihood Estimation
MMI	Maja Michel Ioannis
MPI	Max Planck Institute
MUG	Multimedia Understanding Group
NE	Neutral
NN	Nearest Neighbour
NP	Non-deterministic Polynomial-time
NVIE	Natural Visible and Infrared Facial Expression Database
ORL	Olivetti Research Laboratory
OS	Operating System
PC	Principal Component
PCA	Principal Component Analysis
PD	Plain Domain
pdf	*probability density function*
PR	Pattern Recognition
PTSD	Post-Traumatic Stress Disorder
RAM	Random Access Memory
RDT	Randomized Decision Tree
SA	Sadness
SC	Sparse Coding
SF	Scaling Factor

Acronym	Definition
SRC	Sparse Representation Classifier
SSS	Small Sample Size
SU	Surprise
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TBM	Transferable Belief Model
ULDA	Uncorrelated Linear Discriminant Analysis
UNBC	University of Northern British Columbia
USTC	University of Science and Technology, China
UT	University of Texas
VAM	Vera am Mittag
VLSI	Very-Large-Scale-Integration

List of Symbols

Generally, scalar variables are denoted by plain lower-case letters, (e.g., x), vectors by bold-face lower case letters, (e.g., $\boldsymbol{x}$), matrices and groups of matrices are denoted by bold-face upper-case letters (e.g., $\boldsymbol{X}$). Below are some of the frequently used symbols and notations:

ω	State of nature
$P(\cdot)$	A priori probability
$p\ (\cdot \mid \cdot)$	Class-conditional probability density function (Likelihood)
$\boldsymbol{\mu}$	Mean vector
$\boldsymbol{\Sigma}$	Covariance matrix
$\mathcal{D}$	Random variable samples
$\boldsymbol{\theta}$	Unknown parameter vector
$\boldsymbol{W}$	Optimal projection matrix
$\lvert\cdot\rvert$	Determinant
Φ	Nonlinear map
$(\cdot)^T$	Transpose operation
λ	Eigen-values
$\lVert\cdot\rVert_2$	Euclidean norm
$[\![\cdot]\!]$	Encryption operation
$(\cdot)^{-1}$	Inverse operation
$\lVert\cdot\rVert_1$	ℓ^1- norm

List of Figures

List of Tables

Chapter One

Introduction

1.1 Introduction

Research into the facial expressions of humans generally can be traced as far back as the late 19th century. In 1872, Charles Darwin in his book [1] documented a series of captivating observations relating to facial expressions of humans and animals including dogs, cats and horses. Darwin offered explanations for the workings of the observed expressions and provides stimulating analysis of the underlying emotions. He also posed and discussed interesting philosophical questions, for example;

1. Do humans *learn* which expressions to make when experiencing an emotion such as anger or sadness or is such *knowledge* innate?

2. Are expressions like words that are different in every language or are they the same for all people independent of upbringing, culture or language?

Darwin argued that expressions of emotion are instinctive and a product of evolution hence universal. Even though the face has been the focus of research over the years, expressions generally relate to an exhibition of emotions and typically involve not only the face but also voice and posture or body language albeit less significantly.

Early psychological research [2] suggests that up to 55% of information is passed across using facial expressions, 38% through auxiliary language such as speech

rhythm, tone and only 7% of total information is passed by language underlining the importance of understanding this *universal facial language*.

The problem of understanding this *universal facial language* was more clearly defined by the pioneering work of Ekman and Friesen [3, 4] in which they defined *six basic* expression categories for humans: anger, disgust, fear, happiness, sadness and surprise which are discussed in more detail in Chapter Three. In defining these *basic* expression categories over three decades ago, Ekman laid the foundation for the subsisting research interest in being able to recognize these categories automatically using a machine – a task that is trivial and intuitive for most humans.

There are parallel research streams and sub-streams on the issue of (facial) expressions and the emotions they convey – in addition to the psychological and technological angles, there is also the neurological perspective. There has been debate about whether affect precedes cognition or vice-versa and what role emotions play in the decision making process [5]. There is also debate about how well people understand their own emotions – automatic nervous system responses such as facial expressions are believed to capture subtle discrepancies in emotion that are otherwise missed when participants self-report [6]. Therefore, although mentioned earlier that the task of recognizing facial expressions is mostly trivial for humans to perform, potentially, there are salient aspects of this task that could be enhanced using technology – machine learning.

The above, in part forms the general *justification* for automatic Facial Expression Recognition (FER) without alluding to the numerous applications and their benefits. Advances in image processing and pattern recognition have facilitated FER research, which dates back to the early 90s (the phrase abbreviated – FER will loosely be used to refer to automatic FER in the course of this thesis). Mase [7] in 1991 proposed a new theory that used optical flow-based techniques for FER, more or less launching the trend of using computer technology for *feature extraction* and *classification* – the primary tasks of a recognition algorithm.

In the extraction of features for the purpose of FER, there are two main approaches: the geometric feature-based methods and the appearance-based methods [8]. Geometric feature extraction includes information about the shape and location of parts of the face such as the eyes, brows, nose and mouth. Appearance-based feature

extraction on the other hand works on the facial images directly to represent facial textures such as wrinkles, bulges and furrows. It is worth mentioning that in building a recognition system, these features are to be extracted from a set of generalized examples. With the above in mind, it can be said that a high level look at the task of automatic FER reveals two main components; the database and the algorithm represented in the block diagram below (Figure 1-1).

In terms of research, the relationship between these two aspects (database and algorithm) is a symbiotic cycle in the sense that advances in the development of new datasets promotes the development of new algorithms which goes round to promote new dataset design – a process which is commonly guided/driven by the demands of industry applications. In this thesis, both aspects will be addressed with the primary focus being on database design, and earlier algorithmic work within the Advanced Signal Processing Research Group (ASPG), School of Electronic, Electrical and Systems Engineering in Loughborough University will provide a foundation [9, 10].

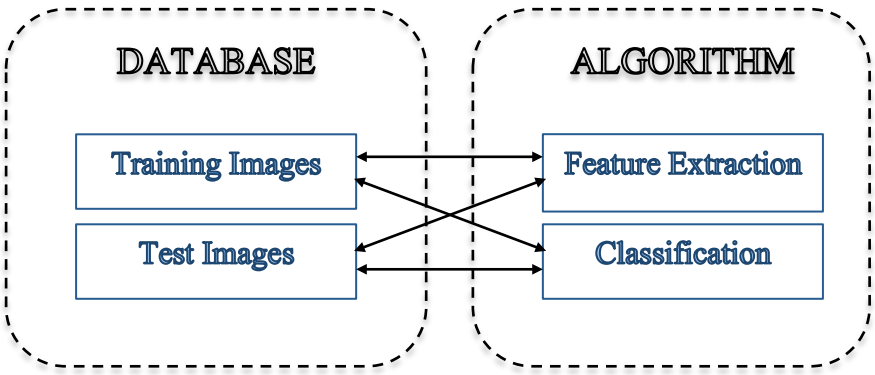


Figure 1–1: Block diagram showing the typical relationship between the two main components of automatic FER.

Consider the image in Figure 1-2 below and the possible contradiction between facial expressions and spoken words. Assume a marketing company is interested in the affective responses of consumers to a publicly displayed digital advertisement campaign (perhaps a digital bill board at a busy bus stop). It may not be practical to interview people effectively and get accurate responses that are representative either due to the volume of people or the likely inconsistent responses that can be expected from people rushing through daily life activities. In this hypothetical scenario, an automatic facial expression system will provide marketing experts a wealth of information. Similarly, there are other sectors that form a growing track of applications for FER from vehicular design through psychology to medical applications, which will be discussed in more detail in Chapter Two.

Figure 1–2: What if facial expressions say something different from words?

1.2 Problem Statement

The algorithmic problem of supervised facial expression recognition is to use labelled training samples from k distinct object classes within a dataset to correctly determine the class to which a new test vector sample $\boldsymbol{y}$ belongs, such as an image stored in a column vector form. The formulation of the algorithmic problem and the efficacy of any proposed solution rely on the availability of a suitable labelled dataset

that adequately describes the object classes. While it must be said that researchers in the general area of image processing sometimes develop datasets for the design of algorithms, literature has shown that the more convincing contributions are based not solely on purpose built datasets but also on independent datasets.

The above underpins another problem – designing a representative dataset on which new and existing algorithms can be evaluated. Facial expression recognition as a research area does not have a sufficient number of good quality training/testing databases that are freely available. This is more so when compared to other areas of image processing such as facial recognition, gender recognition and age estimation. Furthermore, analyses of the available databases reveal limitations, which are detailed and addressed later (Chapter Three) in this thesis.

Informally, the importance of FER is illustrated in Figure 1-3 below where facial expressions can be considered as masks under which corresponding emotions can be found.

Figure 1–3: The problem of FER can be considered as a means of revealing the emotions behind the masks of facial expressions (Image from [11]).

1.3 Aims and Objectives

The primary *aim* of this thesis is to establish and evaluate a labelled naturalistic facial expression database that is freely available, *culturally diverse* and gender balanced as representative data for training and testing facial expression algorithms with a view to bolster the research interest in and applicability of automatic facial expression recognition.

- *Objective One*: To design an experiment that will facilitate the spontaneous display of natural facial expressions that can be captured discretely. The goal is to do this in a semi-controlled environment such that the variability from subject to subject can be reduced without compromising the spontaneity of the expressions being elicited.

- *Objective Two*: To recruit a diverse and balanced set of participants without incentives.

- *Objective Three*: To establish the ground truth labels for the elicited facial expressions.

Chapter Three reports on objectives *one*, *two* and *three*, where newly designed stimuli videos are used to establish a database of spontaneous facial expressions, which contains images from five classes out of the six basic expression classes. Each image in the database is also labelled with the class to which it belongs.

- *Objective Four*: To review and analyse historic methods for facial expression recognition.

- *Objective Five*: To establish baseline results for the recognition of expressions from the newly created database using historic methods.

- *Objective Six*: To perform comparative analysis of the baseline results obtained from the new database and other existing databases.

Chapter Four reports on objectives *four*, *five* and *six*; dimensionality reduction methods are used to extract features that are then classified using a basic distance measure.

- *Objective Seven*: To review and analyse up-to-date techniques for facial expression recognition.

- *Objective Eight*: To establish benchmark results for the robust recognition of expressions from the newly created database using up-to-date techniques.

Chapter Five reports on objectives *seven* and *eight*; robust sparse representation techniques are combined with various feature extractors before the use of a sparse classifier.

1.4 Thesis Structure

This thesis is presented in six chapters, which include this introductory chapter as well as an overall conclusion chapter. Each of the following chapters has an introduction and a brief summary and references are presented at the end of each chapter. The remainder of this thesis is organized as follows:

- *Chapter Two* contains an introduction to pattern classification and pattern recognition systems. An example is used to explain the steps involved in a typical pattern recognition system as well as possible challenges. Fundamental statistical approaches to solving ideal pattern recognition problems are discussed – Bayesian decision theory and maximum-likelihood parameter estimation. Various applications of pattern recognition systems are also discussed as well as a review of literature on some of these applications.

- *Chapter Three* firstly presents a background of facial expressions including the differences between posed and natural facial expressions. A review of existing natural databases is followed by the physical and physiological characteristics of the six basic expressions of interest. The experimental setup for the collection of the new Loughborough University Spontaneous Expression Database (LUSED) is detailed including experimental design, the recruitment of participants,

selection and order of stimuli video, statement of ethics, segmentation of captured expression data and an analysis of the participant demography. Lastly, the tasks relating to the post processing of the images are reported.

- *Chapter Four* is firstly a background review of baseline evaluation methods followed by the mathematical theory of the baseline algorithms – Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Fisher's Linear Discriminant Analysis (FLDA), Kernel Fisher's Discriminant Analysis (KFDA) and the Nearest Neighbour (NN) classifier. Building on the hypothetical application example above (marketing company), in order to mitigate the arising privacy considerations, a system for performing FER in the encrypted domain using FLDA is presented. Details of the simulation setup relating to the database validation are reported prior to results and a comparative analysis.

- *Chapter Five* introduces a sparse representation-based system combined with dimensionality reduction schemes as a technique for facial expression recognition. A background review is given followed by the theory /mathematics of sparse representation. The robustness of this technique to noise and corrupt/occluded images is discussed. The sparse classifier is described in addition to steps taken to compose an underdetermined dictionary from the training images. The simulation setup is described followed by an analysis of evaluation results.

- *Chapter Six* concludes the thesis and offers suggestions for possible future research.

In addition to the above is an Appendix, which contains supplementary information relating to the experimental setup.

References

[1] C. Darwin, *The expression of the emotions in man and animals*. John Murray, London, 1872.

[2] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, pp. 53–56, 1968.

[3] P. Ekman, "Universal and cultural differences in facial expressions of emotion," *Nebraska Symp. Motiv.*, 1971.

[4] P. Ekman and W. Friesen, "Facial action coding system: a technique fir the measurement of facial movements," *Consult. Psychol.*, vol. 2, 1978.

[5] R. B. Zajonc, "Feeling and thinking, preferences need no inferences," *Am. Psychol.*, vol. 35, no. 2, pp. 151–175, 1980.

[6] K. A. Leitch, S. E. Duncan, S. O'Keefe, R. Rudd, and D. L. Gallagher, "Characterizing consumer emotional response to sweeteners using an emotion terminology questionnaire and facial expression analysis," *Food Res. Int.*, 2015.

[7] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. 74, no. 10, pp. 3474–3483, 1991.

[8] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," *Handb. Face Recognit.*, pp. 487–519, 2011.

[9] M. Yu, A. Rhuma, S. Naqvi, L. Wang, and J. Chambers, "Posture recognition based fall detection system for monitoring an elderly person in a smart home environment," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1–13, 2012.

[10] M. Yu, S. M. Naqvi, and J. Chambers, "A robust fall detection system for the elderly in a smart room," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 1666–1669, 2010.

[11] "IndiaToday" [Online]. Available: http://indiatoday.intoday.in/story/a-computer-to-lip-read-and-decode-emotions/1/217015.html. [Accessed: 30-Jun-2015].

Chapter Two

Literature Review

2.1 Introduction

Over the years, researchers have sought to design and develop machine systems that are able to automatically recognize patterns based on data – machine learning. Duda et al. [1] defined pattern recognition as "the act of taking in raw data and making an action based on the category". This chapter will broadly review pattern recognition systems in the context of facial expression recognition as a precursor to the contributions presented in Chapters Three, Four and Five.

Before progressing and for clarity, it is prudent to explicitly define/distinguish the various terminologies relating to pattern recognition that will be referenced in this and subsequent chapters although as is the case in the literature, some of the following terms are used interchangeably;

1. *Pattern* – is essentially an object that is vaguely defined and can be given a name; aptly described by Watanabe as "the opposite of chaos" [2]. Practically, a pattern can be represented as a cluster of data points in a given space.

2. *Recognition* – in a broad sense implies the act of associating a classification with a label [3].

3. *Class* – refers to one of the groups by which the above *pattern* can be characterized i.e. the different clusters from (1) above.

4. *Pattern Recognition* – can alternatively be defined as "the scientific discipline of machine learning (or artificial intelligence) that aims at classifying data (patterns) into a number of categories or classes" [4].

5. *Pattern Classification* – is a phrase often loosely used interchangeably with *pattern recognition, learning* and associated with the definitions above; however strictly speaking, it can be said that while pattern recognition establishes the existence of discernable patterns in data, pattern classification goes the step further to actually establish the different classes and the elements of the data contained in each class.

6. *Pattern Learning* – refers more to evolving iterative systems for pattern recognition and is characteristically linked with *unsupervised learning* (described below).

Less formally, a pattern is characterized by the shared constant among the multiple instances of an entity. For example, commonality in all images of fish species defines the fish pattern; the commonality in images of sea bass defines the sea bass pattern. Contextually, commonality in all human facial images defines the face pattern and the commonality in images of Jane Doe defines the Jane Doe face pattern. In fact, a pattern could be a human face, speech signal, barcode, fingerprint or a handwritten cursive word. Typically, respective patterns may be categorized into a group based on their common properties where the resulting group is also a pattern and is often referred to as a pattern class.

In this chapter, the area of pattern recognition will be described before it is related to the more delimited problem of pattern classification; specifically methods for generalization – how new observations are subjected to derived decision rules. The *human* element will also be discussed; (1) touching on the symbiotic relationship between human learning and pattern learning and (2) reviewing existing applications of pattern recognition systems in different facets of human life.

There are several general approaches to Pattern Recognition (PR) and the different algorithms and methods used in solving the PR problem can broadly be categorized into one of these *approaches*. The applications of the approaches to PR have evolved over time and subsequent sections will discuss the approaches in varying detail based on relevance.

The remainder of this chapter will be organized as follows; Section 2.2 contains the steps involved in a typical pattern recognition system and the challenges associated with such systems is touched on in Section 2.3. Fundamental approaches to pattern recognition are described in Section 2.4. Models for real-world applications as well as actual implementations of facial expression recognition systems are reviewed in Section 2.5 and a summary of the chapter is offered in Section 2.6.

2.2 Pattern Recognition System

Broadly speaking, pattern recognition techniques can be divided into two high-level type groups – supervised learning and unsupervised learning;

2.2.1 Supervised Learning

Supervised learning refers to the training of an algorithm or system using explicitly labelled and class defined examples as inputs (training samples). In supervised learning, the algorithm is trained by generalizing the training samples, the classes to which they belong and the patterns contained therein with the aim of being able to recognize the class to which a new test sample belongs. It is the most common technique for training neural networks and decision trees – methods that are highly dependent on the information given by the predetermined classifications (ground truths).

2.2.2 Unsupervised Learning

Unsupervised leaning on the other hand refers to algorithms that generally cluster samples into a varying number of *natural* classes based on the observed patterns within the data. Some algorithms in this category will accept as input – a hypothesis of the number of clusters present within the data. Given the same dataset, different clustering algorithms will often arrive at a different number of clusters (classes). The primary

difference is that in unsupervised learning, there is no prior knowledge of the number of classes and the class membership of each element; consequently the problem of unsupervised learning can be said to be more difficult.

In order to put the schemes to be discussed into perspective, consider the mock example by Duda et al. [1] that is frequently used and was alluded to in Section 2.1. For this example, suppose that a fish packing plant requires the automatic sorting of incoming fish on a conveyor belt into their respective species – either salmon or sea bass. To achieve this, images of the fish on the conveyor belt were taken as illustrated in Figure 2-1 below, subsequently features such as length, width, and fin structure were observed in a bid to build the ideal classifier. This example will be used in detail to explain recognition systems in general terms; the approaches and the difficulties therein, before being contextualized to the specific problem of facial expression recognition.

Figure 2–1: Image taken from [1] showing the setup of the hypothetical example with fish on a conveyor belt and a camera for sensing.

A pattern recognition system simply refers to the end-to-end operations required for the practical application of a machine able to automatically solve the classification problem taking into account real-world limitations such as noise. These turnkey operations which are largely irrespective of approach represent the process flow from obtaining the data, to sorting, to the highlighting of distinguishing features, culminating in the decision and resulting action. The operations that make up a typical pattern recognition system are encapsulated in the block/flow diagram [1] in Figure 2-2 and explained in further detail in the subsections below.

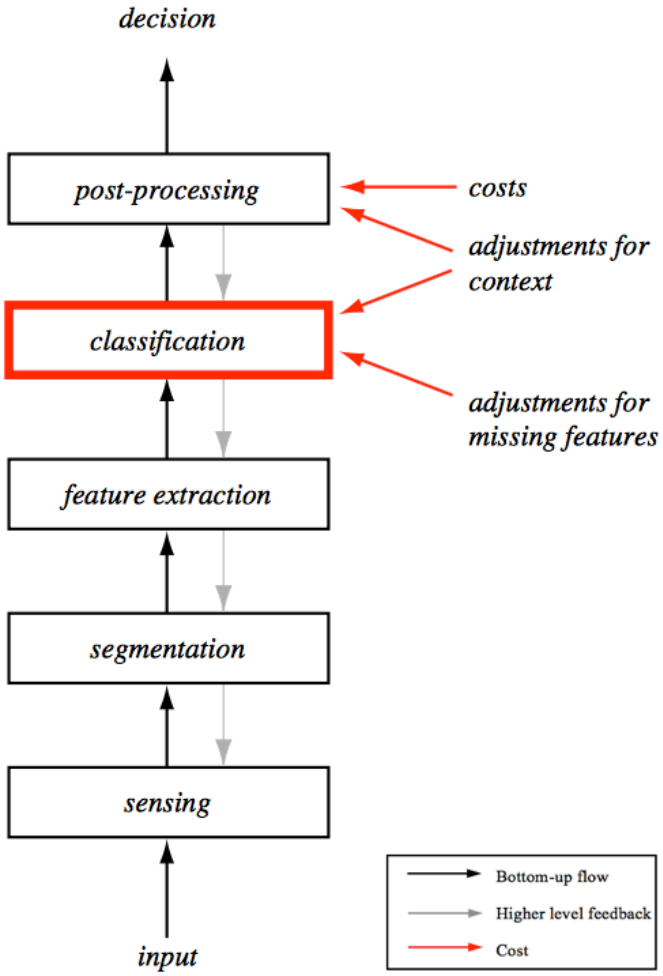


Figure 2–2: Block/flow diagram showing general partitions of the operations involved in a typical pattern recognition system from bottom up including a feedback loop taken from [1].

2.2.3 Input Data Sensing

A microphone array, Electrocardiogram (ECG) or a similar transducer device such as a camera in this example will typically generate the inputs, which a PR system then needs to read prior to processing. It is worth noting that the type/properties of the transducer that generates the data as well as the system's ability to read such data contribute to the difficulty of the problem. In the context of this example, the optical camera is the input device and the limitations could be environmental – a variance in lighting of the captured images, the position of the fish or technical – the specification of the camera. The black rectangle in Figure 2-3 below illustrates the sensing of the fish in the image obtained from the optical camera which will then go on to be analysed as described in subsequent sections.

Figure 2–3: Adapted from Figure 2-1 illustrating the sensing of all the fish on the conveyor belt by a PR system.

2.2.4 Segmentation

Segmentation is a significant problem of pattern classification [5] and is most apparent in problems such as speech segmentation [6]. An implicit assumption of the above example is that ideal conditions exist where each fish on the conveyor belt is in isolation and devoid of occlusion from parts of another fish. However in real-life, the fish may be partly overlapping, as is the case in the illustration of the segmented fish in Figure 2-4. After sensing, the task of the system will then be to identify the perimeters of each fish; subtracting the background or other unrelated objects in the image.

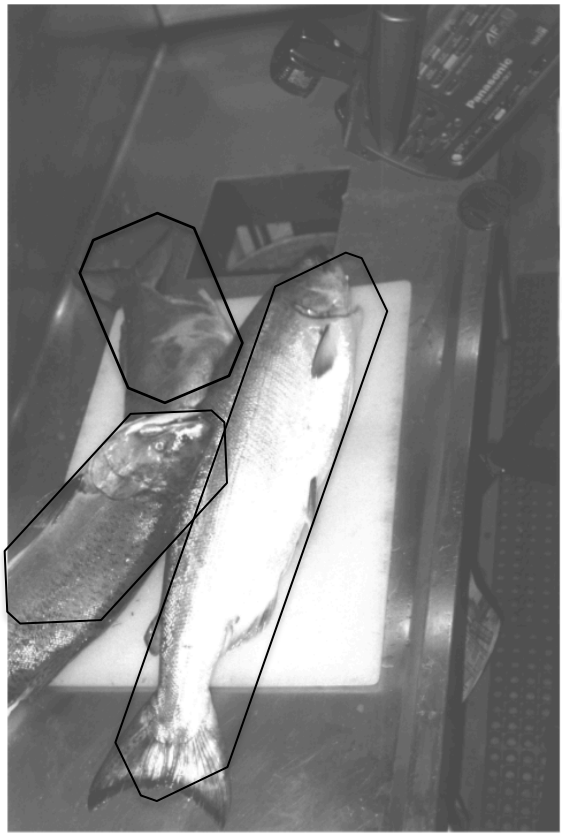


Figure 2–4: Adapted from Figure 2-1 illustrating the segmentation of the sensed fish.

2.2.5 Feature Extraction

This involves selecting measureable quantities that increases the separability of object classes. The aim is to represent the data – for example, the fish using properties such as texture and shape which will have similar values for fish that are of the same species and also be sufficiently different from the property values of different species of fish. In other words, the properties are highlighted (extracted) so that when presented with a new fish for instance, it becomes less difficult for a classifier to identify if it is a salmon or a sea bass. The arrows in Figure 2-5 below demonstrate the use of length as a distinguishing feature of the segmented fish. The process of feature extraction is in effect *training* the recognition system by allowing the property analysis of multiple examples of each of the two fish species. It is essential however that the selected features are not affected by changing variables such as the position and orientation of the fish on the conveyor belt. For example, in the case FER, the extracted features could include the position and shape of the lips.

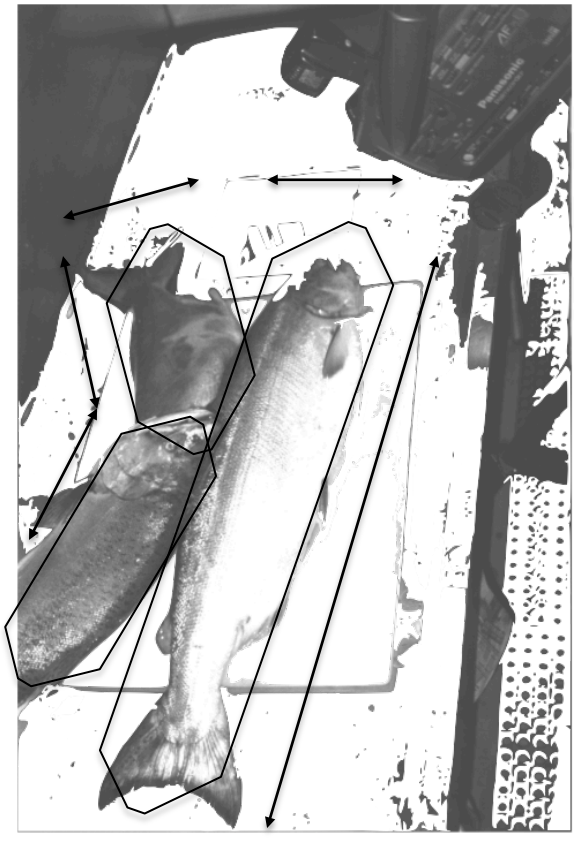


Figure 2–5: Adapted from Figure 2-1 illustrating the extraction of features for the segmented fish.

2.2.6 Classification

Once the features of the *training* objects and the *test* objects have been highlighted as above, the *classifier* will typically then compare these features (training/test) to each other with the aim of assigning a test object to a category/class. In the context of this example, given the texture and shape details of a new test fish, the classifier aims to compare those features to those of the training samples of salmon and sea bass in order to determine the species of the test fish. This is illustrated in Figure 2-6 where the different coloured outlines represent different species of fish. The potency required from a classifier can be said to be inversely proportional to the potency required from the feature extractor; the classifier does not need to be exceedingly sophisticated if the feature extraction method is effective. Conversely, if the feature extractor is deficient, the system will benefit from a more sophisticated classifier. Some of the classifiers that will be discussed in subsequent chapters include Nearest Neighbour (NN) and Support Representation Classifier (SRC).

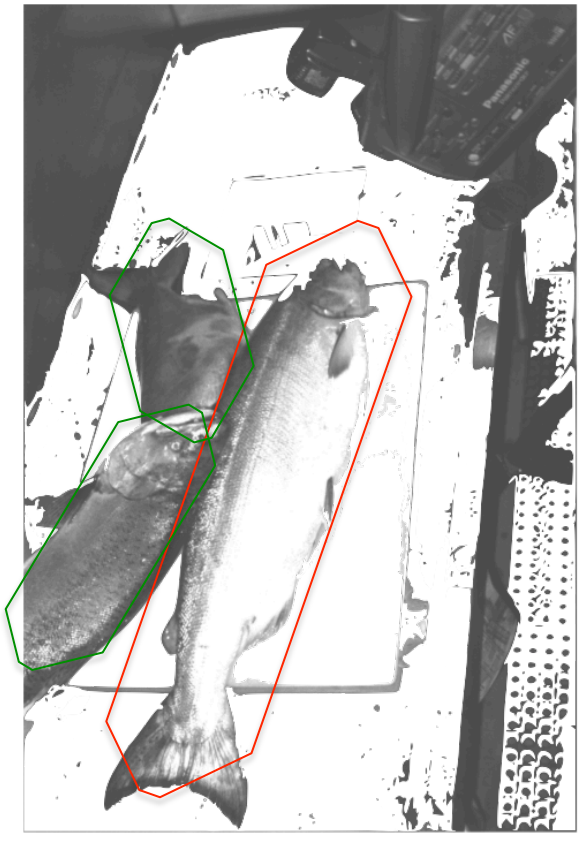


Figure 2–6: Adapted from Figure 2-1 illustrating the recognition of fish (coloured highlights) using extracted features.

2.2.7 Post-Processing

In completion of the system, the post-processing component primarily uses the decision from the classification step above to perform some action such as sending the salmon to a pink bucket and sending the sea bass to a grey bucket. Additionally, post-processing could involve slightly more complex tasks such as using parameters such as context and error cost to act on the decisions or evaluating the results of multiple classifiers with conflicting decisions, as reflected by the red and grey arrows in Figure 2-1 where feedback from each stage can influence other stages in subsequent iterations of the process flow.

A basic measure for the performance of a classifier is the *error rate (%)* – the percentage of test samples wrongly recognized. Intuitively, a minimum-error-rate classifier is typically sought, however, it maybe better to seek a classifier with the lowest *risk*. The *risk* is termed as actions that minimize the total expected *cost*; hence a post processor could use knowledge of *costs* to assess the *risk* of a given classifier.

2.3 Challenges of a Pattern Recognition System

Training the system is an important element of the recognition process that can come with challenges when being applied practically. In the majority of systems, training occurs during the feature extraction step and the most integral element of the training process is the training samples themselves. As such, selecting the training samples that best generalize the variation in each class is a challenge albeit essential to recognition performance.

Noise is commonplace in practical applications of a PR system and has an effect on class variability. Noise in this context can broadly be defined as "any property of the sensed pattern which is not due to the true underlying model but instead to randomness in the world or the sensors"[1].

Effective/efficient representation of objects can be challenging; typically with images, the best samples for recognition are those that maximize redundancy – large raw images, however, such images are not computationally efficient in the recognition process. On the other hand, feature extraction schemes that reduce dimensionality sometimes result in the loss of data relevant to separability. This challenge can be

exasperated by *the problem of overtraining* where an overly complicated system can learn details not important to the recognition task at the cost of not being able to learn common details that actually define the pattern – essentially memorizing the training set rather than learning the general pattern contained in each class.

2.4 Approaches to Pattern Recognition

Pattern recognition problems vary in difficulty bordering on the amount of information about a given problem. Consequently, the approaches taken to solve the problem have varied over time both in terms of technique and complexity. Before exploring the various approaches applied in the literature, it is important to understand the basic approach to pattern recognition given an *ideal* problem.

2.4.1 Bayesian Decision Theory

While the general aim of PR is to determine the pattern to which new data belongs, one of the basic tasks in achieving this aim relates to *how* the decision is taken to assign such data to a given class. Decision theory is the study of this question (*how* the decision is taken) with a view to minimizing the *cost* of the decision.

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. It is based on quantifying the trade-offs between various classification decisions using probability and the costs that accompany such decisions based on the assumption that the decision problem is posed in probabilistic terms and that all the relevant probability values are known [1].

An *ideal* pattern recognition problem will be formulated with a large amount of information about the model such as the probability structure underlying the categories, class labels and so on. While this *ideal* problem is rarely the case in practice, the Bayes classifier has served as an optimal basis with which literature has compared other classifiers.

In Bayes decision theory, it is assumed that the decision problem is posed in probabilistic terms and all the relevant probability values are known. To better explain the concept of decision theory, recall the hypothetical example from Section 2.2 – separating salmon from sea bass, consider the probabilistic variable ω, which denotes

the *state of nature* of the fish arriving on the conveyor belt such that $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon. Based on the number of salmon and sea bass present in the haul, there is some *prior probability* given by $P(\omega_1)$ that the next fish is sea bass and also some *prior probability* that the next fish is salmon given by $P(\omega_2)$ both of which sum up to one and is based on prior knowledge of the *likelihood* of one appearing next over the other.

A classifier can be built on the combination of the above prior probability and a class specific feature pattern (such as lightness) x that is dependent on the *state of nature* of the arriving fish expressed as the *class-conditional probability density* function $p(x|\omega)$. As such, the difference in the lightness between sea bass and salmon can be expressed in the difference between $p(x|\omega_1)$ and $p(x|\omega_2)$. This is graphically illustrated in Figure 2-7 below.

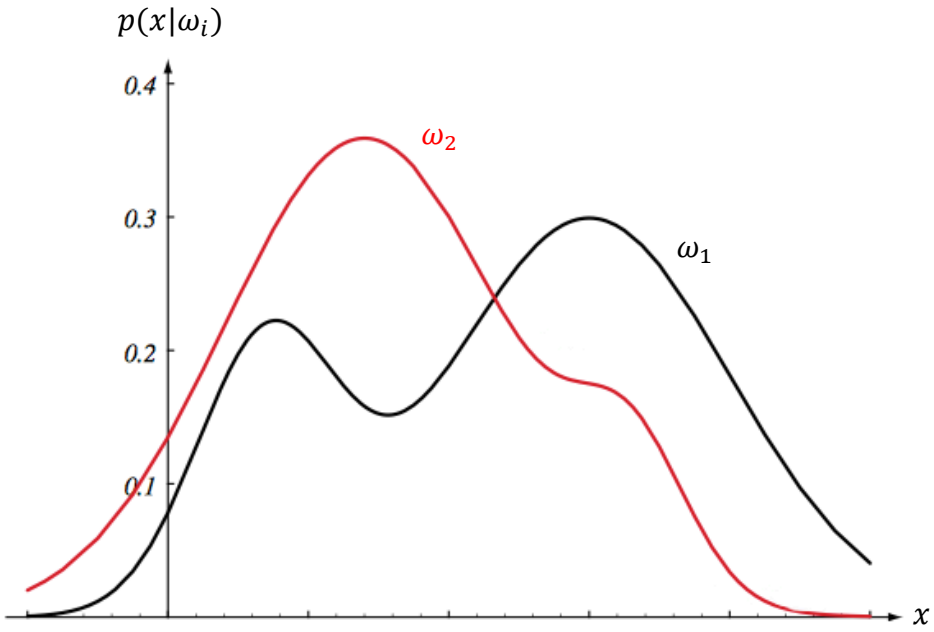


Figure 2–7: Class-conditional probability function for hypothetical example where the two curves can represent the difference using probability density of the lightness of populations x of two types of fish given the pattern is in category ω_i taken from [1].

From the above, x represents the lightness of a fish. Let both prior probabilities be $P\ (\omega_j)$ and the conditional densities be $p\ (x|\omega_j)$ for $j = 1,2$. The probability of the *state of nature* being ω_j given the measurement of x can be derived using Bayes formula, which converts the *prior* probability to the *posterior* probability and is given by:

$$posterior = \frac{likelihood \times prior}{evidence} \tag{2.1}$$

More formally,

$$P\ (\omega_j|x) = \frac{p\ (x|\omega_j)\ P\ (\omega_j)}{p\ (x)} \tag{2.2}$$

where in this two-category case;

$$p(x) = \sum_{j=1}^{2} p\ (x|\omega_j)\ P\ (\omega_j) \tag{2.3}$$

$p\ (x|\omega_j)$ is the *likelihood* of ω_j with respect to x. As such, the true class is ideally more *likely* to be the category ω_j corresponding to a large conditional density $p\ (x|\omega_j)$. The variation $P\ (\omega_j|x)$ with respect to x can be observed in Figure 2-8 for the instances of $P\ (\omega_1) = {}^{2}/_{3}$ and $P\ (\omega_2) = {}^{1}/_{3}$.

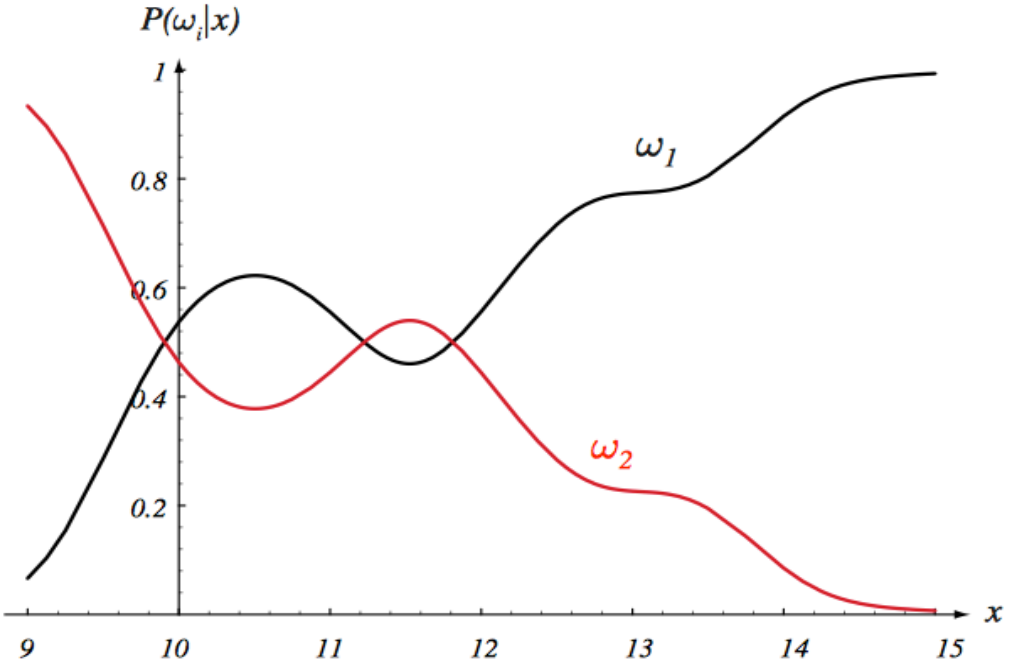


Figure 2–8: *Posterior* probabilities for the *priors* $P(\omega_1) = {}^{2}/_{3}$ and $P(\omega_2) = {}^{1}/_{3}$ for the class-conditional probabilities shown in Figure 2-7 taken from [1].

To ensure that the *posterior* probability is not skewed and to rationalize the decision process, the probability of making an error when a decision is made can also be calculated. For the observation of a given x, the probability of error can be given as:

$$P(error|x) = \begin{cases} P(\omega_1|x) & if\ \omega_2\ is\ decided \\ P(\omega_2|x) & if\ \omega_1\ is\ decided \end{cases} \tag{2.4}$$

Even though the same value of x may not be obtained more than once, if $P(error|x)$ is reasonably small, then the integral that represents the average probability of error must also be reasonably small and is given by:

$$P(error) = \int_{-\infty}^{\infty} P(error, x)\, dx = P(error|x)\, p(x)\, dx \tag{2.5}$$

The above justifies the following *Bayes decision rule* for minimizing the probability of error:

$$Decide\ \omega_1\ if\ P\ (\omega_1|x) > P\ (\omega_2|x);\quad else\ decide\ \omega_2 \tag{2.6}$$

using this rule, eq. (2.4) becomes

$$P(error|x) = \min\ [\ P\ (\omega_1|x), P\ (\omega_2|x)] \tag{2.7}$$

While the decision rule in eq. (2.7) above highlights the relevance of the *posterior* probability in the decision making process, it is worth noting that the *evidence* $p(x)$ in eq. (2.2) is less relevant – being more of a scale factor that ensures the *posterior* probability equates to one. An equivalent decision rule can be obtained even after removing the evidence.

Bayes rule can be said to be reflective of learning, "the transformation from the *prior* to the *posterior* formally reflects what has been learned about the validity of the hypothesis from consideration of the data" [7]. The above concepts can be expanded to allow the consideration of a greater number of features and more *states of nature*. Additionally, decisions other than the *state of nature* can be reached while a loss function can be introduced as a more robust approach than probability of error.

2.4.2 Maximum-likelihood Parameter Estimation

Bayes decision theory is based on an *ideal* problem where a classifier can be designed using knowledge of the *prior* probabilities and class-conditional densities. However, in real-life applications of pattern recognition systems (such as recognizing facial expressions), it is rarely the case that these values are known. This limitation can be mitigated by estimating the unknown values using the samples and using the estimates in place of the true values. Consider a basic supervised learning problem – the estimation of the *prior* probability is straightforward however estimating the class-conditional densities is represented as an instance of the statistical problem of parameter estimation. Assume for example that $p\ (\boldsymbol{x}|\omega_i)$ is a normal density with mean vector $\boldsymbol{\mu}_i$ and covariance (matrix) $\boldsymbol{\Sigma}_i$ the task will be to estimate the *parameters* $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$.

Maximum-Likelihood Estimation (MLE) is one of the common techniques for solving the problem of parameter estimation. With this technique, the parameters to be estimated are defined as quantities whose values are fixed and not known. The best estimate maximizes the probability of obtaining the real samples.

More formally, given a set of independent and identically distributed random variable samples $\mathcal{D}_j$ for $j = 1, \dots, c$ classes, assume that $p\,(\boldsymbol{x}|\omega_i)$ has a known parametric form – distinctively determined by the values of a parameter vector $\boldsymbol{\theta}_j$ consisting of the components of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. Noting that the feature $\boldsymbol{x}$ under consideration in this section is a vector, the task for each class is to obtain the best estimates for the unknown parameter vectors $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_c$.

For each class, let $\mathcal{D}$ be made up of n samples $[\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]$, as such, a set of $\mathcal{D}$ training samples drawn independently from the probability density $p\,(\boldsymbol{x}|\boldsymbol{\theta})$ can be used to estimate the unknown parameter $\boldsymbol{\theta}$ – consequently, the *maximum likelihood estimate* is given by:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p\,(\boldsymbol{x}_k|\boldsymbol{\theta}) \tag{2.8}$$

The main advantage of MLE is that it is asymptotically unbiased and of minimum variance. This is particularly important as *bias* and *variance* are units of measure with which the efficacy of a learning algorithm towards a classification problem can be judged. In other words, how well the learning algorithm *matches* the classification problem – the *bias* measures the quality or accuracy of the classifier; high bias suggests a poor match. The *variance* is a measure of precision and specificity of the match; a high *variance* suggests a weak match. It is worth noting that the results obtained using MLE are often nearly the same as the results obtained using *Bayesian estimation* – another commonly used technique for parametric estimation [1].

The preceding historic approaches to pattern recognition are founded on the flawed assumption that prior probabilities, class-conditional densities and/or the forms of underlying density functions relating to the problem are exactly known. In practice, this is very rarely the case if ever. Expectedly, over time several other approaches (both supervised and unsupervised) have been explored including but not limited to nonparametric techniques, linear discriminant functions, multilayer neural networks, stochastic methods, nonmetric methods, algorithm-independent machine learning as well as unsupervised learning methods and clustering [1]. Additionally, data sizes and computational complexities corresponding to different feature extractors and classifiers have made some systems more attractive than others. For example, the methods applied in chapters four and five of this thesis are largely based on linear discriminant functions which are optimal when it is assumed that the underlying distributions are *cooperative*, such as Gaussians having equal covariance. Essentially, the parameters are estimated by the formation of the covariance matrix and the extracted features approximate the maximum likelihood. The following section will look at pattern recognition as the specific problem of human Facial Expression Recognition (FER) and take a high level conceptual view of some literary works that have applied FER in one way or another. In subsequent chapters (three, four and five), a technical review of literature related to the topics contained in each chapter is included in the corresponding background sections.

2.5 Applications of Facial Expression Recognition

In this section, models for real-world application as well as actual implementations of FER systems will be conceptually reviewed. As long as advances continue in the technical area of FER, there is bound to be a proportional increase in the application of such systems, to date FER systems have been applied in a variety of areas including but certainly not limited to human-computer interaction, automotive design, gaming and entertainment, human psychology, assistive technologies, medical applications, marketing and consumer analysis (discussed below). In addition to the review of selected scholarly works on specific application efforts, this section outlines recently emerging professional facial expression recognition software that is being made available for commercial use.

2.5.1 Robot Design

In HCI, the design of humanised robots and the ways they interact with humans is one of the prevalent areas where FER is applied. Moreover, robot design serves as a general template on which HCI concepts can be proven prior to subsequent application in other areas of interest. Wimmer et. al. affirmed that "to be effective in the human world, robots must respond to human emotional states" [8], in their work, they proposed a prototype for human-robot interaction. They adapt model-based techniques to solve the FER problem using the Cohn-Kanade (CK) database taking into cognizance the mobile nature of robots and the resultant difficulties – variable lighting, variable head pose, separability of humans from other robots and real time computational constraints. Recognition rates of 67% and 70% were obtained on the robot scenario and CK database respectively.

Similarly, in [9] an interactive system was proposed for a robot to reconstruct the facial expressions of a human. The system features the use of a mass-spring model to simulate the tensions of twenty-two facial muscles during the display of facial expressions – the model was based on the assertion in [10] that the mechanical law of soft tissue points are modelled by a nonlinear function. The elastic forces of these tensions are then grouped into a vector, which is used as the input for an expression recognition system. Due to the simulation model, the robot is able to imitate six types of facial expressions. As with [8] above, the recognition system encompasses face detection, feature extraction, classification and artificial emotion – expression generation.

Facial expressions were combined with other concepts including face pose and hand gestures in the proposition of techniques for human–robot interaction in [11] where the emphasis was on gaining an understanding of the positions of facial feature points to aid FER. This was achieved by estimating the 3D positions of each feature point by constructing 3D face models fitted on the user. An Active Appearance Model (AAM) was constructed for variations of the facial expression before estimating the depth information at each feature point from both the frontal and side views. The combination of AAM and the depth estimate fits the 3D model to the user and the (basic) facial expressions recognized are used to change the colours of foreground and background objects in the robot display as well as other robot responses.

In [12], Electrooculographic (EOG) signals were used to recognize the facial expressions, detecting six different movements of the eyes which were then re-enacted by an avatar with the aim of analysing how facial expressions can be better characterized. Time and frequency domain features were extracted from the EOG by the algorithm and were classified in real-time by a multiclass Linear Discriminant Analysis (LDA) classifier with a success rate of 85%. Classification was also performed offline with around 92% accuracy.

Littlewort et al. in [13] presented the progress on a perceptual primitive to automatically detect frontal faces in a video stream and code them with respect to seven facial expressions in real time. The aim was to deploy the system as a tool to measure the quality of human-robot social interaction. This was done in view of the fact that robots bring social considerations to HCI and a desire to replicate the real-time nature of human-human communications in the way humans relate to robots. The novelty at the time was the combination of Adaboost and Support Vector Machines (SVMs) to build a faster and more accurate system that automatically found faces in a visual video stream and automatically coded facial expression dynamics in real-time. The combination of Adaboost and SVMs was able to correctly recognize facial expressions around 95% of the time.

2.5.2 Human Psychology

Human psychology is another field of study that has applied FER in order to gain inroads into the human mind and the implicit messages that are passed across through facial expressions. Marketing and consumer behavioural experts in particular have taken a keen interest. For example, recently researchers in [14] measured the reaction and emotions of consumers in a test in order to aid in the decision making process on the ingredients of sweetener alternatives that function and taste like sucrose in sugar-sweetened beverages but without the associated health risk of sucrose. In their test, the authors used a 9-point hedonic scale, the responses from an explicit emotion term questionnaire and implicit facial expression reactions to sweeteners in tea to measure the relationship between consumer acceptability and emotional responses. FaceReader 5.0 (professional software – mentioned later in this section) was used to capture the facial expressions. This software was also used in [15] for a similar application of FER

– using autonomic nervous system responses and facial expressions to the sight, smell and taste of liked and disliked food.

Along the same lines, in [16], facial responses to online media content were collected and analysed in terms of viewer smile responses to three commercials. The intensity and dynamics of the smile responses showed that there is a substantial difference in the facial responses between participants who reported liking the commercials and those who reported not liking the commercials. The authors were also able to distinguish between groups who were previously familiar with a commercial and those that were not thereby proposing a link to the virality of the commercials. Assertions were also made on the relationship between head movements and facial actions. The results of the analysis using the collected dataset were also compared to those obtained from using the Cohn-Kanade+ (CK+) and MMI datasets. Other notable works in the area of recognizing smiles, online media and marketing include [17–22].

A more peculiar application of automatic FER in human psychology was in [23] where the authors investigated how human attention is allocated in the presence of stimuli that represent threats even when no actual threats were present. Specifically, the authors sought to test the hypothesis that there is a bias toward facial expressions of fear and disgust in animal phobic people. A gaze-cuing paradigm in which participants' attention was driven by the task-irrelevant gaze of a centrally presented face was used, employing negative facial expressions of disgust, fear and anger. They found an increased gaze-cuing effect in people that had a phobia for snakes as compared to the control subjects irrespective of their facial expression – their results point to a general hyper vigilance in animal phobic people.

2.5.3 Medical and Assistive Technologies

FER was integrated into a proposed system [24] designed to aid the communication of patients with tetraplegia, brain injury, cerebral palsy, neurological injury or stroke. Standard computer input devices such as a keyboard or a mouse may be difficult for such patients to use, moreover where speech is impaired, the use of a computer may represent the primary means of communication. The system is capable of detecting the position of the face/head and the status of the mouth (open/closed) combined with facial expressions in order to control and navigate the cursor of a mouse

as a computer user input interface. The system dubbed Facial position and expression Mouse system (FM) employs a fast and robust Randomized Decision Tree (RDT) to automatically detect the position and expression from an individual image. The position of the face/head and/or the detected facial expression is then mapped to mouse movements or command which are triggered after detection.

Affect and expression variability were used as two of the matrices for exploring gender differences in automatically recognizing nonverbal behaviour indicators of depression and Post-Traumatic Stress Disorder (PTSD) in [25]. Behavioural scientists have already established a connection between psychological disorders and nonverbal behaviour [26]. The authors of [25] sought to go a step further by successfully showing that a gender-dependent approach significantly improves the performance over a gender-independent scheme. They achieved this by identifying a directly interpretable and intuitive set of predictive indicators selected from three general categories of nonverbal behaviours; affect, expression variability and motor variability.

Being able to recognize the facial expressions relating to pain also has potential benefits in terms of medical applications. In [27] twenty participants were videotaped while undergoing thermal heat stimulation at various intensities ranging from non-painful to painful. The pain was induced using a Peltier-based computerized thermal stimulator with a contact probe. The aim was to automatically recognize the videos where pain was induced – this was achieved by adopting a Transferable Belief Model (TBM) based machine-learning system previously applied in expression recognition problems involving the six basic expressions.

Similarly, [28] investigated the medical application of spontaneous facial expressions of pain, the aim was to distinguish between when pain was really being felt and when it was being acted. Twenty-six participants were videotaped under three experimental conditions; a baseline, posed pain and real pain. The real pain condition was elicited by submerging the arm of the participant in ice water. A two-stage machine learning approach was employed – the first stage detected twenty Action Units (AUs) from the Facial Action Coding System (FACS) while the second stage featured a classifier trained to detect the difference between expressions of fake pain and real pain.

2.5.4 Security

Another area FER has been applied is in security. The authors of [29] proposed improving current surveillance systems by adding facial expression recognition to make a system that can flag up a person of interest based on perceived expressions that reflect a possible desire to cause harm or perform unauthorized actions.

In some cases, automatic FER schemes serve as a secondary modality to other facial image processing tools. For example, given an authentication problem, a facial recognition system could be made more robust by being independent of facial expressions. This could be achieved by integrating a facial expression recognition classification tool into the facial recognition system. The designers of a new product currently undergoing beta testing – *Chui* described as "an intelligent doorbell that uses facial recognition to make your home keyless, secure and individualized" [30] allowing doors to be unlocked using facial identity. However the product also uses FER not only to ensure robustness but also to attempt to detect entries from authorized persons under duress.

2.5.5 Professional Software

Following the recent launch of commercial FER solutions, the following is a list of notable professional software that can be used for automatic FER related tasks.

Software [Reference]	**Description**
Affdex [17, 18]	FER in Real-time online via Webcam/mobile
Emotient [33]	FER over the Cloud/offline
Eyeris [34]	Deep learning based emotion analytics
FaceReader 5.0 [35]	FER in Real-time from video/still images
iMotions [36]	Biometric research platform
Kairos [37]	FER in videos, crowd analytics

Nviso [38]	Online/offline emotion analytics
Sightcorp [39]	Real-time offline face analytics

Table 2–1: List and description of professional/commercial software for facial expression recognition.

The idea of professional software being available for FER is exciting as it should increase application in areas where there is no interest in the technical workings of FER systems, this increase in application should then in turn boost the research interest of automatic FER. However, this adoption is likely to be mitigated by factors such as the very high price of these products currently. Additionally, there are perceived limitations to the practical performance of these solutions, for example in [14] (reviewed above), there was an unexpected lack of significant correlation between subjects levels of likeness and emotional response. It can be hypothesized that this may be due in part to potency of the software (FaceReader 5.0) or lack thereof. This hypothesis is also shared in an academic forum [40] where a review stated in part "it was quite accurate for happiness, but less accurate for other emotions unless they are very stereotypical (eyes wide open, etc.)". The above points to the fact there is still the need for continued research in the area of FER to boost the adoption of FER systems particularly by non-technological consumers.

2.6 Summary

In this chapter, the following was presented; descriptions of a typical pattern recognition system and the challenges associated with such systems. Historical approaches to PR as well as a review of literature and applications of PR in FER are also presented.

The aim of this chapter was to lay a foundation and introduce the notable schemes on which the subsequent chapters are based; as such it does not represent an exhaustive review of the literature. In addition to referenced content in this chapter and further reviews of relevant literature in subsequent chapters, further resources can be found in [27–31].

From the review of the literature in this chapter, it is clear there is still a gap in the literature with respect to research that uses representative training samples partly due in part to the number of available training datasets. Consequently, in the next chapter a new spontaneous facial expression database will be presented.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Second Ed. John Wiley & Sons, Inc., 2001.

[2] S. Watanabe, *Pattern recognition: human and mechanical*. Wiley, New York, 1985.

[3] M. Vento, "What is pattern recognition?," *IAPR Newsl.*, vol. 25, no. 1, 2003.

[4] K. Kpalma and J. Ronsin, "An overview of advances of pattern recognition systems in computer vision," *Vis. Syst. Segmentation Pattern Recognit.*, pp. 169–194, 2007.

[5] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[6] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimed.*, vol. 4, no. 1, pp. 23–37, 2002.

[7] B. Olshausen, "Bayesian probability theory," *Redw. Cent. Theor. Neurosci. Helen Wills Neurosci. Inst. Univ. Calif. Berkeley, Berkeley, CA*, pp. 1–6, 2004.

[8] M. Wimmer, B. A. Macdonald, D. Jayamuni, and A. Yadav, "Facial expression recognition for human-robot interaction – a prototype," *Robot Vision, Springer Berlin Heidelb.*, pp. pp. 139–152., 2008.

[9] S. S. Ge, C. Wang, and C. C. Hang, "Facial expression imitation in human robot interaction," *RO-MAN 2008 - 17th IEEE Int. Symp. Robot Hum. Interact. Commun.*, pp. 213–218, 2008.

[10] Y. Z. Y. Zhang, E. C. Prakash, and E. Sung, "A new physical model with multilayer architecture for facial expression animation using dynamic adaptive mesh," *IEEE Trans. Vis. Comput. Graph.*, vol. 10, no. 3, pp. 339–352, 2004.

[11] M. H. Ju and H. B. Kang, "Emotional interaction with a robot using facial expressions, face pose and hand gestures," *Int. J. Adv. Robot. Syst.*, vol. 95, no. 9, 2012.

[12] A. Cruz, D. Garcia, G. Pires, and U. Nunes, "Facial expression recognition based on EOG toward emotion detection for human-robot interaction," *ResearchGate*, vol. 2, 2013.

[13] G. C. Littlewort, M. S. Bartlett, I. R. Fasel, J. Chenu, T. Kanda, H. Ishiguro, and J. R. Movellan, "Towards social robots: automatic evaluation of human-robot interaction by facial expression classification," *Adv. Neural Inf. Process. Syst. 16*, pp. 1563–1570, 2004.

[14] K. A. Leitch, S. E. Duncan, S. O'Keefe, R. Rudd, and D. L. Gallagher, "Characterizing consumer emotional response to sweeteners using an emotion terminology questionnaire and facial expression analysis," *Food Res. Int.*, 2015.

[15] R. A. de Wijk, V. Kooijman, R. H. G. Verhoeven, N. T. E. Holthuysen, and C. de Graaf, "Autonomic nervous system responses on and facial expressions to the sight, smell, and taste of liked and disliked foods," *Food Qual. Prefer.*, vol. 26, no. 2, pp. 196–203, 2012.

[16] D. McDuff, R. El Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 456–468, 2012.

[17] D. McDuff, R. El Kaliouby, E. Kodra, and L. Larguinat, "Do emotions in advertising drive sales? - use of facial coding to understand the relationship between emotional responses to ads and sales effectiveness," *Affectiva*, 2013.

[18] T. Senechal, J. Turcot, and R. El Kaliouby, "Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience," *10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, 2013.

[19] D. McDuff, R. El Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," *10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, 2013.

[20] T. Teixeira, R. Picard, and R. el Kaliouby, "Why, when and how much to entertain consumers in advertisements? a web-based facial tracking field study," *Mark. Sci.*, vol. 33, no. 6, pp. 809–827, 2014.

[21] D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard, "Automatic measurement of ad preferences from facial responses gathered over the internet," *Image Vis. Comput.*, vol. 32, no. 10, pp. 630–640, 2014.

[22] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. Picard, "Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads," pp. 1–13, 2014.

[23] C. Pletti, M. Dalmaso, M. Sarlo, and G. Galfano, "Gaze cuing of attention

in snake phobic women : the influence of facial expression," *Front. Psychol.*, vol. 6, pp. 1–8, 2015.

[24] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Facial position and expression based human computer interface for persons with tetraplegia," *IEEE J. Biomed. Heal. Informatics*, vol. 2194, no. c, pp. 1–11, 2015.

[25] G. Stratou, S. Scherer, J. Gratch, and L. P. Morency, "Automatic nonverbal behavior indicators of depression and PTSD: exploring gender differences," *Proc. - 2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact. ACII 2013*, pp. 147–152, 2013.

[26] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 1989.

[27] Z. Hammal, M. Kunz, M. Arguin, and F. Gosselin, "Spontaneous pain expression recognition in video sequences," in *BCS Int. Acad. Conf.*, pp. 191–210, 2008.

[28] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1797–1803, 2009.

[29] A. Al-Modwahi, O. Sebetela, and L. Batleng, "Facial expression recognition intelligent security system for real time surveillance," *Elrond. Inform.*, vol. 1, pp. 1–8, 2012.

[30] "Chui" [Online]. Available: https://www.getchui.com/. [Accessed: 23-Jun-2015].

[31] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): naturalistic and spontaneous facial expressions collected 'in-the-wild,'" in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 881–888, 2013.

[32] "Affectiva" [Online]. Available: http://www.affectiva.com/. [Accessed: 20-Jun-2015].

[33] "Emotient" [Online]. Available: http://www.emotient.com/. [Accessed: 20-Jun-2015].

[34] "Eyeris" [Online]. Available: http://emovu.com/e/. [Accessed: 21-Jun-2015].

[35] "FaceReader" [Online]. Available: http://www.noldus.com/human-

behavior-research/products/facereader. [Accessed: 21-Jun-2015].

[36] "iMotions" [Online]. Available: http://imotionsglobal.com/software/add-on-modules/attention-tool-facet-module-facial-action-coding-system-facs/. [Accessed: 21-Jun-2015].

[37] "Kairos" [Online]. Available: https://www.kairos.com/. [Accessed: 21-Jun-2015].

[38] "Nviso" [Online]. Available: http://www.nviso.ch/index.html. [Accessed: 21-Jun-2015].

[39] "Sightcorp" [Online]. Available: http://sightcorp.com/. [Accessed: 21-Jun-2015].

[40] "ResearchGate" [Online]. Available: http://www.researchgate.net/post/Does_anybody_know_a_free_software_for_emotional_facial_recognition. [Accessed: 19-Jun-2015].

[41] C. Chibelushi and F. Bourel, "Facial expression recognition: a brief tutorial overview," *On-Line Compend. Comput. Vis.*, 2003.

[42] V. Bettadapura, "Face expression recognition and analysis: the state of the art," *arXiv Prepr. arXiv1203.6722*, 2012.

[43] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[44] A. Ryan, J. Cohn, and S. Lucey, "Automated facial expression recognition system," *43rd Annu. 2009 Int. Carnahan Conf. Secur. Technol.*, pp. 172–177, 2009.

[45] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition : a review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–38, 2000.

Chapter Three

New Spontaneous Expression Database

3.1 Introduction

As computing continues to become a natural part of everyday life, there has been a concomitant rise in the development of autonomous systems and the applications which they serve – the early evolution of which can be observed in [1]. Further to this is the corresponding need to keep up with an imminent shift in the design of how humans interact with ever-growing automated environments. This shift can be represented by systems that will be designed to take into account factors implicitly portrayed by the user and not just the explicit inputs through the use of traditional input devices such as the mouse and keyboard. As facial expressions are one of the instinctive ways humans communicate with other humans, it is important in the context of these emerging trends (humans communicating and interacting with computers in the same way as they communicate and interact with each other) that human emotions can be automatically deduced. Integral to understanding these emotions is the need to correctly recognize facial expressions thereby being able to infer the emotions that induced them. As a result there has been a substantial research interest in the area of automatic Facial Expression Recognition (FER) by researchers in the areas of image processing, human-computer interaction, affective computing, pattern classification and machine learning amongst others [1–3].

The task of automatically classifying facial expressions can be defined as an object recognition problem that involves the use of labelled training samples from k different object classes to accurately establish the class to which a new test sample $\boldsymbol{y}$ belongs [4]. However, in comparison to other areas of facial image processing such as facial recognition, facial expression recognition has a somewhat limited number of publicly available databases containing labelled training data. Moreover, the majority of previous research on the recognition of facial expressions employs the use of posed image databases such as the JAFFE database [5]. The images contained in a posed expression database are artificial and often exaggerated and/or non-realistic because they are elicited by asking subjects who are aware they are being captured on camera to act out a sequence of facial expressions [6]. These posed images as such do not best represent the natural way humans communicate expressions on their faces. Whereas in real day-to-day life, the display of facial expressions is more spur-of-the-moment, these are better represented in spontaneous facial expression databases which are often more subtle and less intense but more representative for use as training/test data.

A problem that may be one of the main factors slowing down the practical deployment of FER systems is the limited number of freely available spontaneous (natural) facial expression databases. Where they exist, they are limited by factors including; availability, number of participant subjects, number of expression classes, number of images per class and good baseline recognition results. The work in this chapter will bridge some of these research gaps.

This chapter proposes and establishes the Loughborough University Spontaneous facial Expression Database (LUSED) which is a labelled database of natural facial expression images which will be made publicly available for the design, training and testing of algorithms and systems for automatically inferring emotional and affective states through the classification and recognition of human facial expressions. In subsequent chapters, we present results of evaluations using the new LUSED database obtained from the use of several common classification-based techniques including Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA); in addition, we evaluate using Sparse Representation-based Classification (SRC) techniques and various combinations of SRC and other techniques.

3.1.1 Statement of Novelty

This chapter provides details of the work relating to the creation of the database; LUSED. The new database represents a novel contribution to knowledge as it addresses specific gaps in literature relating to public availability, the demography of the participants that make up the database, the number of participants, the way in which the data are labelled and the methods by which the expressions are elicited from the participants to begin with. Specifically, in this Chapter the following are presented;

1. A comprehensive and up-to-date review of existing spontaneous (natural) facial expression databases.

2. Details on the creation of a new facial expression database with established ground truths from five out of the six basic expression classes.

The remainder of this Chapter will be organized as follows; Section 3.2 contains a background of facial expressions. An overview of noteworthy databases and an up to date description of natural (spontaneous) databases is contained in Section 3.3, although already briefly touched on in Chapter 2 above. Section 3.4 discuses the characteristics, both affective and physical appearance of the six basic expressions (of interest). The experimental setup for the creation of the new database is detailed in Section 3.5 followed by a statement of ethics in Section 3.6. An analysis of the participant demography is contained Section 3.7 while the steps taken for post processing are described in Section 3.8. A summary of the Chapter is presented in Section 3.9.

3.2 Background on Facial Expressions and Databases

Facial expression databases – labelled images of facial expressions are requisite in the design of any autonomous/semi-autonomous affect recognition systems. They also allow researchers in the community to objectively compare results. Most of the existing studies on the automatic analysis of human affective displays have been based on the "artificial" material of deliberately expressed emotions [7], and there is considerable literature on the development and use of several posed databases especially with regards to the six prototypical expressions i.e. anger, disgust, fear, happiness,

sadness and surprise. Sometimes a neutral expression is added in which case it will be referred to as the seven-prototypical expressions.

Notwithstanding the above, in this section, firstly the background of the six basic expression classes in the context of FER is discussed then the important distinction between the posed expressions Section (3.2.1) and natural expressions Section (3.2.2) is re-established before presenting a taxonomy of existing databases in both categories (Posed and Natural).

3.2.1 Posed Facial Expressions

Posed expressions by definition are often exaggerated and/or non-realistic because they are elicited by asking subjects who are aware that they are being captured on camera to act out a sequence of facial expressions (sample images can be observed in Section 3.4). They are likely given examples and direction as to the form and intensity of the expressions they are asked to elicit. Most of the resulting posed databases are characterized by clear constrained input examples in the form of high quality visual recordings and images, non obstructed/occluded faces which are in full frontal view and in the desired orientation. A consequence of the above is that images within the same expression class are very similar to each other at the same time as being very dissimilar to images from the other expression classes. This high intra-class correlation and concurrent low inter-class correlation forms an artificial 'best case scenario' dataset for the object recognition problem. Results obtained by algorithms and researchers testing on such datasets are misrepresentative in comparison to what should be expected from real-life applications.

3.2.2 Natural Facial Expressions

Subsection (3.2.1) above encapsulates the reasons behind the aspiration of researchers to use natural facial expression in algorithm design. More so in light of the development of multimodal approaches to affect detection, where for example, the recognition of facial expressions could be combined with speech analysis and/or gait recognition amongst other modalities to better discern the holistic affective state of a subject. These other natural modalities (speech and gait in this example) already have natural training datasets and as such should be combined with natural facial expressions.

In addition, as articulated in [7], it is well documented in fields such as psychology and neuroscience that spontaneous and purposely displayed facial behaviour has differences both in terms of facial muscles used and their dynamics (e.g., [8]). For example, many types of natural smiles (e.g., polite) are smaller in amplitude, longer in total duration, and slower in onset and offset times than posed smiles (e.g., [7–9]). Likewise, it has been shown that spontaneous brow actions (AU1, AU2, and AU4 in the FACS system) have different morphological and temporal characteristics (intensity, duration, and occurrence order) than posed brow actions [9]. It is not surprising, therefore, that methods of automated human affect analysis that have been trained on deliberate and often exaggerated behaviours usually fail to generalize to the subtlety and complexity of spontaneous affective behaviour.

However, authentic affective expressions are difficult to collect because they are somewhat uncommon, short lived, and laden with subtle contextual changes that make it difficult to successfully elicit affective displays without unduly influencing the resultant affective displays. In one or two cases, the natural database is formulated using clips collected from TV recordings or other events that have been captured live e.g. talk shows as is the case in [11] and some parts of [12]. An added difficulty that is peculiar to the collection of a natural database is the need for manual labelling of expressions for establishing the 'ground truths' which is very time consuming, error prone, and expensive. As a result, in the past researchers tend to use the limited amount of data like the authors of [13] who explore one-class classification application in analysing spontaneous facial expressions.

3.3 Taxonomy of Facial Expression Databases

There are approximately 50 existing databases containing deliberately exhibited (acted) affective behaviour with facial expressions with some including varying modalities e.g. speech. These are composed of images of subjects, with the number of images and participants varying but mostly comprising of the 7-prototypical expression classes. There is a changing degree of availability and accessibility of these databases to the research community; as a result, some are more popular than others. Worthy of mention firstly, is the Cohn-Kanade (CK) facial expression database [14], which contains Action Unit (AU) coded images (also grouped into 6 prototypical classes) from

over 100 students aged from 18 -30. It is one of the most widely used databases for facial expression recognition. The same research group released an extension of the original CK database in 2010. The extended Cohn-Kanade database (CK+) [15] contained data from the onset through the apex and to the offset of the expressions as well as some spontaneous images. There is also the Japanese Female Facial Expression (JAFFE) database [5] from 10 subjects each contributing 7 pictures in each of the 7 classes. Similar to this in size is the Yale Face database of 165 images from 15 subjects [16]. Others include the AT&T database (formerly the ORL database). Due to reasons articulated in Section 3.1 above, the development and publication of posed databases have been few and far between in recent times except in cases where the dataset is required to include parameters not yet obtainable spontaneously. An example of recent work in the posed domain is [17] where the authors collect 21 distinct emotion categories, that they describe as compound expressions from 230 adults, e.g. happily surprised and angrily surprised. Also fairly recent is the MPI database [18], which also comprises of 3-dimensional facial images from 19 participants who elicited 55 different facial expressions.

There is much less in terms of well-annotated, publicly available, spontaneous (natural) databases for facial expression recognition. The MMI facial expression database [19, 20] is one of the most comprehensive data sets of facial behaviour recordings. The database, which contains both static images and videos, where a large part of the video recordings were recorded in both the frontal and the profile views of the face, represents a facial behaviour data repository that is available, searchable, and downloadable via the Internet. However it is collected from only 29 subjects, which is an inherent limitation. There is also the NVIE database [6], developed at USTC which contains images from 215 subjects, even though images from all 6 expression classes and neutral are included in the database, the authors only certify the images of disgust, fear and happiness as being truly spontaneous. Other spontaneous databases that have been developed include the Geneva Airport Lost Luggage Study [21] where subjects were filmed at the lost luggage counter and interviewed subsequently. It contains 5 emotion classes from 109 subjects. 1872 facial images were also collated from footage of a television talk show and labelled into 3 groups Valence (positive vs. negative), Activation (excited vs. calm) and dominance (strong vs. weak) to form the VAM database [11]. One of the more recent contributions is the BP4D database [22] where

multidimensional data were collected from 41 subjects participating in social interviews, playing games and watching clips. Similarly is the recent DISFA database [23] which comprises images from 27 adults (12 women and 15 men) that vary in age from 18 to 50 years. Three were Asian, 21 Euro-American, two Hispanic, and one African-American. Participants viewed a 4-minute video clip (242 seconds in length) intended to elicit spontaneous AUs in response to videos intended to elicit a range of facial expressions of emotion.

Table 3-1 below, presents a comprehensive list/details of known, existing natural facial expression databases. Even though it is not claimed the review is exhaustive, it does offer an up-to-date overview. The list contains 17 databases, which is much less than what is available for posed facial expression research and more or less adequate when compared to other branches of human facial analysis e.g. face recognition and age estimation. This is drawn on, as justification for the development of the new database, which is presented in this chapter. Further justification is obtained after examination of the following parameters that characterize the existing natural databases;

1. Method of elicitation
2. Labelling type
3. Participant size
4. Participant demography
5. Public availability

A look at the freely available databases shows limitations in the diversity of the participant demographic e.g. along the lines of age range, ethnicity and gender. For example, the AvID database [24] has 15 participants which comprise of 12 females and only three males. The NVIE database [6] is made up of participants, 99% of whom have an Asian appearance. On the other hand, the DISFA database [23] mentioned above has a fairly balanced demographic but has a total of only 15 subjects. The method/format for labelling the ground truths in about 30% of the datasets is the Facial Action Coding System (FACS) while the remainder are split between the community agreed ‘basic expressions’ (also referred to as the six prototypical expressions) and variants of the basic expressions suited for the intended purpose of the authors’ research. For example,

both [11] and [25] are labelled by measure of valence i.e. positive or negative expressions. The method of stimulating participants as well as the material used also contributes to the quality of spontaneity in the expressions elicited. Most datasets adopt the use of videos and/or tasks or similar interactions to stimulate participants. The authors of [26] performed and recorded tests on sufferers of shoulder pain albeit with the aim of labelling the data with AUs relating to pain.

Database [Reference]	**Method of elicitation**	**Labelling/Classes**	**Size**	**Free Access**
AAI [27]	Subject interviews – asked to describe childhood experiences	11 Classes; 6 prototypical expressions, embarrassment, contempt, shame, general positive and negative	60 Subjects – each with a 30-60 minute audio/video clip	N/A
AM-FED [28]	Webcam recordings over the internet of people viewing online media	10 Symmetrical and 4 Asymmetric FACS AUs, head movements, gender	242 Video sequences (168,359 frames) from 200+ people	Yes
AvID [24]	Subjects describe neutral photographs, play a game of Tetris and solve cognitive tasks	4 Classes; Neutral, relaxed, moderately and highly aroused	15 Subjects – each with a 60 minute audio/video clip	Yes
Belfast [29]	Interactive chats	Multiclass; Anger, fear	125 Subjects – 298 audio/video clips	N/A
BP4D [22]	Subjects participated in social interviews, game, watching video clips etc.	FACS; 27 AUs coded	41 Subjects	N/A
DISFA [23], [30]	Subjects watched emotion- inducing videos	FACS; 12 AUs intensity coded	27 Subjects – 130,000 video frames	N/A
DynEmo [25]	Subjects performed emotion-inducing tasks	Valence (positive vs. negative)	358 Subjects	Yes
Geneva Airport Lost Luggage Study [21]	Subjects filmed at Geneva airport lost luggage counter + follow up interview	5 Classes; Anger, good humour, indifference, stress and sadness	109 Subjects – audio/video	N/A
HUMAINE [12]	Television recordings	Multiclass; positive and negative, active and passive,	Multi-subject – 48 clips between 3 – 120	N/A

		consistent, co-existent emotion, emotional transition over time	seconds in length	
MMI [20]	Adults watch emotion-inducing videos while children were told jokes	FACS; 79 AUs and their combinations	29 Subjects [18 adults + 11 children] – 65 videos	Yes
NVIE-USTC [6]	Subjects watched emotion-inducing videos	Multiclass; Disgust, fear, happiness	215 Subjects	Yes
RU-FACS [31]	Subjects tried to convince the interviewers that they were telling the truth	FACS; 33 AUs	100 Subjects	N/A
SALAS [32]	Subjects talk to artificial listener with changing emotional states as a result of interaction	Variety of non-intense emotions/emotion related states	20 Subjects	N/A
SMARTKOM [33]	Subjects solving tasks with system- human machine in Woz scenario	9 Classes; Joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise and neutral	224 Subjects – 4/5 minute clips	N/A
UA-UIUC [34]	Subjects watched emotion-inducing videos	4 Classes; Neutral, joy, surprise and disgust	28 Subjects – each with a video clip	N/A
UNBC-McMaster Pain Archive [26]	Subjects underwent shoulder motion tests for pain	FACS; AUs related to pain coded	129 Subjects – 200 video clips	Yes
UT-Dallas [35]	Subjects watched emotion-inducing videos	Multiclass; Happiness, sadness, fear, anger, boredom	284 Subjects - from which about 1540 5-second standardised clips are obtained	N/A
VAM [11]	Television talk show	3 Groups; Valence [positive vs. negative], Activation [excited vs. clam] and dominance [strong vs. weak]	104 Subjects – 1872 Facial images and 1421 segmented utterance videos	N/A

Table 3–1: Comprehensive List with Details of Natural Expression Databases

3.4 The (Basic) Facial Expressions of Interest: Characteristics

In this section, a brief psychological review of the 'basic expressions' of the face that was sought to elicit is given. 'Discrete categories' is one of the most consistent ways of quantifying *affect* by psychologists, an approach that is embedded in the semantics of daily life [35–37]. It is one of the ways affect has been conceptualized in psychological research. Correspondingly, an understanding of the basic structure and description of affect is important in that these conceptualizations provide information about the affective displays that automatic emotion recognition systems are designed to detect [7]. A standard illustration of this description is the prototypical (basic) emotion classes, which include anger, disgust, fear, happiness, sadness and surprise. Cross-cultural studies conducted by Ekman [36, 39], corroborated this description of basic emotions indicated that humans perceive certain basic emotions with respect to facial expressions in the same way, regardless of culture. It is worth pointing out at this point that the assertions in respect of culture do not represent an opinion on ethnicity. The influence of this basic emotion theory can be seen in the fact that many of the initial research in facial expression recognition focused on the recognition of theses emotions. Even though the prototypical emotions cover a rather small part of our daily emotional displays (as they often lack context), they are key points of emotion reference. The positive is that this structure of labelling based on category is very intuitive and hence matches human experiences. As research continues to progress in natural FER, the basis for future work will be to attempt to elicit more detailed natural expressions than the basic six e.g. as was done in the posed case where 21 'compound expressions' were defined in [17]. However prior to this, a solid foundation, which includes datasets, such as the LUSED, which is being reported on in this chapter, needs to be laid.

There are many parameters, varying in complexity and technicality that have been used to analyse the structure of each of the prototypical expressions, which are well documented in the literature. In Section (3.4.1–3.4.6), for completeness, broad visual characteristics of each of the six expressions are offered. It is noteworthy that these physical and psychological reviews are by no means exhaustive in their description but are a generalization as relates to most people.

3.4.1 Anger

An intense emotion, anger can be aroused in a number of different ways e.g. frustration. It is not only one of the most difficult expressions to elicit in a controlled experiment such as this work but it is also fairly difficult to interpret solely from facial expressions. The appearance of anger is defined by eyebrows that are lowered and drawn together, tensed eyelids, hard stare from the eyes and lips that are either tightly pressed together or parted in a square shape. Figure 3-1 shows examples of posed depictions of anger from two databases.

Anger can be caused in a number of different ways [40], for example; (1) the threat of physical hurt to someone could arouse anger, which will often be combined with other emotions such as fear or contempt. (2) A more common cause of anger is frustration, which could result from an interference with routine actions or aspirations. (3) Provocation to anger is the threat or effect of psychological hurt rather than physical as in the first example. (4) The final example is when anger is prompted by observing something which contradicts values held in high esteem, be it; moral, religious or cultural. This final example is leveraged on in the selection of the stimuli video that participants watch during the experiments in the development of LUSED (more details in Section 3.5).

(a) (b) (c)

Figure 3–1: Examples of Images Displaying Posed Anger from the JAFFE Database (a) and the MUG Database (b) and (c).

3.4.2 Disgust

This is a feeling of repugnance that is close to the feeling of *contempt* except that disgust relates to not only to people and their actions, but can relate to tastes and smells for example or even the thought of the taste or smell of something being watched or observed. The appearance of disgust is more defined in the lower part of the face where the upper lip is raised while the lower lip may be raised or lowered, the nose is wrinkled and in the upper part, the lower eyelids are pushed up and the eyebrow is lowered. Figure 3-2 shows examples of posed depictions of disgust from two databases while examples of natural (spontaneous) expressions of disgust can be seen in Figure 3-13(a) and (b).

Disgust usually involves getting-rid-of and getting-away-from responses. The goal is to either remove the object or remove oneself from the cause of the disgust. Nausea and vomiting can occur with the most extreme experiences of disgust [40]. It is worth noting that the examples of things that could induce disgust mentioned above (thoughts, tastes and smells) are more subjective from person-to-person based on factors such as culture. An easy example is that of food; some people find the thought of eating dog meat disgusting while to others this is normal. This formed the consideration for the use of multiple clips within the stimuli video to attempt to induce disgust.

(a) (b) (c)

Figure 3–2: Examples of Images Displaying Posed Disgust from the JAFFE Database (a) and the MUG Database (b) and (c).

3.4.3 Fear

Often triggered in advance of perceived harm, fear is a purely negative emotion, which is one of the ways it differs from surprise (which can be either positive or negative) even though they are similar in appearance and often confused. The facial reaction to fear features the raising and drawing together of the eyebrows, opening of the eyes with the tensing of the lower lid and the stretching back of the lips. Figure 3-3 shows examples of posed depictions of fear from two databases while examples of natural (spontaneous) expressions of fear can be seen in Figure 3-13 (c) and (d).

Fear is an evolving emotion; this evolution is sometimes in terms of its intensity, which ranges from apprehension to terror. As with anger, fear is also a difficult expression to elicit from all participants in a controlled experiment environment. This can be partly attributed to the fact that it is an emotion that occurs with advance warning of danger and/or the sudden commencement of harm mostly manifested by pain. The imminent danger could be either physical or psychological. With the stimuli video (discussed in Section 3.5) designed to initially make subjects comfortable and implicitly safe, it is challenging to spontaneously induce fear. This difficulty is evident by the number of participants who spontaneously displayed fear as a function of the total number of participants in the LUSED experiment and can be observed in Figure 3-12, which shows the number of images from each class contained in the database – expressed in percentage.

(a) (b) (c)

Figure 3–3: Examples of Images Displaying Posed Fear from the JAFFE Database (a) and the MUG Database (b) and (c).

3.4.4 Happiness

This is a positive emotional state attainable via a number of routes e.g. pleasure and excitement although those examples are emotional states sometimes referred to as happiness. In the context of this work however, and in the wider context of facial expression recognition, pleasure is regarded as positive physical sensations, the opposite to the physical sensation of pain; even though the English language synonymizes happiness with pleasure. Excitement on the other hand is the opposite of boredom; often triggered when interest is aroused by something.

It is one of the more straightforward expressions to interpret and elicit, partly because it is the emotion most people want to experience [40]. Happiness is facially expressed with corners of the lips drawn back and upwards. The mouth may be parted with teeth exposed, as is the case with grinning and/or laughter. In the case of a smile, mouth and lips may not be parted. Wrinkles may run down from the nose to the outer edge beyond the lip corners, raised cheeks and wrinkling below the lower eyelids. Figure 3-4 shows examples of posed depictions of happiness from two databases while examples of natural (spontaneous) expressions of happiness can be seen in Figure 3-13 (e) and (f).

(a) (b) (c)

Figure 3–4: Examples of Images Displaying Posed Happiness from the JAFFE Database (a) and the MUG Database (b) and (c).

3.4.5 Sadness

Sadness is an affective state that is also difficult to interpret and elicit and is often confused with anger in terms of physical appearance. Likely induced by muted suffering frequently associated with some form of loss, disappointment or hopelessness; it is considered a passive emotion mostly associated with non-physical pain [40]. In terms of the physical appearance of sadness, the inner corners of the eyebrows are raised and may be drawn together, the inner corner of the upper eyelid is drawn up and the lower eyelid may appear raised. The corners of the lips are drawn down, or the lips appear to tremble.

Sadness is a negative emotion; it is a variant or form of distress except for the fact that distress is more associated with physical pain. Sadness often occurs after distress. For example, if a loved one were to be killed suddenly in an accident, the immediate reaction will be that of distress likely combined with shock and/or anger. This could be accompanied by audible crying/weeping, which represents a sign of disbelief and protest against the loss. Following the physical displays of the pain of the loss is the grief and mourning which better represent sadness; a resignation to the reality of the loss. Sadness can also be felt as a form of empathy when witnessing or observing others sorrow. Figure 3-5 shows examples of posed depictions of sadness from two databases while examples of natural (spontaneous) expressions of sadness can be seen in Figure 3-13(g) and (h).

(a) (b) (c)

Figure 3–5: Examples of Images Displaying Posed Sadness from the JAFFE Database (a) and the MUG Database (b) and (c).

3.4.6 Surprise

This emotion is sudden in onset and brief in duration. Surprise is triggered both by *unexpected* and *misexpected* events. The facial expression corresponding to surprise is characterized by the raising of the eyebrows, the widening of the eyes and pupils, and the dropping open of the jaws and parting of the lips [40]. Figure 3-6 shows examples of posed depictions of surprise from two databases while examples of natural (spontaneous) expressions of surprise can be seen in Figure 3-13(i) and (j).

The unexpected event(s) that trigger the feeling of surprise can be in almost any form; smells, sounds, sights, tastes and even non physical feelings; thoughts, ideas etcetera. Surprise varies in intensity from mild to extreme and can be either positive or negative. Although relatively one of the easier expressions to induce, the expression of surprise on the face is often very brief commonly evolving to the expression of another related expression; for example one could express surprise at finding something that had been long lost, this expression may quickly morph into that of happiness once the original cause of the surprise has been mentally processed. It is not considered surprise if the person has enough time to think about whether or not they are surprised.

(a) (b) (c)

Figure 3–6: Examples of Images Displaying Posed Surprise from the JAFFE Database (a) and the MUG Database (b) and (c).

These physical and psychological characteristics of the six 'basic' expressions; outlined in Section 3.4.1 – 3.4.6 combined with feedback from pilot experiments formed the basis for the design of the experiments to collect the new database. The experimental setup is detailed in Section 3.5 below.

3.5 Experimental Setup

In this section, the process of the collection of the novel Loughborough University Spontaneous Expression Database (LUSED) is described; mainly the experimental setup and procedures. Also related but in subsequent sections is a 'statement of ethics' Section (3.6), an analysis of the participant demographic Section (3.7) and post processing of the data Section (3.8). Figure 3-7 depicts the process flow in the development of the database from the design of the experiment contained in this section to the simulations and testing using baseline algorithms (Chapter 4) and more recent recognition techniques (Chapter 5).

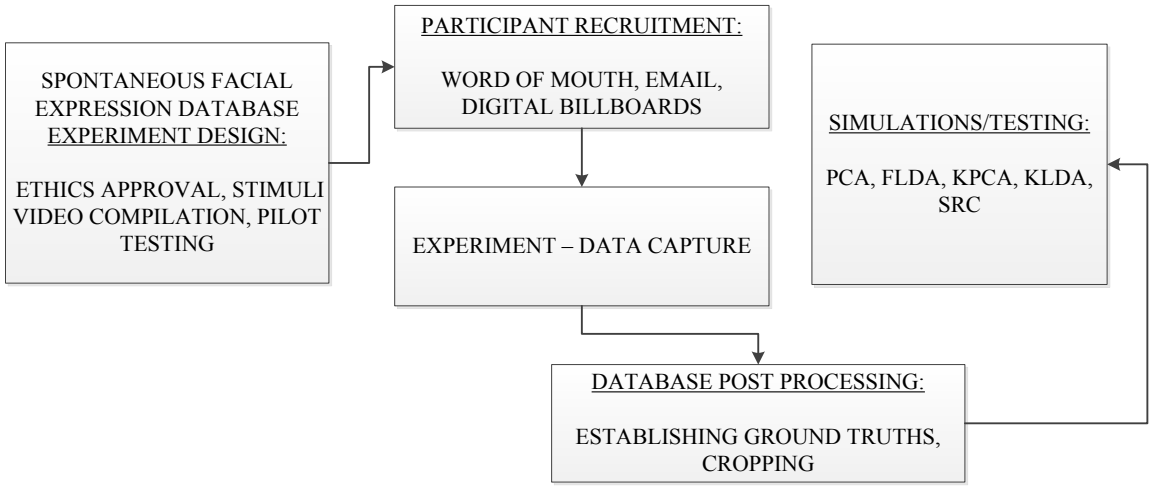


Figure 3–7: Flow Chart/Block Diagram showing the process of the LUSED development.

Participants of the experiment were primarily staff and/or students of Loughborough University. The campaign for participant recruitment was based on word of mouth, the use of fliers and digital display boards around the school building as well as the use of emails to various mailing lists. A website was set up to manage experiment time bookings. In order to preserve the spontaneity of the experiment, subjects were recruited under the auspices of taking part in a memory test otherwise described as an observation test. No participants were offered any kind of compensation for their participation, although this made recruitment more difficult, it guaranteed that participants were not induced to take part for reward.

The stimuli video used to elicit the expressions of the subjects were selected from YouTube and arranged with a view of maximizing the potential to emotionally induce the subjects. For example, the first few clips were funny pranks selected not only to make subjects express happiness through laughter but to relax the participants and distract from the setting of the experiment. The following clips were also pranks making subjects think they were watching more of the same thing, however there were unexpected endings to these pranks often leading to the feeling of surprise, the next clips were prime to inducing expressions of fear as the subjects were now relaxed but still unsure of the direction of coming clips while still oblivious to the true purpose of the experiment. The next set of clips sought to make the participants squirmish leading to the feeling of disgust and the corresponding facial expression. After a largely light-hearted start, a video of animal cruelty was placed towards the end to produce a sober mood with the potential to anger and/or sadden the participants. Two final clips were included; one with a sob storyline and the other a heart-warming yet touching story of soldiers returning to surprise family members, both selected also to induce the expression of sadness. Table 3-2 shows a more detailed description of each video in the order that they appear along with the sought expression and web links.

The length of each of the clips was also of consideration; if the clips were too short and passed by quickly, the subject may not have enough time to process the contents and react accordingly. On the other hand, if the clips were too long, the experiment time would be lengthened making it harder to recruit willing volunteers and creating the likelihood that participants could become bored and lose focus. Feedback from the detailed debrief of 12 subjects who took part in the pilot experiment which was conducted during the design of the experiment was used to achieve a balance in the length and content of the stimuli video. It is worth mentioning at this point that none of the images from the pilot experiment are contained in the database used for evaluations.

The experiments were conducted in the multimedia lab of the School of Electronic, Electrical and Systems Engineering of Loughborough University. Figure 3-8 below is a block diagram that depicts the experiment area while Figure 3-9 shows an image of a mock participant. The room measures 5.6m x 4.5m, and the back wall has a green screen background intended to make it easier for researchers and users to perform background subtraction.

Order	Description	Expression	URL
1	Compilation of funny accidents and blips	HA, SU	http://www.youtube.com/watch?feature=endscreen&NR=1&v=791Dr-VH-Zw
2	Compilation of funny accidents and blips	HA, SU	http://www.youtube.com/watch?v=ZS38N7RhA_0&feature=relmfu
3	Compilation of pranks and practical jokes	HA, SU	http://www.youtube.com/watch?v=Gq10aIXNOeQ
4	Large snake that suddenly attacks the camera	FE, SU	http://www.youtube.com/watch?feature=endscreen&NR=1&v=IogEiKhEJFo
5	Serene scene of landscape with a scary face suddenly appearing	FE	http://www.youtube.com/watch?v=AMj6L9_eHD0&feature=BFa&list=PL596186DC00211F43
6	A doctor pulling out an oversized bogey from patients nose	DI	http://www.youtube.com/watch?v=NRAYccnyfkY
7	Top 10 countdown of disgusting meals	DI	http://www.youtube.com/watch?v=0kWvaxtc_0g
8	A seemingly funny bike accident with gruesome injury at the end	DI, SU	http://www.youtube.com/watch?v=AEnCZPPtkpA
9	Animal cruelty to chicks at a poultry farm	AN, SA	http://www.youtube.com/watch?v=JJ--faib7to
10	Story of a lost loved one	SA	http://www.youtube.com/watch?v=6vjTp9-cj9g
11	Compilation clip of soldiers returning home to surprise loved ones	SA	http://www.youtube.com/watch?v=VKyaif6lgjs&feature=relmfu

Table 3–2: The video clips that make up the stimuli video, the order in which they appear, the expression(s) intended to elicit and the web links for each clip.

Once participants arrived and prior to the start of the experiment, they were given a verbal briefing, stating they are to participate in a memory test/observation experiment during which they will watch a video lasting approximately 19 minutes containing random video clips, they are advised certain measurements such as their pulse will be taken while they watch the video. Participants are then required to sign an *'Informed Consent Form'*, which confirms they have been given information about the impending experiment and that they consent to taking part.

It is worth mentioning at this point that the consent to participate was given on the basis of *'false'* information about the experiment which raises ethical concerns, these are addressed in the 'statement of ethics' contained in Section 3.6. Also worthy of mentioning at this point is the novel approach that led to taking the pulse of participants – during the pilot experiments conducted in the design stages of the experiment, one of the consistent problems experienced was that participants will often occlude their own face with their hands (palms) and thus occlude the expressions being depicted. The use of a portable pulse monitor worn on the right wrist (the hand often placed on the cheek or chin) of each participant provided the justification to ask participants to keep both arms placed on the arms of the chair without unduly arousing their suspicion as to the true purpose of the experiment.

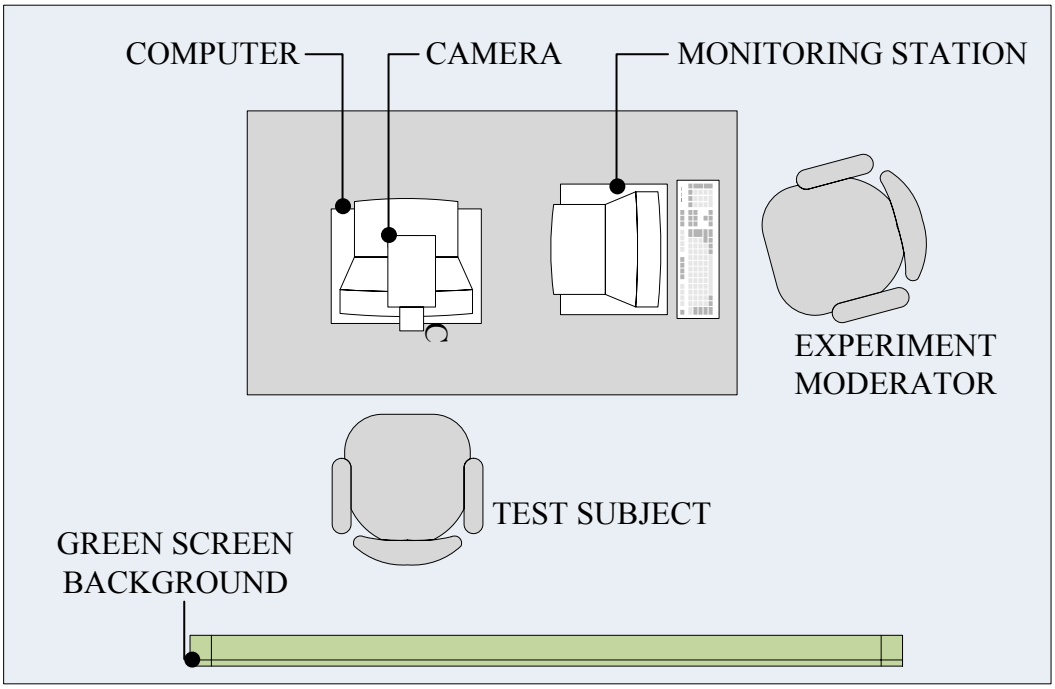


Figure 3–8: Block Diagram showing the experiment area of the Multimedia Room.

To begin the experiment, the participant is invited to sit in the position of the test subject while the experimenter sits in the position of the moderator both depicted in Figure 3-8. The test subject is provided with a pair of headphones. The camera, which is operated from the experimenter's station, is set to start recording. For capture, a mini High-Definition (HD) camera was used; it was placed on top of the monitor but with the recording light disabled. While the participant watched the video, the experimenter remained seated in the room and completed the 'subject information' form which included assigning a 'subject number' to the participant (as required by the ethics committee), subject gender, age group and notes relating to what parts of the video elicited what expressions as observed in real time.

Figure 3–9: Showing a *mock* participant in the experiment setup. (*Inset* is a closer view of blood pressure monitor used to discourage hand movement).

Upon completion of the video, the experimenter advises the participant the true purpose of the experiment while standing behind the participants screen to ensure a frontal view of the participants face is maintained while they receive the news of being deceived, as the camera was still recording, in many cases truly spontaneous facial reactions were captured at this stage. Subjects are then interviewed/debriefed about what emotions they felt at different parts of the video. Subsequently, each subject was asked to attempt to act out each of the prototypical expressions in front of the camera. Lastly, subjects were asked to sign a *'Release Form'* confirming they have been advised the true purpose of the experiment and consent to the use of the recordings for none commercial purposes.

3.6 Statement of Ethics

The procedures above, adopted in the collection of the facial expression database which involved human participants raised some ethical considerations.

The Loughborough University Ethics Approvals (Human Participants) Sub Committee granted ethics approval for the experiment in December 2012 with reference

number R12-P170. In obtaining ethical approval, a research proposal along with an ethical clearance checklist was submitted. These were intended to give assurances that the design of the experiment met with requisite standards; some of the noteworthy issues highlighted and the corresponding mitigations are discussed below.

One of the peculiar concerns were the use of a 'bogus experiment' to recruit participants and the recording of participants without their prior knowledge and explicit approval. This was an integral part of the experiment as subjects needed to be completely oblivious to the intended purpose of the experiment so as to guarantee the spontaneity of the expressions they went on to display, this was mitigated by the use of a carefully worded '*Informed Consent Form*' which contained only general information signed prior to the experiment and a more detailed *'Release Form'* signed at the end of the experiment (both described in more detail in Section 3.5).

Other considerations included the anonymity of the participant's identity, the information they provided during their debrief and their rights to rescind any permissions granted to take part. These concerns were mitigated by firstly adopting a 'subject number – based' coding nomenclature to represent each subject, all their data and documentation, secondly, the signed release form contained a clause stating that participants reserved the right to withdraw from the experiment at any time in the future at which point reasonable steps will be taken to delete all their data. During the experiments, only one participant refused to sign the release form at the end of the experiment and all their information was instantly deleted and their subject number was reallocated to the next volunteer.

The safety and security (physically and psychologically) of both the experimenter and volunteer partly informed the location of the experiment while also ensuring a consistently controlled environment for each participant. The secure location of the experiment and equipment which had restricted access also mitigated concerns about the security of the data collected; a requirement for complying with the Data Protection Act 1998.

Lastly, during the debriefing part of the experiment, apologies were offered for the ominous nature of some of the videos. It was confirmed that participants were not lastingly affected by the emotional induction required to elicit the desired expressions. The workings, aims and objectives of the database were also carefully explained to

participants as well as practical applications of facial expression recognition systems. As well as capturing subject reactions to the true purpose of the experiment, as an ethical requirement, full recordings including the debriefing contents can be made available on request.

3.7 Participant Demography Analysis

During the experiment, data were collected from a total of 125 volunteer participants. One participant chose not to sign the release form at the end of the experiment leaving a total of 124 usable subjects. As mentioned earlier in Section 3.3, one of the main drawbacks of the few existing natural databases that are freely available is the diversity of the demography. As such this section will analyse the demography of the subjects contained in our newly collected LUSED database.

Each participant was recorded as being a member of one of three broad age groups to reduce the amount of unnecessary data being collected. The groups were selected on the basis of what can reasonably be discerned. About 83% of the volunteers were aged 18 – 30 and formed the first group reflective of the fact that they were mostly students, around 15% were aged between 31 – 60 making up the second group, and almost 2% made up the final group of subjects aged 61 and above.

Participants were divided into five ethnic groups; Asian, Black, Caucasian, Mixed and other. ‘Asians’ was the broadest grouping, consisting of Chinese, Pakistanis, Indians, etcetera. As such it was the largest group, making up about 51% of the database, roughly 26% were Black and around 19% belong to the Caucasian group. Around 4% were split between the ‘Mixed’ and ‘Other’ groups.

The most even split in the demography was along gender lines, where 50% of the subjects were male and the other 50% female.

Age Group	Number of Subjects	Percentage %
18 – 30	103	83
31 – 60	19	15
61 & <	2	2
Ethnic Group	**Number of Subjects**	**%**
Asian	63	51
Black	32	26
Caucasian	24	19
Mixed	3	2
Other	2	2
Gender	**Number of Subjects**	**%**
Male	62	50
Female	62	50
Obscurity	**Number of Subjects**	**%**
Glasses	31	25
Hair	18	15
Clothing	2	2
None	65	52
Multiple	8	6

Table 3–3: Analysis of Participant's Demography – showing the number and percentage of participants belonging to each group

As would be the case in a 'real-life' application scenario, little or no emphasis was attached to variables such as head orientation, lighting conditions and so on. In addition, participants were not asked to remove items of clothing or accessories that may have partially obscured parts of their face before taking part in the experiment. As such, at least 25% of the participants were wearing glasses, a minimum of 15% had some part of their frontal face obscured by hair (not including facial hair), clothing e.g. scarfs, burkas obscured some part of the face of 2% of the participants. About 6% had multiple forms of obscurity, for example, a person wearing glasses who also had their face obscured by hair. Roughly 52% had no notable obstruction to the face. Table 3-3, contains a more detailed breakdown of the participant demography.

3.8 Post Processing

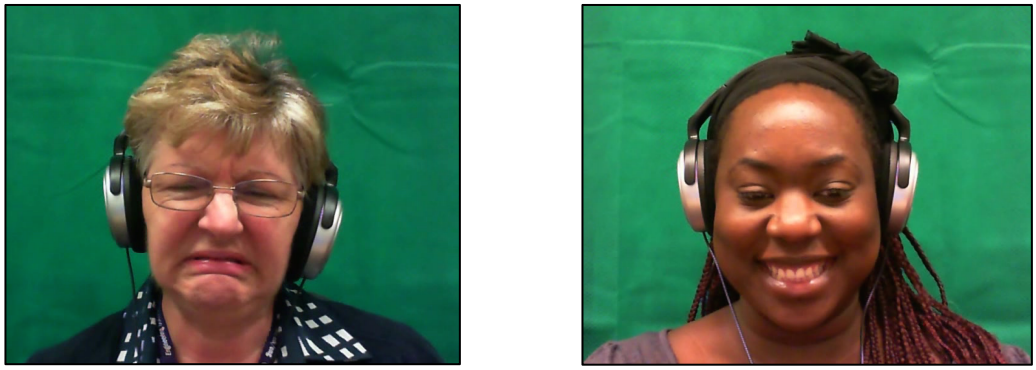
Following the recordings of each participant, the recorded video clip underwent processing prior to being added to the database. Part of this processing included establishing the *ground truth* data for each of the images in the database, i.e. establishing an expression class label for each image. As mentioned in the introduction, having correctly labelled data with which to design and assess facial expression recognition systems is integral to the research community and forms the thrust of this work. As such, this section will discuss the steps taken to prepare the recoded video clip for inclusion in the database. It is worth mentioning that as not all the participants in the experiment displayed facial expressions in all the desired classes, there are fewer images than the total number of participants in some of the classes. The number of images in each class of the database was also affected by the level of difficulty in eliciting some emotional expressions (e.g. Anger) as compared to others (e.g. Happiness).

The video recordings of each of the participants (typically around 25 minutes – including the period of the debrief) was watched by the experimenter and then manually cut up into shorter clips using Corel video-studio pro X6 editing software of a few seconds long where the experimenter observed the subject displayed any of the desired expressions of interest (starting just before the onset and ending just after the dissipation of the given expression).

The experimenter made a decision as to what expression class each of the shorter clips belonged based on the characteristics of the expression using given criteria, some of which are described in Section 3.4. Because the class decision can be said to be somewhat subjective, three added levels of validation were used to assess the experimenters expression class decision; (1) Notes collected during the experiment and debrief were used as a reference, i.e. checking to see if labelled expressions were consistent with any of the emotions the subjects expressed that they experienced. (2) The timing (during the experiment) of the expression was compared to the stimuli video to see if the labelled expression was consistent with the expected reaction in that part of the video clip, i.e. cross checking that subjects smiled/laughed (happy face) during the parts of the stimuli video where funny clips were shown. (3) The final validation was the use of a majority determined three-person judging panel (composed of PhD students from within the department) to allocate each short clip to an expression class without prior knowledge of the experimenter labelled class. Where any two out of the three modes of validation were consistent with the experimenter-labelled class, the short clip retained the existing label while inconsistent labels were re-evaluated.

Once the correct labels were established for each short clip, the video was converted into *.jpg* image files sized 1280 x 720 (around 25 images/expression). Images relating to the *onset* being the beginning of the expression, the *peak* being the image where the expression was most intense and the *end* being the point just after the dissipation of the expression are identified. The *peak* images are the primary images of interest for this work and are added to the database. The other images in the expression sequence are catalogued in the *extended database* and will be made available for researchers interested in the evolution of an expression.

Following the processing and subsequent labelling of all the captured expressions, the final step before a performance analysis of the database was to manually crop each image in the database.



(a) (b)

Figure 3–10: Showing samples of the original .jpg images from the LUSED after establishing the ground truths but before post processing. (a) is a subject from the Disgust class while (b) is from the class relating to Happiness.

(a) (b)

Figure 3–11: Showing samples of the subjects in Figure 3-10 after post processing – cropping and conversion to greyscale.

As mentioned earlier, not all of the 124 consenting participants displayed expressions in all the desired basic classes of interest (anger, disgust, fear, happiness, sadness and surprise). In addition it was concluded that the inducing of expressions relating to anger were unsuccessful and thus those images were excluded from the database. This was determined due to lack of expressions consistent with the characteristics of anger and supported by feedback from participants, many of whom felt they were not genuinely angered. Each subject contributed only one image to a given class ensuring no false positives when using the database to test algorithms. The main database contains a total of 559 images over six classes broken down as follows;

Disgust – 102, Fear – 43, Happiness – 121, Neutral – 124, Sadness – 71 and Surprise – 98 images. The percentage of the images in each class as a function of the total number of participants (124) is indicative of the difficulty in inducing some of the expressions spontaneously. This is graphically illustrated in Figure 3-12 where it can be seen that the expression of happiness had the most images – with approximately 98% of participants eliciting an expression while only around 35% displayed an expression corresponding to Fear.

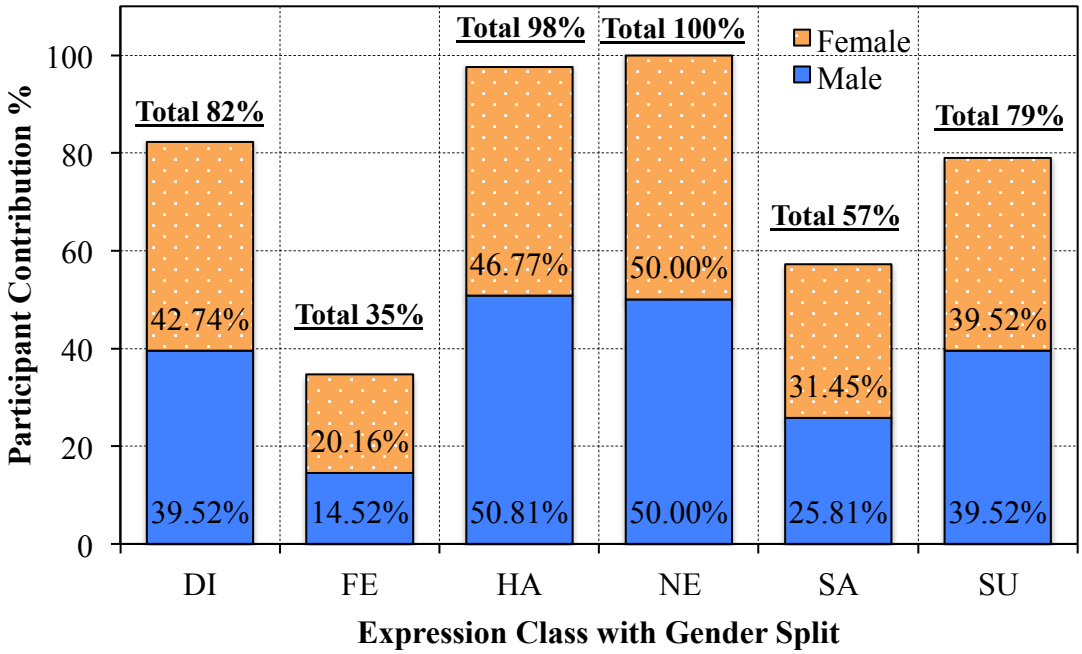


Figure 3–12: Number of images from each class (split according to gender) contained in the final database – expressed in percentage.

The total percentages from Figure 3-12 above are subdivided by gender however Table 3-4 below contains a more detailed breakdown of the demography of participants included in the database.

Demographic Group		CLASS %					Database Average
		DI	FE	HA	SA	SU	
AGE	18 – 30	87	53	84	72	79	75.09%
	31 – 60	11	42	14	25	19	22.29%
	61 & <	2	5	2	3	2	3.62%
ETHNICITY	Asian	58	49	52	52	57	53.60%
	Black	27	28	26	18	22	24.51%
	Caucasian	10	21	17	27	15	18.03%
	Mixed	3	2	2	0	3	2.16%
	Other	2	0	2	3	2	1.69%
GENDER	Male	48	42	52	45	50	47.41%
	Female	52	58	48	55	50	52.59%

Table 3–4: Analysis of Participant's Demography – showing the number and percentage of participants belonging to each group.

The extended database, made up of the sequence of images before and after each peak expression, contains around 14,000 images and will be made available alongside the main database.

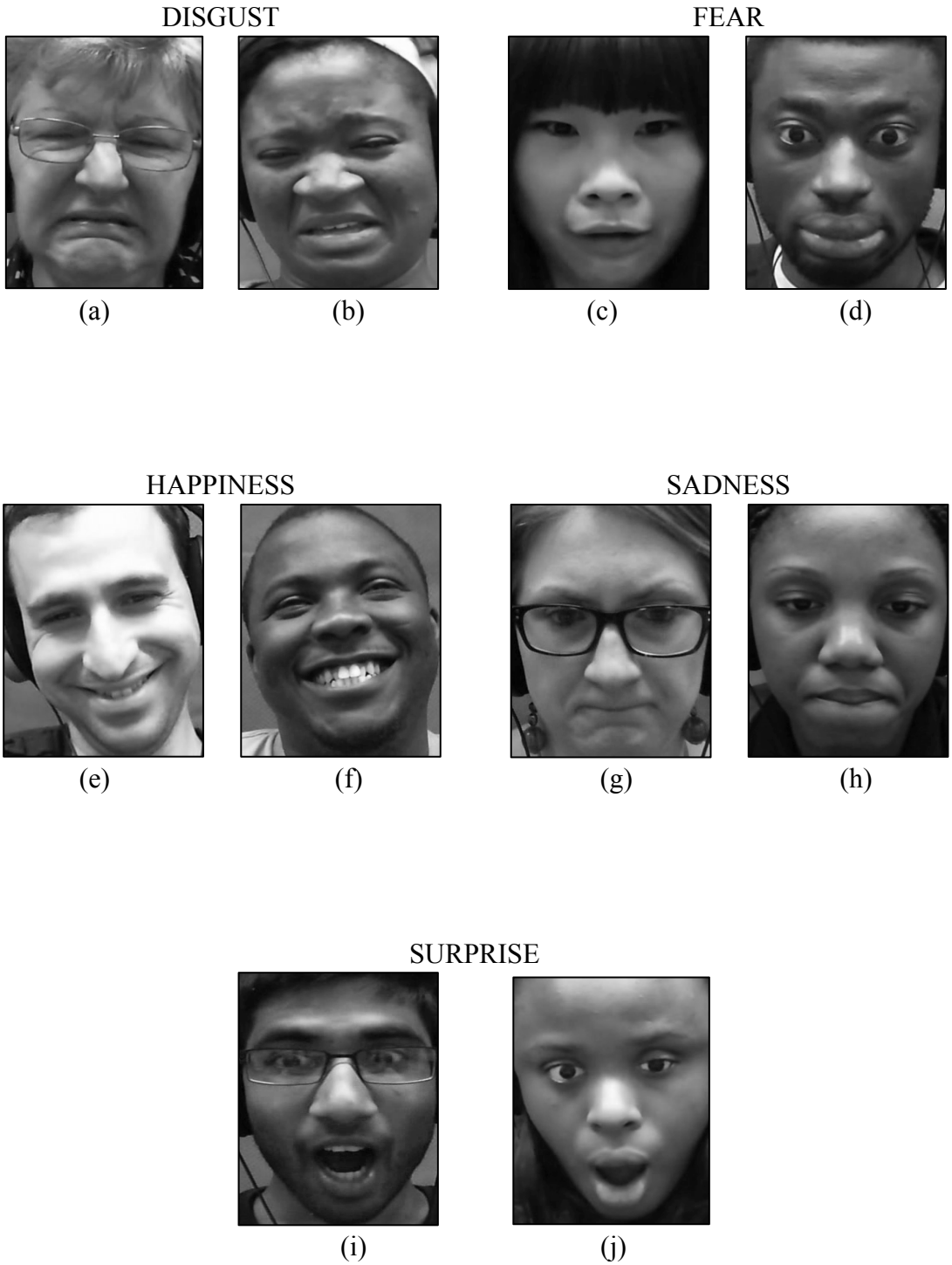


Figure 3–13: Showing two sample images from each of the classes contained in the LUSED.

3.9 Summary

Traditionally, Facial Expression Recognition (FER) has been performed using laboratory-controlled data (posed databases), while undoubtedly worthwhile at the time, such lab controlled data poorly represent the environment and conditions faced in real-world situations. One of the main challenges faced in FER research is the access to well-annotated, publicly available spontaneous databases. Where they exist, they suffer from drawbacks some of which are highlighted in Section 3.3.

In this chapter, the newly collected Loughborough University Spontaneous Expression Database (LUSED) was presented, which (1) Contains labelled data for five out of the six basic expressions and will be made publicly available for the development and assessment of facial expression recognition systems. (2) Will include an extended database containing images corresponding to the evolution of each expression in the main database. (3) A comprehensive and up to date review of existing spontaneous (natural) facial expression databases.

The contribution this database represents to the community will further bolster the pace and quality of research in the area of natural expression recognition leading to more practical deployments of such systems. The database can be accessed via [http://www.lboro.ac.uk/departments/eese/research/communications/lused.html].

In addition to having labelled data, a valuable addition to the community is baseline results with which researchers can compare results when algorithms are being developed. To this end, Chapter Four and Chapter Five of this Thesis evaluate and validate the database discussed in this chapter in comparison with other existing databases.

References

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, 2003.

[2] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[3] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, 1992.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Second Ed. John Wiley & Sons, Inc., 2001.

[5] M. Lyons and S. Akamatsu, "Coding facial expressions with gabor wavelets," *Autom. Face Gesture*, pp. 200–205, 1998.

[6] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimed.*, vol. 12, no. 7, pp. 682–691, 2010.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[8] P. Ekman and E. L. Rosenberg, *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system*, Second Edi. Oxford University Press, 2005.

[9] J. Cohn and K. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *Int. J. Wavelets*, pp. 1–12, 2004.

[10] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," *Proc. ninth Int. Conf. Multimodal interfaces - ICMI '07*, p. 38, 2007.

[11] "Vam." [Online]. Available: http://emotion-research.net/download/vam. [Accessed: 21-Sep-2015].

[12] E. Douglas-Cowie, "D5i: Final report on WP5 - HUMAINE human-machine interaction network on emotions," Belfast, 2008.

[13] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, "Spontaneous emotional facial expression detection," *J. Multimed.*, vol. 1, no. 5, pp. 1–8, 2006.

[14] T. Kanade, J. F. J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Autom. Face Gesture Recognition, 2000. Proceedings. Fourth IEEE Int. Conf.*, 2000.

[15] P. Lucey, J. Cohn, and T. Kanade, "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," *Comput. Vis. Pattern Recognit. Work. (CVPRW), 2010 IEEE Comput. Soc. Conf. on. IEEE.*, pp. 94–101, 2010.

[16] "Yale" [Online]. Available: http://vision.ucsd.edu/content/yale-face-database. [Accessed: 27-Nov-2014].

[17] S. Du, Y. Tao, and A. Martinez, "Compound facial expressions of emotion," *Proc. Natl. Acad. Sci.*, 2014.

[18] K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven, "The MPI facial expression database - a validated database of emotional and conversational facial fxpressions," *PLoS One*, vol. 7, no. 3, p. E32321, 2012.

[19] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *2005 IEEE Int. Conf. Multimed. Expo*, pp. 317–321, 2005.

[20] M. Pantic and M. Bartlett, *Machine analysis of facial expressions*. I-Tech Education and Publishing, 2007.

[21] K. Scherer and G. Ceschi, "Lost luggage : a field study of emotion-antecedent appraisal," *Motiv. Emot.*, vol. 21, no. 3, pp. 211–235, 1997.

[22] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[23] P. Trinh and J. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, 2013.

[24] R. Gajšek, V. Štruc, and F. Miheli, "Multi-modal emotional database : AvID," *Informatica*, vol. 33, no. 1, pp. 101–106, 2009.

[25] A. Tcherkassof and D. Dupré, "DynEmo: A video database of natural facial expressions of emotions," *Int. J. Multimed. Its Appl.*, vol. 5, no. 5, pp.

61–80, 2013.

[26] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: the UNBC-McMaster shoulder pain expression archive database," in *Face and Gesture*, pp. 57–64, 2011.

[27] G. I. Roisman, J. L. Tsai, and K.-H. S. Chiang, "The emotional integration of childhood experience: physiological, facial expressive, and self-reported emotional response during the adult attachment interview," *Dev. Psychol.*, vol. 40, no. 5, pp. 776–789, 2004.

[28] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): naturalistic and spontaneous facial expressions collected 'in-the-wild,'" in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 881–888, 2013.

[29] E. Douglas-Cowie, R. Cowie, and M. Schroeder, "The description of naturally occurring emotional speech," *Proc. 15th Int. Congr. Phonetic Sci. Barcelona*, pp. 2877–2880, 2003.

[30] "DISFA," *Denver Intensity of Spontaneous Facial Action (DISFA) Database*, 2013. [Online]. Available: http://www.engr.du.edu/mmahoor/DISFA.htm. [Accessed: 27-Nov-2014].

[31] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 568–573, 2005.

[32] "Ermis" [Online]. Available: http://www.image.ntua.gr/ermis/. [Accessed: 11-Dec-2014].

[33] "SmartKom" [Online]. Available: http://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom; [Accessed: 15-Mar-2015].

[34] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image Vis. Comput.*, vol. 25, no. 12, pp. 1856–1863, 2007.

[35] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi, "A video database of moving faces and people," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 812–6, 2005.

[36] P. Ekman, "Universal and cultural differences in facial expressions of emotion," *Nebraska Symp. Motiv.*, 1971.

[37] P. Ekman, *Emotion in the human face*, Second Edi. Cambridge University Press, 1982.

[38] P. Ekman, "Facial expression and emotion," *Am. Psychol.*, vol. 30, pp. 527–554, 1993.

[39] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique," *Am. Psychol. Assoc. Inc.*, vol. 115, no. 2, pp. 268–287, 1994.

[40] P. Ekman and W. V. Friesen, *Unmasking the face: a guide to recognizing emotions from facial expressions*. Prentice-Hall, INC., Englewood Cliffs, NJ, 1975.

Chapter Four

Baseline Evaluation: In the Plain and Encrypted Domains

4.1 Introduction

Further to the need for a labelled dataset for the spontaneous Facial Expression Recognition (FER) problem is the need for well-established bases by which subsequent attempts at recognition can be measured. A consideration of spontaneous FER, like other image classification problems is that of excessive dimensionality and one approach to this problem is to reduce dimensionality by combining features. 'Linear combinations are particularly attractive because they are simple to compute and are analytically tractable' [1]. Effectively, linear methods form a smaller representation of an image as a function of the whole image. In other words, given an image, linear methods project the higher dimensional data in the *plain* space onto a lower dimensional *feature* space. There are two classical approaches to finding effective linear transformations [1]. The first being Principal Component Analysis (PCA), which obtains a projection that best represents the data in a least-squares sense. The second classical approach is generally dubbed Multiple Discriminant Analysis (MDA) – which obtains a projection that best separates the data in a least-squares sense, an instance of which is Fisher's Linear Discriminant Analysis. The primary difference between the two approaches is that while PCA seeks to best represent the data in the feature space, FLDA will seek to best discriminate the data, noting that the best feature representation

of data (PCA) may not necessarily be the best for discriminating between data in different classes.

Scholkopf et al. extended the classical PCA approach to Kernel Principal Component Analysis (KPCA) in [2]. They showed empirically that KPCA is able to extract non-linear features and thus provided better results on digit recognition. Baudat et al., Roth et al., and Mika et al. [3, 4] then applied the kernel extension to the classical FLDA approach aptly dubbed KFDA. Their experiments showed that KFDA is able to extract the most discriminant features in the feature space, which is equivalent to extracting the most discriminant non-linear features in the original plain space.

On the basis of the above, in this chapter, the new database – LUSED; introduced in Chapter Three will be evaluated using established historic methods for image classification problems (PCA, FLDA), which are known to have been applied to the FER problem [5–11] some of which include the spontaneous FER case. The novel application of the kernel variants – KPCA and KFDA to the spontaneous FER problem is also documented along with a practical application case scenario.

The process flow of the evaluation process is depicted in Figure 4-1, which is generally representative of the steps required for the pattern classification problem.

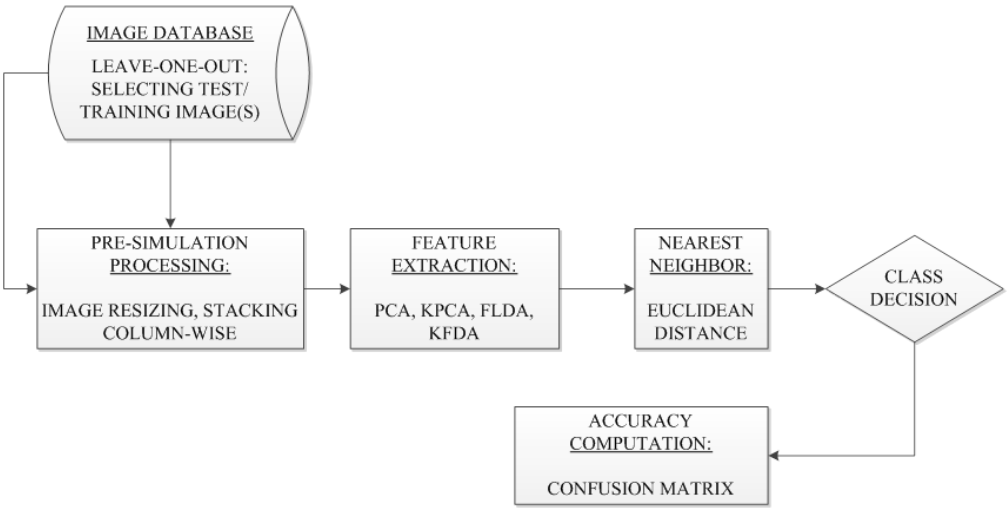


Figure 4–1: Flow Chart/Block Diagram showing the steps in the evaluation of the LUSED.

4.1.1 Statement of Novelty

This chapter validates the new database; LUSED, presented in Chapter Three. The database is evaluated using Principal Component Analysis (PCA), Fisher's Linear Discriminant Analysis (FLDA) and their kernel variants; KPCA and KFDA combined with a basic classifier. The benchmark results obtained from the analyses using PCA and FLDA combined with the comparisons with existing literature contribute a valuable resource to the facial expression recognition community, while the evaluations using KPCA and KFDA represent the first known application of the kernel variants to spontaneous FER. In addition, a conceptual model that addresses privacy issues related to the practical application is verified. This was achieved by a scheme that allows the recognition of natural facial expressions in the encrypted domain and represents a significant and novel contribution to knowledge. More clearly, the following are presented in this chapter:

1. Evaluation of LUSED using PCA, KPCA, FLDA and KFDA.

2. Verification of a new scheme for performing facial expression recognition in the encrypted domain.

3. Comparative analysis of LUSED with another natural expression database (NVIE).

The remainder of this chapter is organized as follows; Section 4.2 contains background on the use of the baseline methods for facial expression analysis. The mathematical theory of the algorithms used in this chapter's evaluation of LUSED is presented in Section 4.3, including both feature extraction and classification methods. Section 4.4 contains a model for a practical implementation of FER, which mitigates privacy concerns. The setup of the simulation is described in Section 4.5 while results and corresponding discussions are presented in Section 4.6. A summary of the chapter is given in Section 4.7.

4.2 Background Review of Baseline Evaluation Methods

The background of evaluating facial expression databases is largely based on the same methods and techniques used in facial recognition. This is not unexpected as correctly classifying the facial expression of a test image given training samples from the different expression classes is a similar pattern classification problem mathematically to correctly classifying the identity of a test image given training samples. However facial recognition is an area of image processing that is far more researched than the recognition of facial expressions and it is evident from the literature that some of the methods used for evaluation in this chapter; particularly PCA and FLDA are quite consistent for use in the validation of databases and baseline result comparisons (as will be seen in this section). In addition, they have formed the basis for the development of other algorithms. It is however worth pointing out that this does not suggest PCA and/or FLDA are the most superior in performance.

Perhaps the simplest scheme with regards to the classification problem is a nearest neighbour classifier in the image space [12]. In this scheme, a test image is classified by assigning it the label of the closest image from the training data. This 'closeness' is represented by the smallest Euclidean distance measure between the pixel values of the actual test image and the pixel values of each of the images in the training set (in the image space). This process is in effect selecting the image from the training set that best *correlates* with the test image assuming all the images have zero mean and unit variance (if the images are normalized). However, as articulated in [13], this process referred to as correlation has a number of notable drawbacks; firstly, it lacks robustness in that if the lighting of training images and test images vary from each other, then the pixel values that are supposed to be compared to each other may not be closely clustered enough to allow correct recognition. Secondly and more important practically, is the large storage requirements and excessive computational expense associated with classifying in the image space. For example, the algorithm described in [12], which was implemented by the authors of [14], required them to develop special purpose Very-Large-Scale-Integration (VLSI) hardware.

The above, formed part of the justification for pursuing dimensionality reduction schemes for image processing generally and FER by extension. One of such schemes is PCA also related to the Karhunen-Loève theorem [13]. It can be defined as 'an

orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on' [15]. It is the most widely used method for feature extraction in the area of face recognition and classification of facial images based on various properties including but not limited to facial expression, age and gender. According to work carried out by O'Toole, A.J., Abdi, H. and others, [5–15] PCA has been shown to encode the relevant features of the face. Also significant to the success of PCA is that Abdi, O'Toole and Deffenbacher amongst others [10–15] all imply that PCA extracts facial features in a way that is consistent with the way we as humans perceive faces. It is worth noting most of these studies use relatively small datasets with less variation and consider the data in regards to only one or two properties of faces e.g. identity and/or expression.

Buchala et al. [27] used a much larger number of faces and tested if PCA could encode properties such as gender, ethnicity, age, and identity efficiently. They employed the use of FLDA to determine what component(s) performed best for each of the different properties being examined, in addition, they analysed how these different properties varied on the different components of PCA. They found that PCA encodes facial image properties efficiently. It was also asserted that properties of the face are encoded by different principal components.

Calder et al. [17] used PCA in the analysis of facial expressions. They applied PCA to the pixel intensities of images showing facial expressions, also using FLDA; they found that a PCA-based system could be successfully used for the task of facial expression recognition. They also highlighted the difference in the encoding power of each component for different properties where it is stated that 'components for coding facial expression information are largely different to components for facial identity information' [17]. It is worth noting at this point that determining the best eigen vectors for encoding facial properties such as expression is a process that is more subjective than objective and best carried out experimentally. In Section 4.4, experimentally obtained assertions are made as to the best eigen vectors for spontaneous FER. The common method for selecting the components to define the data that represent an image without significant loss of information is to order the components based on their importance in representing the variance of the data where the most important will be the

component that represents the most variance and consider the first few components which represent a percentage (normally over 80%) of the cumulative variance. However, if components that are not significant in defining the data but represent the features relating to the properties of interest are located within the last few components, then PCA will be ineffective because facial image data usually encode multiple properties.

The concept of PCA encoding characteristics is well illustrated in Buchala's example, 'the data shown in Figure 4-2 below have different properties and can be classified by colour – orange, blue, by shape – circle, square, or both colour and shape. PCA on the data would indicate maximum variance in the direction of $\boldsymbol{W}_1$. This first Principal Component (PC) encodes only shape information and it fails if the property of interest is colour, whereas the second component in the direction of $\boldsymbol{W}_2$, though it accounts for lesser variance, would be successful in this regard. This shows that the selection of the components should be based on its importance for a given task, rather than its importance in accounting for the total variance. This example would be apt for face data which have multiple properties, such as identity, gender, ethnicity, age and expression' [27].

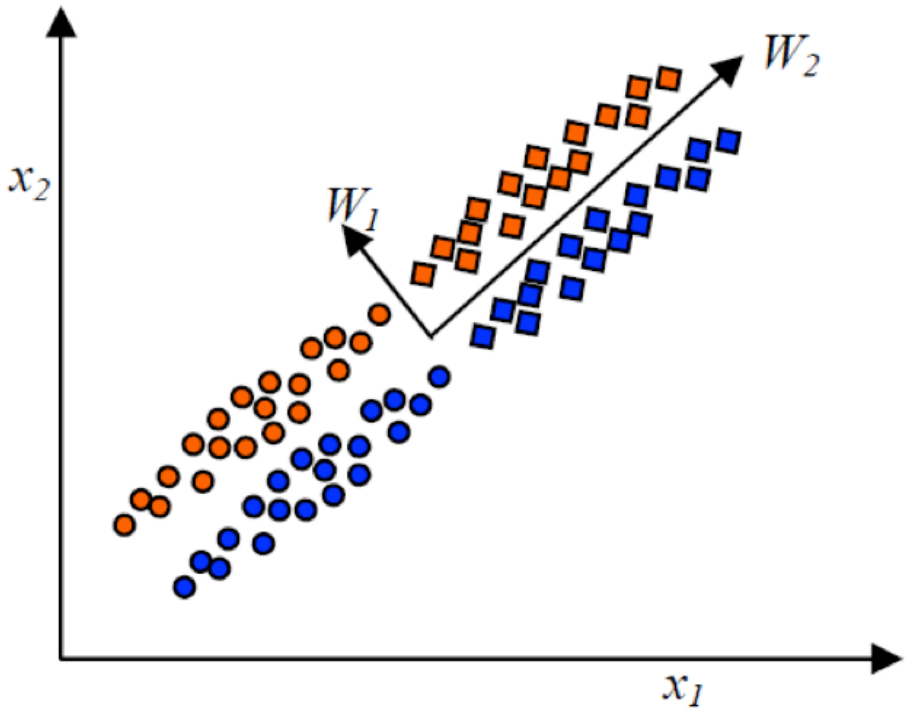


Figure 4–2: An illustration of a dataset with multiple properties [27].

The first component in the direction of $\boldsymbol{W}_1$, despite accounting for the maximum variance, fails to encode information regarding the colour of the data, whereas the second component in the direction $\boldsymbol{W}_2$, would be successful.

Other algorithms have come close to PCA in repute, for example, Independent Component Analysis (ICA), which maximizes the degree of statistical independence among output variables using contrast functions. Although Draper et al. [28] found ICA to be better than PCA for the representation of faces (feature extraction) when cosines were used as a measure of similarity which was backed up by Bartlett et al. [29], the improvement ICA offers comes at a significantly higher computational cost as was discovered by Yang [30] who also used Kernel PCA for face feature extraction and face recognition and showed that the Kernel Eigenfaces method outperforms the conventional PCA (Eigenfaces) method. Yang's results showed that the ratio of the computation time required by ICA, Kernel PCA, and PCA is, on average, 8.7: 3.2: 1.0 showing that ICA is over three times as demanding as PCA and twice in the case of KPCA with respect to time. Barlett et al. [29] also found implementations of ICA produce marginally better results than PCA.

Further work has been done to try and improve the efficiency of PCA. In 2004, Yang et al. [31] proposed a new method dubbed 'Two-dimensional principal component analysis (2DPCA)'. As opposed to conventional PCA, 2DPCA is based on two-dimensional matrices rather than one-dimensional vectors, the primary difference between PCA and 2DPCA being that, with conventional PCA, the two dimension image matrices need to be transformed into a vector before processing while in 2DPCA, the covariance matrix is constructed directly using the original bi-dimensional image matrices. 2DPCA has a much smaller sized image covariance matrix in comparison to the covariance matrix of PCA as such making it computationally easier to construct the image covariance matrix accurately. It is noteworthy however that in the practical application of PCA to a dataset, a surrogate matrix is used (more details in Section 4.5 – Simulation Setup). More recently, the 2DPCA model was applied to FER in [32] where a model was proposed to firstly detect a face and subsequently recognise the facial expression. Similarly, [33] proposed two-dimensional Uncorrelated Linear Discriminant Analysis (2D-ULDA) for FER as a variant to FLDA to mitigate ULDA's ineffectiveness when faced with the *small sample size* problem – where the number of training samples is less than the dimension of the feature vectors.

Finally, it interesting to note once again that the majority of the literature reviewed in this section that applies PCA-based methods to FER employs the use of posed facial expression databases. Exceptions include [34] where the AAI database was used.

4.3 Theory of the Algorithms

The line between the tasks of feature extraction and classification is somewhat blurred. As articulated in [1], 'the conceptual boundary between feature extraction and classification proper is somewhat arbitrary: an ideal feature extractor would yield a representation that makes the job of the classifier trivial'. As such varieties of methods exist and have been applied for feature extraction with the aim of increasing separability between each of the proposed expression classes. This section contains the mathematical theory of the feature extraction techniques employed, as well as the *trivial* classifier.

4.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) characterizes objects to be measurable by machines and reduces the dimensionality of objects by combining features. This is advantageous due to the computational expense, amongst other drawbacks of *correlation*-based methods. Effectively, PCA chooses a dimensionality reducing linear projection that maximizes the scatter matrix [35]. To achieve that, a scatter or covariance matrix of all the training samples is formed.

Mathematically, assume a set of N training images from a facial expression dataset $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ taking values in an n-dimensional image space. Given ground truth class labels that assign each of those images to one of C classes $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_C\}$. Also consider a linear transformation that maps the original n-dimensional image space into an m-dimensional feature space, where $m < n$. The new lower dimensional feature space representation $\boldsymbol{y}_k \in \mathbb{R}^m$ of a given image space vector $\boldsymbol{x}_k$, is given by:

$$\boldsymbol{y}_k = \boldsymbol{W}^T \boldsymbol{x}_k \quad k = 1,2, \dots, N \tag{4.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ is a matrix with orthogonal columns. It is given that the covariance matrix $\boldsymbol{S}_T$ is formed by:

$$\boldsymbol{S}_T = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T \tag{4.2}$$

given that $\boldsymbol{x}_k$ is the column vector created by reshaping each original image (stacking the columns of each image to form one long column), $\boldsymbol{\mu}$ is the mean of all $\boldsymbol{x}_k$, N is the total number of training samples, and $(\cdot)^T$ denotes the transpose operation. Applying linear projection $\boldsymbol{W}^T$, the scatter of the projected feature vector $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$ can be rewritten as $\boldsymbol{W}^T \boldsymbol{S}_T \boldsymbol{W}$. For the PCA the projection matrix $\boldsymbol{W}_{opt}$ is chosen to maximize the determinant of the scatter of the projected feature vector:

$$\boldsymbol{W}_{opt} = arg \max_{\boldsymbol{W}} |\boldsymbol{W}^T \boldsymbol{S}_T \boldsymbol{W}| = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_m] \tag{4.3}$$

where $\{\boldsymbol{w}_i | \; i = 1,2, \dots, m\}$ is a set of n-dimensional eigen vectors of $\boldsymbol{S}_T$ corresponding to the m largest eigenvalues.

A known drawback of PCA is that the covariance (scatter) matrix being maximized during the eigen decomposition is due not only to the between-class scatter that increases the ease of separability of classes, but also the within-class scatter. It is noted in [36] that 'much of the variation from one image to the next is due to illumination changes'. It can therefore be said that PCA also lacks robustness to images with a high variance in lighting conditions (albeit not as much as the *correlation* method mentioned above). Based on the comments in [36] and similar assertions, it has been suggested that the first few most significant principal components be discarded to reduce the variation in lighting. Though it is plausible that these first few principal components capture the lighting variation, suggesting that if they (the first few PCs) are ignored, it will allow for better clustering when projected. It is however unlikely that these ignored principal components do not encode any other information useful for separability. Such are the factors that make eigen vector selection a subjective process, assertions on which are experimentally obtained and discussed later in this chapter.

4.3.2 Kernel Principal Component Analysis (KPCA)

Kernel variants of the two algorithms mentioned above (PCA and FLDA) were also used for feature extraction. The first of the two variants is Kernel-Principal Component Analysis KPCA. In general, KPCA maps data into a nonlinear space $\boldsymbol{R}^F$, which contains more information due to its higher order correlations [2], which enables the extraction of nonlinear features, thereby achieving better classification performance [30, 37]. The nonlinear mapping function for this projection is given by $\Phi: \boldsymbol{R}^M \rightarrow \boldsymbol{R}^F$ where $\boldsymbol{R}^M$ is the original data space. In space $\boldsymbol{R}^F$, the covariance matrix of $\Phi(\boldsymbol{x})$ is then given by:

$$\boldsymbol{S}_T^{\Phi} = \sum_{k=1}^{N} \Phi(\boldsymbol{x}_k)\, \Phi(\boldsymbol{x}_k)^T \tag{4.4}$$

where $\Phi(\boldsymbol{x}_k)$ is the projection of image $\boldsymbol{x}_k$ in feature space $\boldsymbol{R}^F$. Notice that the dimensionality of the feature space $\boldsymbol{R}^F$ can be arbitrarily large [2], [37]. The eigen-decomposition problem of $\boldsymbol{S}_T^{\Phi}$ can be simplified as: $\boldsymbol{KX\alpha} = \mathrm{N}\lambda\boldsymbol{\alpha}$, and the projection matrix is given by:

$$\boldsymbol{W}^{\Phi} = \sum_{i=1}^{N} \alpha_i \Phi(\boldsymbol{x}_i) \tag{4.5}$$

where $\alpha_i, \ldots, \alpha_N$ are the elements of vector $\boldsymbol{\alpha}$. With the projection matrix $\boldsymbol{W}^{\Phi}$, we can then project the images in $\boldsymbol{R}^F$ to a dimensionally reduced space. Assuming that $\boldsymbol{x}_k$ is a sample image, its projection via the kernel method is then given by:

$$\boldsymbol{W}^{\Phi}\Phi(\boldsymbol{x}_k) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_k) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i k(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{4.6}$$

Note that whether or not $\boldsymbol{x}$ is mean-centred in its original space, there is no guarantee that it has zero mean in space $\boldsymbol{R}^F$. Therefore, the mean-centred version of the kernel matrix is given as:

$$\boldsymbol{Kx'} = \boldsymbol{Kx} - \mathbf{1}_N\boldsymbol{Kx} - \boldsymbol{Kx}\mathbf{1}_N + 1_N\boldsymbol{Kx}\mathbf{1}_N \tag{4.7}$$

where $\mathbf{1}_N$ is an $N \times N$ matrix whose elements take the value of $1/N$.

4.3.3 Fisher's Linear Discriminant Analysis (FLDA)

FLDA is a class specific method in that it attempts to structure the scatter to make it more consistent for classification. While PCA provides a basis for feature extraction by performing dimensionality reduction with a linearly selected projection matrix, the dimensional reduced images are not always perfectly linearly separable. As such, FLDA was introduced, this transforms the original database into a dimension reduced space where the differences between classes are maximised and the differences within each class are minimised. To achieve that, let the between-class scatter matrix be:

$$\boldsymbol{S}_B = \sum_{i=1}^{c} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{4.8}$$

and the within-class scatter matrix be defined as:

$$\boldsymbol{S}_W = \sum_{i=1}^{c} \sum_{\boldsymbol{X}_k \in \boldsymbol{X}_i} (\boldsymbol{X}_k - \boldsymbol{\mu}_i)(\boldsymbol{X}_k - \boldsymbol{\mu}_i)^T \tag{4.9}$$

where $\boldsymbol{\mu}_i$ is the mean of class $\boldsymbol{X}_i$, $\boldsymbol{\mu}$ is the mean of the entire training set, N_i is the number of samples in class $\boldsymbol{X}_i$ and c is the number of classes. The optimal projection matrix is formed of orthogonal columns which maximise the ratio of the determinant of $\boldsymbol{S}_B$ to the determinant of $\boldsymbol{S}_W$, given as:

$$\boldsymbol{W}_{opt} = arg\max_{\boldsymbol{W}} \frac{|\boldsymbol{W}^T\boldsymbol{S}_B\boldsymbol{W}|}{|\boldsymbol{W}^T\boldsymbol{S}_W\boldsymbol{W}|} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_m] \tag{4.10}$$

where $\{\boldsymbol{w}_i | \; i = 1,2, \dots, m\}$ is the set of generalized eigenvectors of $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ which correspond to the m largest eigenvalues $\{\lambda_i | \; i = 1,2, \dots, m\}$.

To practically obtain the optimal FLDA projection matrix, PCA is first applied; as such, the FLDA solution is given by:

$$\boldsymbol{W}_{opt} = \boldsymbol{W}_{pca}\,\boldsymbol{W}_{flda} \tag{4.11}$$

where $\boldsymbol{W}_{pca}$ is given by the optimal PCA projection matrix in equation (3) and $\boldsymbol{W}_{flda}$ is given by:

$$\boldsymbol{W}_{flda} = arg\max_{\boldsymbol{W}} \frac{\left|\boldsymbol{W}^T\boldsymbol{W}_{pca}^T\boldsymbol{S}_B\boldsymbol{W}_{pca}\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{W}_{pca}^T\boldsymbol{S}_W\boldsymbol{W}_{pca}\boldsymbol{W}\right|} \tag{4.12}$$

The resulting optimal projection matrix $\boldsymbol{W}_{opt}$ is used to transform a given image $\boldsymbol{x}_k$ into the feature space.

4.3.4 Kernel Fisher's Discriminant Analysis (KFDA)

The second of the kernel variants is Kernel-Fisher Discriminant Analysis (KFDA). While the FLDA method solves Fisher's discriminant problem in linear space, the KFDA solves it in nonlinear feature space $\boldsymbol{R}^F$. Effectively, because the extra nonlinear features can easily be extracted in space $\boldsymbol{R}^F$, applying the kernel variant produces better results. Similar to the process of the FLDA method, all the scatter matrices of the KFDA will have KPCA pre-applied for dimensional reduction [38]. Let $\boldsymbol{S}_B^{\Phi}$ denote the between-class scatter matrix and $\boldsymbol{S}_W^{\Phi}$ denote the within-class scatter matrix, which is defined as Fisher's discriminant function in feature space $\boldsymbol{R}^F$ and is given by:

$$\boldsymbol{W}_{opt}^{\Phi} = arg\max_{\boldsymbol{W}} \frac{\left|\boldsymbol{W}^T\boldsymbol{S}_B^{\Phi}\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_W^{\Phi}\boldsymbol{W}\right|} = [\boldsymbol{w}_1^{\Phi}, \boldsymbol{w}_2^{\Phi}, \ldots, \boldsymbol{w}_m^{\Phi}] \tag{4.13}$$

where, $\boldsymbol{S}_B^{\Phi}$ and $\boldsymbol{S}_W^{\Phi}$ are respectively defined by:

$$\boldsymbol{S}_B^{\Phi} = \sum_{j=1}^{c} l_j(\boldsymbol{M}_j - \boldsymbol{M}_0)(\boldsymbol{M}_j - \boldsymbol{M}_0)^T \tag{4.14}$$

and

$$\boldsymbol{S}_W^{\Phi} = \sum_{j=1}^{c} \boldsymbol{K}_j(\boldsymbol{I} - \boldsymbol{1}_{lj})\boldsymbol{K}_j{}^T \tag{4.15}$$

where $(\boldsymbol{K}_j)_{nm} = k(\boldsymbol{X}_n, \boldsymbol{X}_m^j)$. The number of images in class $\boldsymbol{X}_j$, is l_j. The mean of the entire training set is:

$$(\boldsymbol{M}_0)_i = \frac{1}{N}\sum_{k=1}^{N} k(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{4.16}$$

and the mean of class $\boldsymbol{X}_i$ is:

$$(\boldsymbol{M}_j)_i = \frac{1}{1_j}\sum_{k=1}^{l_j} k(\boldsymbol{X}_i, \boldsymbol{X}_k^j) \tag{4.17}$$

The above allows the projection of the training and test images into the feature space using the optimal projection matrix $\boldsymbol{W}_{opt}^{\Phi}$ prior to classification using the nearest neighbour classifier.

4.3.5 Nearest Neighbour Classifier (NN)

The Nearest Neighbour (NN) classifier is often used because of its simplicity and robustness particularly in the hope that the given feature extraction technique applied prior to classification would have effectively clustered the data into classes in the feature space thereby making the classification problem trivial. NN uses a Euclidean distance measure between the elements of the projections of the test samples and the elements of the projections of training samples to judge their correlation. The projection of the test sample $\boldsymbol{s}$ is said to belong to the same expression class as the projection of training image $\boldsymbol{t}$ for the lowest value of the Euclidean distance given by:

$$D(samp, training) = \sqrt{\sum_{i=1}^{n} (s_i - t_i)^2} = \|\boldsymbol{s} - \boldsymbol{t}\|_2 \tag{4.18}$$

where for a given comparison (between test image and training image projections), n is the number of elements in the projected images. Following the classification of all test images, the accuracy of the used algorithm is computed.

4.4 FER in the Encrypted Domain using FLDA

Recalling the example given in Chapter One of a practical application of a FER system (a digital advert display capturing the facial expression reactions of observers), one of the primary considerations relating to practicality is the privacy of the observers. As such, this section justifies how the classification of facial expressions can be applied to encrypted images. This is achieved using the same principles employed in [39] in which encrypted images of people's faces were recognised by leveraging the homomorphic properties of the Paillier cryptosystem [40]. Using these same principles in addition to a cryptographic protocol for the comparison of two encrypted values, the facial expressions of encrypted images can be classified in such a way that it can be done by a server hosted database without revealing the contents of the image to the server.

The Paillier cryptosystem is an additively homomorphic public-key encryption scheme, where its security is based on the decisional composite residuosity problem [40]. For example, given encryption $[\![a]\!]$ and $[\![b]\!]$, for all operations performed with plaintext or cyphertext, the following corresponding encryption can be obtained where $[\![a + b]\!] = [\![a]\!] \cdot [\![b]\!]$. Similarly, multiplying an encryption $[\![a]\!]$ with a constant c can be calculated as $[\![a \cdot c]\!] = [\![a]\!]^c$ [40].

4.4.1 Projection in the Encrypted Domain

In the case of a training image, a string Exp_i corresponding to the expression class is assigned to the lower dimensional feature vectors $\boldsymbol{y}_i$ for $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M\}$ where the server in order to setup the facial expression database obtains $\boldsymbol{W}_{opt} = [\boldsymbol{w}_1, \boldsymbol{w}_2 \dots \boldsymbol{w}_m]$ using (4.11). As a requirement for Paillier encryption, all the individual elements in $\boldsymbol{W}_{opt}$ and $\boldsymbol{x}_i \; for \; i \in \{1, \dots, M\}$ are denoted by integers. To achieve this, elements in $\boldsymbol{W}_{opt}$ are scaled by a factor SF and subsequently quantized to the nearest integer. Elements of $\boldsymbol{x}_i$ are simply quantized to the nearest integer. The client generates a set of private and public keys, the latter of which is sent to the server.

The encrypted $[\![\boldsymbol{x}]\!]$ is obtained and can be sent to the server when the client encrypts each pixel value of that image $\boldsymbol{x}$ using the earlier generated public key. At this point, the encryption $[\![\boldsymbol{x}]\!]$ is obtained using the client's public key; as such neither the server nor anyone else is able to decrypt the image hence keeping the identity of the subject completely private and confidential from everyone. The server is able to perform linear operations to determine the expression class e.g. operations (4.1) and (4.18) on the encrypted image by leveraging the homomorphic properties of the Paillier cryptosystem described above. The resultant expression class, encrypted by the server using the client's public-key is then sent to the client and is decrypted using their private-key.

Formally, facial expression classification in the encrypted domain requires the evaluation of equations (4.1), which will be the projection of an encrypted test image and (4.18), the Euclidean distance measure in order to match the image with an expression class. For projection of an encrypted test image, equation (4.1) can be rewritten as:

$$\boldsymbol{y}_i = \sum_{j=1}^{n} w_{i,j} \left(\boldsymbol{x}_j - \boldsymbol{\mu} \right) \tag{4.19}$$

where from (4.1), $\boldsymbol{x}$ is re-represented as $\left(\boldsymbol{x}_j - \boldsymbol{\mu} \right)$ being the mean centred image while the projection matrix is correspondingly redefined as: $\boldsymbol{w}_i = [w_{1,i} \; w_{2,i} \dots w_{n,i}]^T$ and the elements of the mean vector are defined as $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_n]^T$. On the side of the server, only the encrypted value of a given test image $[\![\boldsymbol{x}_j]\!]$ is known, as such,

homomorphic properties allow the evaluation of encrypted value of $\boldsymbol{y}_i$ (obtained using clients public key), given by:

$$[\![\boldsymbol{y}_i]\!] = \prod_{j=1}^{n} \left([\![\boldsymbol{x}_j]\!]\,[\![-\boldsymbol{\mu}_j]\!]\right)^{w_{i,j}} \tag{4.20}$$

In (4.18), the n encrypted values of $[\![\boldsymbol{y}]\!]$ that make up the projection of encrypted plain images $[\![\boldsymbol{x}]\!]$, in order to associate a given test sample image with an expression class, we compute the encrypted distances $[\![D_i]\!], i = 1, \dots, \mathrm{M}$ between the elements of the feature vectors of the test sample $\boldsymbol{s}$ and the elements of the feature vectors of the training images $\boldsymbol{t}$. For this, (4.18) can be rewritten as:

$$\begin{aligned} D_i(samp, training) &= \sum_{j=1}^{n} \left(s_j - t_{i,j}\right)^2, i = 1, \dots, \mathrm{M}, \\ &= \sum_{j=1}^{n} t_{i,j}^2 + \sum_{j=1}^{n} \left(-2t_{i,j}\right)s_j + \sum_{j=1}^{n} s_j^2, i = 1, \dots, \mathrm{M}. \end{aligned} \tag{4.21}$$

Noting that (the elements of) test samples and training samples are represented as $\boldsymbol{s}$ and $\boldsymbol{t}$ respectively, the homomorphic properties allow the encrypted distances to be computed as:

$$[\![D_i]\!] = [\![\sum_{j=1}^{n} t_{i,j}^2]\!]\,[\![\sum_{j=1}^{n} (-2t_{i,j})s_j]\!]\,[\![\sum_{j=1}^{n} s_j^2]\!], \qquad i = 1, \dots, \mathrm{M}. \tag{4.22}$$

where the server can obtain $[\![\sum_{j=1}^{n} t_{i,j}^2]\!]$ by encrypting the value $\sum_{j=1}^{n} t_{i,j}^2$ and $[\![\sum_{j=1}^{n} (-2y_{i,j})s_j]\!] = \prod_{j=1}^{n} [\![s_j]\!]^{(-2t_{i,j})}$. The server participates in a two-party computation protocol with the client in order to obtain the value of $[\![\sum_{j=1}^{n} s_j^2]\!]$ as only $[\![s_j]\!]$ is known to the server. During this exchange, the server is also keen to keep the contents of the training database private. As such, the server additively blinds each feature vector component $[\![s_j]\!]$ with a random element $[\![r_j]\!]$, to obtain $[\![\sum_j]\!] = [\![s_j +$

$r_j\rrbracket = \llbracket s_j \rrbracket \llbracket r_j \rrbracket$ which is sent to the client where it is decrypted to calculate $\sum_j^2$ and subsequently $\sum_{j=1}^{m} \sum_j^2$ within the plain domain. Once this is done, the client encrypts $\llbracket \sum_{j=1}^{m} \sum_j^2 \rrbracket$ and sends it to the server who then uses it to deduce $\llbracket \sum_{j=1}^{n} s_j^2 \rrbracket$, given as:

$$\llbracket \sum_{j=1}^{n} s_j^2 \rrbracket = \llbracket \sum_{j=1}^{m} \sum_j^2 \rrbracket \cdot \prod_{j=1}^{m} \left(\llbracket s_j \rrbracket^{-2r_j} \llbracket -r_j^2 \rrbracket \right). \tag{4.23}$$

In doing so, the server has calculated the encrypted distances in (4.22). The next step in associating a test image with an expression class is to identify the image corresponding to the lowest encrypted distance.

4.4.2 Nearest Neighbour Classification in the Encrypted Domain

The objective is to establish the lower of two encrypted l-bit values $\llbracket D_i \rrbracket$ and $\llbracket D_j \rrbracket$. The server calculates $\llbracket z_{i,j} \rrbracket = \llbracket 2^l + D_i - D_j \rrbracket = \llbracket 2^l \rrbracket \llbracket D_i \rrbracket \llbracket D_j \rrbracket^{-1}$, where $z_{i,j}$ is a positive $(l+1)$-bit value. Let the most significant bit of $z_{i,j}$ be represented as $\tilde{z}_{i,j}$, then $\tilde{z}_{i,j} = 0 \Leftrightarrow D_i < D_j$ and $\tilde{z}_{i,j} = 2^{-l} \cdot \left(z_{i,j} - \left(z_{i,j} \bmod 2^l \right) \right)$. Homomorphic properties allow $\tilde{z}_{i,j}$ to be calculated as $\llbracket \tilde{z}_{i,j} \rrbracket = \left(\llbracket z_{i,j} \rrbracket \llbracket z_{i,j} \bmod 2^l \rrbracket \right)^{-2^{-1}}$. The server needs to engage the client to calculate $\llbracket z_{i,j} \bmod 2^l \rrbracket$ as only $\llbracket z_{i,j} \rrbracket$ is known. As previously done, the server generates and applies a random blinding value as $\llbracket z_{i,j} + r \rrbracket = \llbracket z_{i,j} \rrbracket \llbracket r \rrbracket$ which is sent to the client. Once received, the blinded value is decrypted and $z_{i,j} + r \bmod 2^l$ is reduced. The result is encrypted and sent back to the server who retrieves it as:

$$\llbracket z_{i,j} \bmod 2^l \rrbracket = \llbracket z_{i,j} + r \bmod 2^l \rrbracket \llbracket r \bmod 2^l \rrbracket^{-1}. \tag{4.24}$$

Again using a collaborative two-party calculation protocol, the server obtains the encrypted minimum as $\llbracket \tilde{z}_{i,j} \cdot \left(D_i - D_j \right) + D_j \rrbracket$ and the encrypted expression class matching that minimum distance, given by $\llbracket \tilde{z}_{i,j} \cdot \left(Exp_i - Exp_j \right) + Exp_j \rrbracket$. This is then returned to the client who decrypts it to find the expression class of the test image. Further details on FER in the encrypted domain can be found in [41].

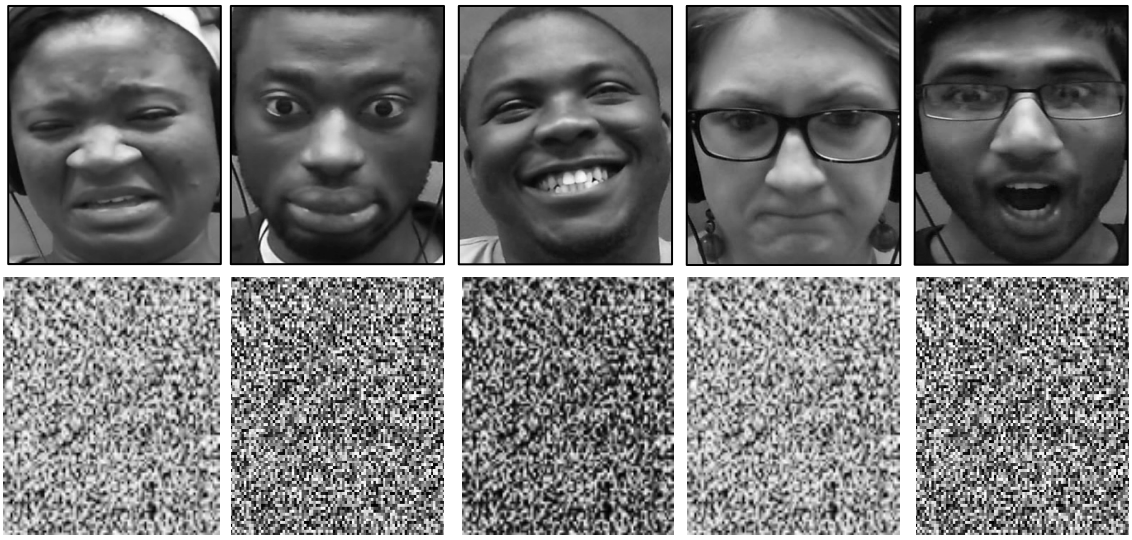


Figure 4–3: Showing sample images from each class of LUSED in the plain domain (top row) and the corresponding image in the encrypted domain (bottom row).

4.5 Simulation Setup

In this section, the simulation setup including notable functions is described in more detail. All the algorithms and simulations using the algorithms detailed in Section 4.3 were run on Matlab R2012b. The hardware was a Viglen desktop computer with the following specification: OS: 64-bit, Windows 7 Enterprise, CPU: Intel(R) Core i5-650, 3.20 GHz, RAM: 4GB. A total of 435 images from 5 classes of the new LUSED (database) were used for the evaluations using each method broken down as follows: DI – 102, FE – 43, HA – 121, SA – 71 and SU – 98. Recalling from Chapter Three that the main database contains 559 images when the 124 Neutral images are added to the above breakdown. Noting that leave-one-out was used for cross validation, it was prudent not to include images from the NE class in order not to trivialize the FER problem and artificially increase recognition rates.

Before simulations, each image was closely cropped manually around the face at the top of the forehead, by the ears and just below the chin. During input, images were converted to greyscale (as depicted earlier in Chapter Three: Figure 3-11) using the function `rgb2gray`, while the `imresize` function was used to change the size of each image to 48×36.

4.5.1 Leave-one-out (LOO)

This is a validation technique that maximizes the number of training images by using the $i-th$ image as the test sample and all the other images as training images for $i = 1, \dots, N$ where N is the total number of images in the dataset.

Considering that when using LOO for PCA, as each image is selected in turn as a test sample, the covariance matrix needs to be re-computed for the remaining training images in that iteration. This represents a significant computational expense, which is mitigated by the way the covariance matrix S_T is computed practically. From equation (4.2), for ease of explanation, assume `A` is a variable in Matlab that contains all the mean centered images $(\boldsymbol{x}_k - \boldsymbol{\mu}), k = 1, \dots, N$. As such, the covariance calculation in Matlab can be written as: `C = A*A'`. However, since the number of training images (N) is usually less than the number of pixels $(n * N)$, the most non-zero eigenvalues that can be found are equal to $(N - 1)$. So we can calculate eigenvalues of `A'*A` (an $N \times N$ matrix) instead of `A*A'` (an $n * N \times n * N$ matrix). It is clear that the dimensions of `A*A'` is much larger that `A'*A`. So the dimensionality will decrease. `L = A'*A` is the *surrogate* of the covariance matrix `C = A*A'`. When the eigen decomposition `[V D] = eig(L);` is computed, the diagonal elements of `D` are the eigenvalues for both `L=A'*A` and `C=A*A'`. All eigenvalues of matrix `L` are sorted and those that are less than a specified threshold are eliminated. So the number of non-zero eigenvectors may be less than $(N - 1)$.

4.5.2 Eigen-vector Selection

The approaches discussed above include the process of first applying PCA and KPCA before FLDA and KFDA respectively. It has been proven in several studies that different eigenvectors may carry information that are relevant to certain facial properties [42, 43], this was also mentioned briefly above. Although the eigen vectors shown in Figure 4-4 can be very significant for recognition, they are often not relevant to the facial expression classification. This is likely because those eigenvectors contain the common features in the face, which exist both within classes and between classes; this causes PCA/FLDA based methods to produce less optimal results. For example, the first eigenface shown below represents the common background. The following few eigenfaces reflect the lighting conditions and possibly gender information. It was

experimentally discovered that discarding the first few eigenvectors produces better results in the case of natural expression recognition evident by better recognition rates [43]. As such, all the evaluation results in this chapter as well as Chapter Five were obtained using eigen vectors 5 to 35. In other words, the 5th largest to the 35th largest eigen values inclusive were used.

Figure 4–4: Five eigenfaces obtained using each of the five largest eigen vectors.

4.6 Evaluation Results

The results of evaluation using the aforementioned methods are contained in this section. *Confusion matrices* were used to present the results (except FLDA in the ED) with the matrix rows representing the actual classes while the columns represent the algorithm-assigned class – as such the diagonal of the matrix represents the correctly recognised images in each class. For each method, the first matrix represents the number of images in their actual classes versus their algorithm-assigned classes. For perspective, a second confusion matrix contains values that represent the percentage of the total number of images in that given class. Subsequently, average recognition accuracy in percentage for the method was calculated by averaging the diagonal of the matrix. More formally, $Ave\ Acc = \frac{1}{C}\sum_{i=1}^{C} x_i$ where C is the number of classes and x_i in this case is the recognition rate of the $i - th$ class.

PCA + NN

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**45**	15	17	15	10	102
FE	5	**22**	5	3	8	43
HA	20	13	**78**	3	7	121
SA	9	16	10	**31**	5	71
SU	12	13	14	7	**52**	98

Table 4–1: Confusion matrix showing recognition results (number of images) on LUSED.

PCA + NN %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**44.12**	14.71	16.67	14.71	9.80
FE	11.63	**51.16**	11.63	6.98	18.60
HA	16.53	10.74	**64.46**	2.48	5.79
SA	12.68	22.54	14.08	**43.66**	7.04
SU	12.24	13.27	14.29	7.14	**53.06**
Average Recognition Accuracy				**51.29%**	

Table 4–2: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

PCA combined with the NN classifier provides fairly consistent recognition rates across the five classes and a reasonable average recognition rate of the LUSED (database). The best performance class-wise being the recognition of the HA and the poorest being the SA classes. The highest single instance of wrongful classification was 22% of the images from SA that were wrongly assigned to FE classes. This is not uncommon as the physical properties of the expressions of sadness and fear are often confused. Consistent with expectations, in the lowest instance of wrongful recognition, only 2% of images belonging to the HA class were wrongly assigned to the SA class.

KPCA + NN

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**46**	15	16	15	10	102
FE	4	**25**	4	3	7	43
HA	20	12	**79**	4	6	121
SA	8	16	10	**32**	5	71
SU	10	13	14	6	**55**	98

Table 4–3: Confusion matrix showing recognition results (number of images) on LUSED.

KPCA + NN %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**45.10**	14.71	15.69	14.71	9.80
FE	9.30	**58.14**	9.30	6.98	16.28
HA	16.53	9.92	**65.29**	3.31	4.96
SA	11.27	22.54	14.08	**45.07**	7.04
SU	10.20	13.27	14.29	6.12	**56.12**
Average Recognition Accuracy				**53.94%**	

Table 4–4: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

KPCA when combined with NN outperformed PCA+NN in every class as well as on average (although only by 2.65%), which is consistent with the theory. The best recognition rates could again be seen in the HA class. The lowest rates are again in the SA class, closely followed by the DI class – both of which recorded only 45% success in the recognition task. The highest wrongful classification can be observed where 22.54% of images belonging to the SA were wrongly assigned to the FE class, notably in the case of the wrongly classified images from the DI class; they were spread evenly across the other classes suggesting the algorithm found it difficult to identify the DI expression and was not simply confused by the images from another specific class.

FLDA + NN

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**67**	8	13	6	8	102
FE	3	**28**	4	2	6	43
HA	12	10	**93**	3	3	121
SA	7	10	7	**43**	4	71
SU	10	11	13	5	**59**	98

Table 4–5: Confusion matrix showing recognition results (number of images) on LUSED.

FLDA + NN %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**65.69**	7.84	12.75	5.88	7.84
FE	6.98	**65.12**	9.30	4.65	13.95
HA	9.92	8.26	**76.86**	2.48	2.48
SA	9.86	14.08	9.86	**60.56**	5.63
SU	10.20	11.22	13.27	5.10	**60.20**
Average Recognition Accuracy				**65.69%**	

Table 4–6: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

FLDA was combined with NN and outperformed both PCA and KPCA (in each class and on average). Notably, significant recognition rates can be observed in the HA class where FLDA outperformed all the algorithms considered in this chapter and thus allowing for a higher average recognition rate. The lowest class recognition accuracy was in the SU class and the largest misclassification was 14% of SA faces that were confused with FE, this is closely followed by 13.95% of images from the FE class that were wrongly recognized as SU faces.

KFDA + NN

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**46**	15	16	15	10	102
FE	4	**25**	4	3	7	43
HA	20	12	**79**	4	6	121
SA	8	16	10	**32**	5	71
SU	10	13	14	6	**55**	98

Table 4–7: Confusion matrix showing recognition results (number of images) on LUSED.

KFDA + NN %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**69.61**	7.84	9.80	7.84	4.90
FE	9.30	**55.81**	9.30	6.98	18.60
HA	14.05	10.74	**67.77**	2.48	4.96
SA	8.45	11.27	8.45	**66.20**	5.63
SU	10.20	11.22	13.27	6.12	**59.18**
Average Recognition Accuracy				**63.71%**	

Table 4–8: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

The combination of KFDA and NN did not perform as well as expected, on average underperforming in comparison to FLDA+NN although with better recognition rates than PCA and KPCA. It is worth considering that perhaps the 'better than expected' performance of FLDA particularly in the HA class rather than under performance of KPCA, could be responsible for the difference in performance. The best and poorest recognition rates can be observed in the DI and FE classes respectively. 18.6% of images in the FE class were wrongly assigned to the SU class.

Summary of Results - LUSED (%)

Method/ Class	PCA+ NN	KPCA+ NN	FLDA+ NN	KFDA+ NN
DI	44.12	45.10	65.69	69.61
FE	51.16	58.14	65.12	55.81
HA	64.46	65.29	76.86	67.77
SA	43.66	45.07	60.56	66.20
SU	53.06	56.12	60.20	59.18
Average	**51.29**	**53.94**	**65.69**	**63.71**

Table 4–9: Showing the recognition accuracy percentage of each class using the corresponding method as well as the overall average recognition accuracy for each method.

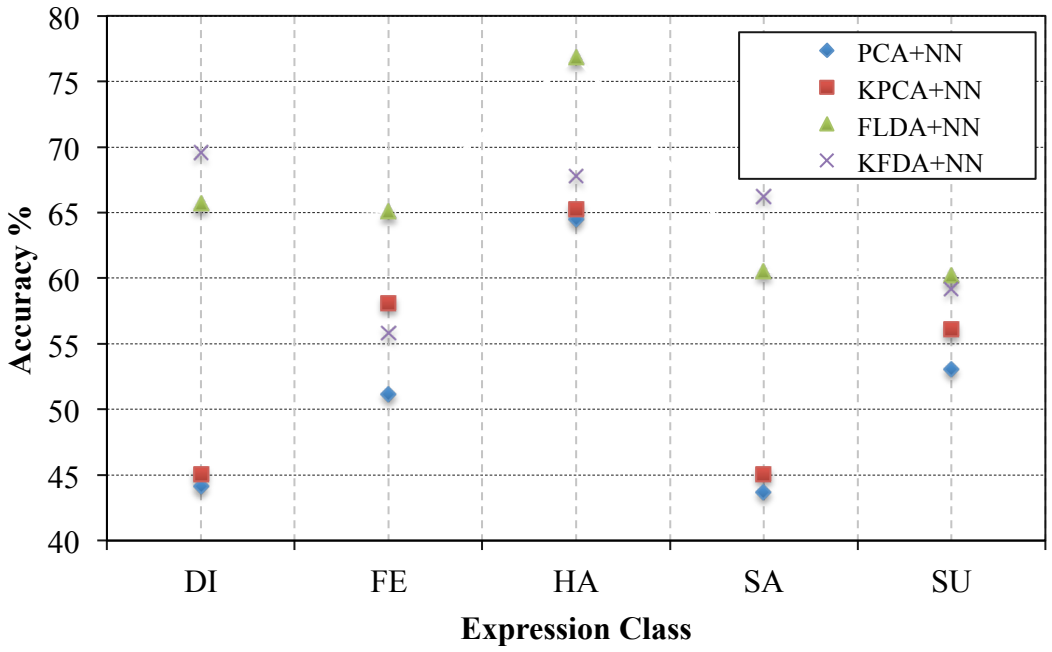


Figure 4–5: Plot of average recognition rates for each class with the use of the different methods to evaluate LUSED.

The HA class had the highest recognition rates consistently across all the algorithm combinations while the SA recorded the lowest recognition rates in two out of the four algorithm combinations covered in this chapter. The wrongly classified facial images represented by the data in the 'off-diagonals' of the resulting confusion matrices above reveal some trends; for example, images from the HA class were mostly confused with images from the DI class – this is likely as a result of the physical characteristics and appearance of both expressions which both feature parted lips, showing teeth and narrowing of the eyes. Interestingly, on the other hand images from the DI class were not consistently wrongly recognized as belonging to the HA class. As expected, images from the HA class were rarely confused with the SA class which can be considered informally as the opposite expression. Also as expected, images from the FE class were often wrongly assigned to the SU class due in part to the physical similarities – raised eyebrows and parted lips. A significant proportion of wrongly classified SU class images were also assigned to the FE class.

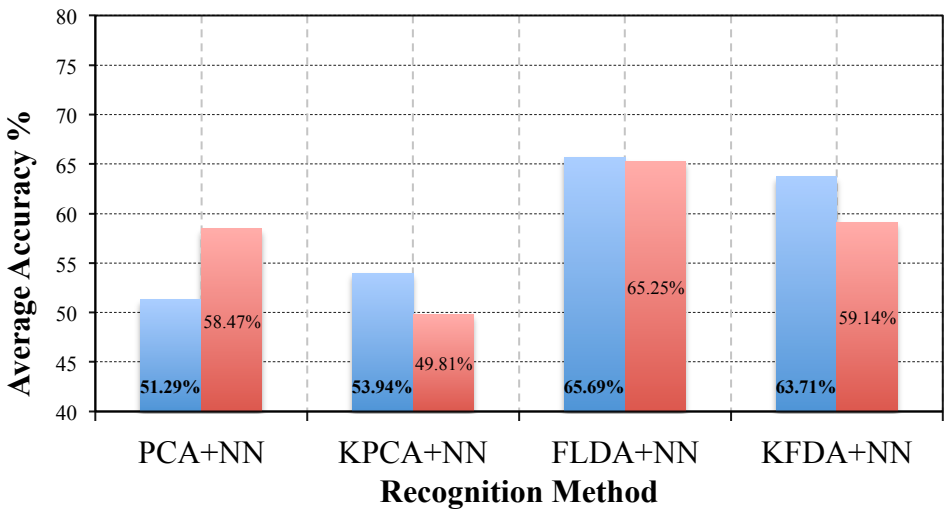


Figure 4–6: Chart of average recognition rates for the corresponding methods used on the evaluation of five classes of LUSED (blue) compared to recognition rates on three classes of the NVIE database (red).

Figure 4-6 summarises the average accuracies for each algorithm on the LUSED alongside results from PCA and FLDA recognition of the NVIE database [44] showing fairly comparable recognition rates between both databases. However for perspective, it is worth pointing out that only three spontaneous classes were classified from the NVIE database while five classes were classified from LUSED making the results from the LUSED evaluation arguably better.

Lastly the classification results (%) obtained in the Encrypted Domain (ED) increase, as the value of the scaling factor SF is increased up to the maximum percentage obtainable in the Plain Domain (PD). The proportional increase in scaling factor values and recognition rates are shown in Table 4-10 below.

ED Scaling Factor	$SF = 1$	$SF = 10^1$	$SF = 10^2$	$SF = 10^3$	$SF = 10^4$
Accuracy %	16.09	61.67	65.69	65.69	65.69
PD Average Recognition Accuracy (FLDA)			**65.69%**		

Table 4–10: Recognition accuracy in % for each of the scaling factor value used in encryption.

4.7 Summary

It is accepted that there are two classical approaches (PCA and FLDA) to finding effective linear transformations in an attempt to solve the image classification problem. It is also common in FER and image processing in general to adopt these methods as a basis for assessing and comparing the performance of image databases and new algorithms to existing literature.

In this chapter, (1) the new natural LUSED (database) presented in Chapter Three was successfully validated and evaluated using two classical approaches as well as kernel variants of both methods recording good recognition rates relative to existing literature. (2) The concept of FER in the encrypted domain was presented and empirically verified addressing a vital privacy concern in the practical application of FER in a 'real-world' scenario. The contribution this evaluation and analysis represents to the community will provide a basis with which future efforts in spontaneous FER can be compared.

While a classic approach basis has been formed in this chapter, the next chapter will apply a more current method to the spontaneous recognition problem using LUSED.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Second Ed. John Wiley & Sons, Inc., 2001.

[2] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 1319, pp. 1299–1319, 1998.

[3] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[4] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," in *Advances in Neural Information Processing Systems 12*, pp. 568–574, 2000.

[5] A. Garg and V. Choudhary, "Facial expression recognition using principal component analysis," *Int. J. Sci. Res. Eng. &Technology*, vol. 1, no. 4, pp. 039–042, 2012.

[6] M. Kaur, R. Vashisht, and N. Neeru, "Recognition of facial expressions with principal component analysis and singular value decomposition," *Int. J. Comput. Appl.*, vol. 9, no. 11, pp. 36–40, 2010.

[7] N. Vretos, N. Nikolaidis, and I. Pitas, "A model-based facial expression recognition algorithm using principal components analysis," *Image Process. (ICIP), 2009 16th IEEE Int. Conf.*, pp. 3301–3304, 2009.

[8] Z. Niu and X. Qiu, "Facial expression recognition based on weighted principal component analysis and support vector machines," *Knowl. Creat. Diffus. Util.*, pp. 174–178, 2010.

[9] A. P. Gosavi and S. R. Khot, "Facial expression recognition using principal component analysis with singular value decomposition," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 158–162, 2013.

[10] Taqdir and J. Kaur, "Facial expression recognition with PCA and LDA," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 6996–6998, 2014.

[11] D. Fadi and F. Davoine, "Facial expression recognition in continuous videos using linear discriminant analysis," in *MVA2005 IAPR Conference on Machine VIsion Applications*, pp. 277–280, 2005.

[12] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1052, 1993.

[13] P. Belhumeur and J. Hespanha, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 19, no. 7, pp. 711–720, 1997.

[14] J. M. Gilbert and W. Yang, "A real-time face recognition system using custom VLSI hardware," *Comput. Archit. Mach. Perception, IEEE Proc.*, no. 617, pp. 58–66, 1993.

[15] I. T. Jolliffe, *Principal component analysis*, Second ed. John Wiley & Sons, Ltd, 2002.

[16] A. M. Burton, V. Bruce, and P. J. B. Hancock, "From pixels to people: a model of familiar face recognition," *Cogn. Sci.*, vol. 23, no. 1, pp. 1–31, 1999.

[17] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions.," *Vision Res.*, vol. 41, no. 9, pp. 1179–208, 2001.

[18] P. J. Hancock, a M. Burton, and V. Bruce, "Face processing: human perception and principal components analysis.," *Mem. Cognit.*, vol. 24, no. 1, pp. 21–40, 1996.

[19] A. J. O'Toole, K. A. Deffenbacher, D. Valentin, and H. Abdi, "Structural aspects of face recognition and the other-race effect.," *Mem. Cognit.*, vol. 22, no. 2, pp. 208–24, 1994.

[20] A. J. O'Toole, H. Abdi, K. Deffenbacher, and D. Valentin, "A perceptual learning theory of the information in faces," *Cogn. Comput. Asp. Face Recognit.*, vol. 8, pp. 159–182, 1995.

[21] H. Abdi, D. Valentin, and B. G. Edelman, "Eigenfeatures as intermediate-level representations: The case for PCA models," *Behav. Brain Sci.*, vol. 21, no. 1, pp. 17 – 18, 1998.

[22] D. Valentin, H. Abdi, B. Edelman, and A. O'Toole, "Principal component and neural network analyses of face images: what can be generalized in gender classification?," *J. Math. Psychol.*, vol. 41, no. 4, pp. 398–413, 1997.

[23] D. K. A. O'Toole A J, Peterson J, "An 'other-race effect' for categorizing faces by sex," *Perception*, vol. 25, no. 6, pp. 669 – 676, 1996.

[24] J. C. O'Toole, A.J., Abdi, H., Deffenbacher, K.A., and Bertlett, "Classifying faces by race and sex using an autoassociative memory trained for recognition," *Proc. Thirteen. Annu. Conf. Cogn. Sci. Soc. Eds. K.J. Hammond D. Gentner, Hillsdale Lawrence Erlbaum.*, 1991.

[25] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 19, no. 7, pp. 696–710, 1997.

[26] M. Turk, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[27] S. Buchala, N. Davey, T. M. Gale, and R. J. Frank, "Principal component analysis of gender, ethnicity, age, and identity of face images," *Proc. IEEE ICMI*, p. 8, 2005.

[28] B. Draper, K. Baek, and M. Bartlett, "Recognizing faces with PCA and ICA," *Comput. Vis. Image Underst.*, vol. 91, no. 1, pp. 115–137, 2003.

[29] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *Neural Networks, IEEE Trans.*, vol. 13, no. 6, pp. 1450–1464, 2002.

[30] M. H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods," *Proc. Fifth IEEE Int'l Conf. Autom. Face Gesture Recognit.*, pp. 215 – 220, 2002.

[31] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–7, 2004.

[32] N. G. Gupta and S. A. Ladhake, "Face detection and facial expression recognition system using 2DPCA," *Int. J. Technol. Res. Eng.*, vol. 2, no. 7, pp. 651–654, 2015.

[33] W. Li, Q. Ruan, and J. Wan, "Two-dimensional uncorrelated linear discriminant analysis for facial expression recognition," in *IEEE 10th International Conference on Signal Processing (ICSP), 2010*, 2002, vol. 10, no. 2, pp. 1362 – 1365.

[34] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, "Spontaneous emotional facial expression detection," *J. Multimed.*, vol. 1, no. 5, pp. 1–8, 2006.

[35] J. Ruiz-del-Solar and P. Navarrete, "Eigenspace-based face recognition: a comparative study of different approaches," *Syst. Man, Cybern. Part C Appl. Rev. IEEE Trans.*, vol. 35, no. 3, pp. 315–325, 2005.

[36] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *Pattern Anal. Mach.*

Intell. IEEE Trans., vol. 19, no. 7, pp. 721–732, 1997.

[37] M. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," *Proc. Int. Conf. Image Process.*, vol. vol. 1, pp. pp. 37–40, 2000.

[38] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–44, 2005.

[39] Z. Erkin, M. Franz, and J. Guajardo, "Privacy-preserving face recognition," *Priv. Enhancing*, pp. 235–253, 2009.

[40] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," *Adv. Cryptology—EUROCRYPT'99*, vol. 1592, 1999.

[41] R. Yogachandran, R. C.-W. Phan, J. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local fisher discriminant analysis," *Affect. Comput. IEEE Trans.*, vol. 4, no. 1, pp. 83–92, 2013.

[42] G. Bebis and S. J. Louis, "Genetic feature subset selection for gender classification: a comparison study," *Proc. Sixth IEEE Work. Appl. Comput. Vision, 2002.*, pp. 165–170.

[43] B. Draper, "Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures," *Empir. Eval. Methods Comput. Vis.*, pp. 1–14, 2002.

[44] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimed.*, vol. 12, no. 7, pp. 682–691, 2010.

Chapter Five

Benchmark Evaluation: Robust Sparsity-based Schemes

5.1 Introduction

In this chapter, the principles of sparse representation theory are explored to solve the Facial Expression Recognition (FER) problem. While Chapter Four was largely based on the use of classical methods to evaluate the Loughborough University Spontaneous Expression Database (LUSED) detailed in Chapter Three, this chapter is motivated by the current relevance of sparse representation-based methods and successes recorded in its application to similar problems such as face recognition [1] and posed FER [2].

"Many computer vision problems can be reformulated as finding a linear representation of an input signal from a dictionary of training samples" [3], as such, the concept of a Sparse Representation Classifier (SRC) is based on the theory of Compressed Sensing (CS). This theory – CS, which is interchangeably referred to as Sparse Coding (SC), has been proposed as a more efficient alternative classification method, "It is based on the premise that a sparse signal can be recovered from a small number of random linear measurements" [2] and is supported by a strong mathematical foundation [4, 5]. With sparse representation, it is possible to represent a signal such as an image accurately and in some cases exactly from a number of samples much smaller than the desired resolution of the image, meaning that a signal can be represented using linear combination of only a few nonzero coefficients taken non-adaptively from

coefficients describing an image. The theory states that an image can be represented by a sparse linear combination of some redundant bases, which constitute an overcomplete dictionary. The advancements in the concept of SRC are fuelled by the discovery that whenever the optimal representation is *sufficiently sparse*, it can be efficiently computed by convex optimization even though the problem can be very difficult in the general case [6].

In addition to the 'facial recognition' background of SRC in image processing along with subsequent applications in posed FER, the fields of pattern recognition and computer vision have also seen applications of SRC to solve problems such as object recognition [7], object tracking [8], image de-noising [9] and super resolution image processing [10].

The process flow of the evaluation process is depicted in Figure 5-1, which is generally representative of the steps required for the use of sparse representation for recognition.

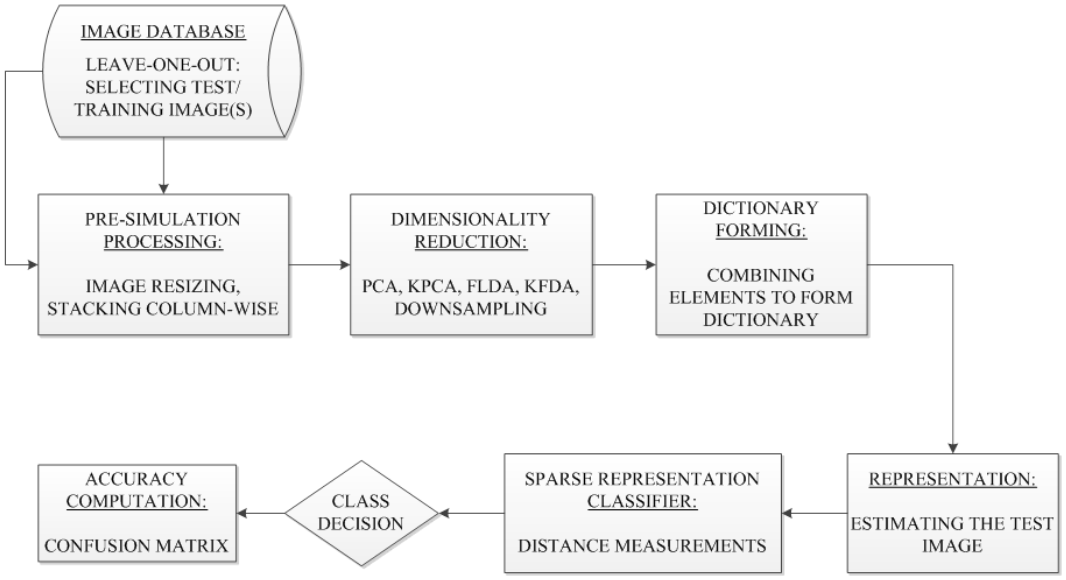


Figure 5–1: Flow Chart/Block Diagram showing the steps in the evaluation of the LUSED using SRC.

5.1.1 Statement of Novelty

Further to the need to validate the LUSED dataset by establishing baseline recognition results using classical methods from the previous chapter is the need to set a benchmark using an up-to-date FER method. In terms of this chapter's contribution, notable is the comprehensive review of research on the use of compressed sensing in solving the FER problem. Further to this is the application of a Sparse Representation-based Classifier (SRC) combined with classic feature extraction methods (PCA, KPCA, FLDA and KFDA) for the recognition of spontaneous facial expressions along with a novel way to expand the dictionary to ensure it is overcomplete. More succinctly, the contributions of this chapter can be summarized as follows:

1. Evaluation of LUSED using SRC combined with classical methods.

2. Expanding the dictionary through the creation of pseudo images from existing images to increase the number of images in the dataset.

3. Comprehensive review of existing research into the use of SRC based methods for FER.

The remainder of this chapter will be organized as follows; Section 5.2 contains background of sparse representation and a review of existing research with its application to FER. The mathematical theory of SRC is contained in Section 5.3 while Section 5.4 contains details of the practical steps taken to achieve an overcomplete dictionary. Section 5.5 provides details of the experiment simulation setup while results are presented in Section 5.6. Finally, a summary of the chapter is offered in Section 5.7.

5.2 Background of Sparse Representation

As mentioned in Chapter Four, the methods employed for FER are largely based on methods already tried and tested in solving other facial image processing problems particularly facial recognition. This is true for both the classical methods and more current signal processing methods. The use of sparse representation for solving the FER problem generally and in the evaluation of LUSED as detailed in this chapter can be

linked back to applications of SRC in face recognition although it is worth mentioning that there are not many applications of SRC for FER.

In the 2000s, researchers began to harness the characteristics of SRC based on the principles of compressed sensing [4], notably, Wright et al. [1] considered the problem of automatically recognizing human faces from frontal views of the face with variances in illumination and facial expression as well as considering the effects of occlusion and disguise. They expressed the recognition problem as one of classifying among multiple linear regression models and argued that a new theory from sparse signal representation obtained from ℓ^1-minimization offered the key to addressing this problem. In so doing, a general classification algorithm for image-based object recognition was established which indicated that in the context of feature extraction, the number of features (whether or not they are sufficiently large) was a more critical consideration than the choice of what features are used as is the case with classical methods. The question of whether or not the features are sufficiently large relates to the dictionary size – having an overcomplete (underdetermined) dictionary (this is explained more explicitly/mathematically in the next section). Wright et al.'s theory that the choice of features was not significant in SRC was confirmed by the similar performances of an overcomplete dictionary when classical approaches e.g. PCA and FLDA were used for feature extraction compared to the performances when unconventional feature extraction approaches were used e.g. down-sampling and projections using random bases.

To this end, schemes varying in complexity have been developed in order to achieve an *ideal* overcomplete dictionary for SRC which can be as simple as selecting a pre-specified transform matrix or to effectively reverse the process by adapting the content of the dictionary. For example, in [11], an algorithm is proposed that generalizes the K-means clustering process for adapting dictionaries in order to achieve sparse signal representation. The authors present an iterative method dubbed K-SVD that alternates between the sparse coding of the examples based on the current dictionary and a process of updating the dictionary elements to better fit the data. The updated dictionary elements are combined with an update of the sparse representations effectively accelerating the convergence of the optimization process on which SRC is based.

In [12], the authors also exploited a sparse representation-based face recognition method for facial expression recognition with 3D face meshes using low-level geometric features. The emphasis was on designing a system that is insensitive to variances in facial expression of the faces being identified. To achieve this, they designed a feature pooling and ranking scheme to collect various types of low level geometric features and ranked them according to their sensitivities to facial expressions.

Converse to facial recognition schemes, FER solutions are best designed to be insensitive to identity and SRC has been applied in many varying ways to FER. The earliest applications of sparse representation to the FER problem were in 2010 [5–7]. Wang et al. [13] compared the performance of SRC on the JAFFE database to other methods such as 2D-PCA and curvelet transform while studying the robustness of SRC to noise. Cotter [15] also used the JAFFE database for FER with a focus on partially corrupted or occluded faces while the authors of [14] also explored the theory of SRC for FER on the CK database. They also showed that the straightforward application of SRC on expressive images posed certain difficulties thereby justifying the use of *difference* images i.e. the images that are derived from the subtraction of the neutral image from the expressive one.

As research progressed, other combinations of SRC for FER were explored for example in [3], where in slight contrast to [14] both accuracy and tolerance to occlusion are increased without the need to perform neutral frame subtraction using the discriminative power of manifold learning combined with the parsimonious power of SRC by utilizing an ℓ^1 reconstruction error and a statistical mixture model. Their experiments were performed on the CK+ and GEMEP-FERA datasets.

Mahoor et al. [16] took a slightly different approach by using SRC for the recognition of facial Action Units (AUs) on the CK database, from which an overcomplete dictionary is formed. The main elements are mean Gabor features of AU combinations and the remaining elements are randomly sampled from a distribution (e.g. Gaussian) that ensures sparse signal recovery.

More recently, both [17] and [2] apply SRC to FER with [17] applying SRC to a multiple Gabor filter and Support Vector Machine (SVM) operation on the JAFFE database while [2] extracted Local Binary Patterns (LBP) and Gabor wavelets prior to

SRC on both the JAFFE and CK datasets and compared the results to other classifiers such as Artificial Neural Network (ANN), K-nearest neighbour (KNN) and SVM.

The recognition rates on the works reviewed above are generally between 70 and 90 percent while those that investigated partially occluded and/or corrupt images recorded results generally around 60 to 90 percent recognition accuracy (dependent on the level of corruption/occlusion). It is however worth pointing out that the majority of these works are on posed (acted) facial expression databases.

5.3 Theory/Mathematics of Sparse Representation

The concept of sparse representation for FER is again consistent with the structure of a basic problem in object recognition, which is to use labelled training samples from k distinct object classes to correctly determine the class to which a new test sample $\boldsymbol{y}$ belongs.

With sparse representation, the test sample $\boldsymbol{y}$ is approximated as a function of the dictionary $\boldsymbol{A}$ and some sparse vector $\boldsymbol{x}$. This is graphically illustrated in Figure 5-2 below.

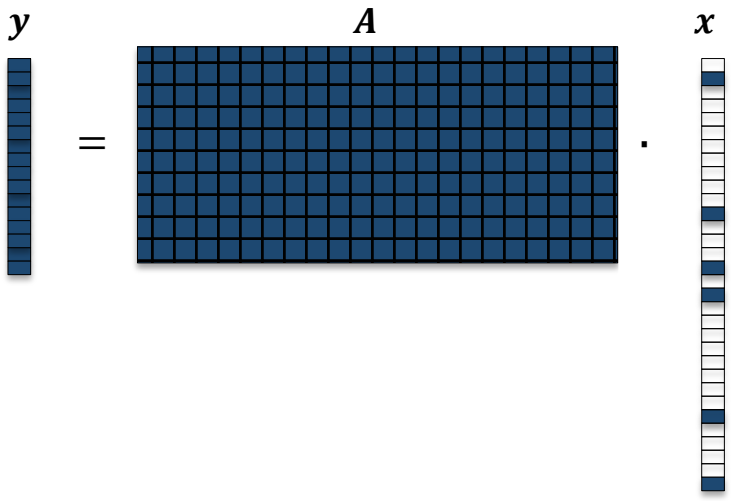


Figure 5–2: Graphical illustration showing the estimation of a test sample.

More formally, given n_i grayscale training images in vector form, each of m-dimension belonging to the i-th class and represented as columns of a matrix:

$$\boldsymbol{A}_i = \left[\boldsymbol{v}_{i,1}, \boldsymbol{v}_{i,2}, \dots, \boldsymbol{v}_{i,n_i} \right] \in \mathbb{R}^{m \times n_i} \tag{5.1}$$

A dictionary comprising of all the training images from all the classes can be composed as:

$$\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_k] \in \mathbb{R}^{m \times n} \tag{5.2}$$

Provided a set of training images, adequately large enough exists for the i-th object class $\boldsymbol{A}_i \in \mathbb{R}^{m \times n_i}$, a new test image $\boldsymbol{y} \in \mathbb{R}^m$ that belongs to the same expression class will approximately lie in the linear span of the training samples corresponding to object i:

$$\boldsymbol{y} = \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \cdots + \alpha_{i,n_i} v_{i,n_i} \tag{5.3}$$

for some scalars, $\alpha_{i,j} \in \mathbb{R},\ j = 1, \dots, n_i$. As such, the linear representation of the test image $\boldsymbol{y}$ can be rewritten in terms of the dictionary as:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 \quad \in \mathbb{R}^m \tag{5.4}$$

where $\boldsymbol{x}_0 = [0, \dots, 0, \alpha_{i,1},\ \alpha_{i,2}, \dots, \alpha_{i,n_i}, 0, \dots, 0]^T \in \mathbb{R}^m$ is a coefficient vector whose elements are zero except for those corresponding to the i-th class.

In the problem of robust FER, usually, $m < n$ making the linear system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ underdetermined (overcomplete), as such, the correct $\boldsymbol{x}_0$ cannot be found as its unique solution. Typically, this difficulty is resolved by selecting the minimum ℓ^2-norm solution:

$$(\ell^2): \quad \hat{\boldsymbol{x}}_2 = \arg\min \|\boldsymbol{x}\|_2 \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{5.5}$$

however, $\hat{\boldsymbol{x}}_2$ is generally *dense*, with a large number of nonzero elements associated to the samples from different classes and thus not particularly useful for recognizing the facial expression test sample $\boldsymbol{y}$. Instead, consider that a test sample $\boldsymbol{y}$ can be adequately

represented with only the training samples from the same class. Such a representation is naturally *sparse*, and bearing in mind that the more sparse the obtained $\boldsymbol{x}_0$, the easier it will be to correctly obtain the class identity of $\boldsymbol{y}$. This prompts the seeking of the sparsest solution to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by solving the optimisation problem:

$$(\ell^0): \quad \hat{\boldsymbol{x}}_0 = \arg\min \|\boldsymbol{x}\|_0 \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{5.6}$$

where the ℓ^0-norm denoted by $\|\cdot\|_0$ counts the number of nonzero elements in $\boldsymbol{x}$. However, obtaining the sparsest solution to an underdetermined problem is NP-hard and difficult to estimate – for the general case, to find the sparsest solution, all subsets for the entries of $\boldsymbol{x}$ need to be exhaustively searched which lacks efficiency.

To mitigate the above, advancements in compressed sensing and sparse representation [6] show that the solution to the ℓ^0-norm problem (5.6) can be closely approximated obtained by solving the following ℓ^1-norm problem:

$$(\ell^1): \quad \hat{\boldsymbol{x}}_1 = \arg\min \|\boldsymbol{x}\|_1 \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{5.7}$$

This is a convex problem and can be solved using standard linear programming methods in polynomial time. The dictionary comprised of all the training samples $\boldsymbol{A}$ can be composed of columns of vectorized raw images or feature vectors obtained from the projection of images using feature extraction techniques such as PCA.

5.3.1 Robustness to Small Dense Noise

It is noteworthy that practically, due to noise that may be present in the data and correlation that may exist between elements from different classes; the solution obtained from (5.7) may not exclusively contain vectors from a single class. This means that in practice, it is unlikely to express the test sample $\boldsymbol{y}$ exactly as a sparse superposition of the dictionary $\boldsymbol{A}$. To accommodate such small dense noise [1], (5.4) can be modified as:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + z \tag{5.8}$$

where $z \in \mathbb{R}^m$ is a noise term. An approximation of the sparse solution can still be obtained by solving the *stable* ℓ^1-minimization:

$$(\ell_s^1): \widehat{\boldsymbol{x}}_1 = \arg\min \|\boldsymbol{x}\|_1 \text{ subject to } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \varepsilon. \tag{5.9}$$

The *stable* ℓ^1-minimization (5.9) is also a convex optimization problem whose solution can be efficiently obtained using second-order cone programming [6].

5.3.2 Robustness to Occlusion/Corruption

Occlusion of part of an image and/or corruption of image pixels poses a significant difficulty to robust facial expression recognition especially in real-world (often spontaneous) scenarios. Furthermore, these potentially randomly sized occlusions may affect any and different parts of an image making such cases unpredictable and increasing the difficulty of the problem. However, these errors can be considered sparse in the standard basis of individual pixels considering that typically such an error represents a fraction of the entire image. As proposed in [1], the basis in which the error is sparse can be considered as a supplementary class of training samples included in the dictionary. The resulting sparse representation of an occluded test image regarding the expanded dictionary (training images plus error basis) naturally separates the component of the test image relating to occlusion from the component relating to the facial expression of the test subject. "In this context, the theory of sparse representation and compressed sensing characterizes when such *source-and-error separation* can take place and therefore how much occlusion the resulting recognition algorithm can tolerate" [1].

Consider a partially occluded test image $\boldsymbol{y}$, in this case, (5.4) above, should be modified as:

$$\boldsymbol{y} = \boldsymbol{y}_0 + \boldsymbol{e}_0 = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}_0, \tag{5.10}$$

where $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a vector of errors whose elements are zero except for a few of them denoted as ρ which model the corrupted pixels of $\boldsymbol{y}$. These corrupted pixels may vary in size and position on a given test image and as such, the scheme for coping with such small dense noise will not mitigate such errors. It is worth noting that redundancy in the

images is key to recognizing corrupted images. As such, the highest possible image resolution is optimal for correctly recognizing occluded or corrupted images conversely if the image is in the feature space, useful information will have been discarded during the feature extraction process. For an occluded image, assume that the affected pixels make up a small portion of the image. Similar to the vector $\boldsymbol{x}_0$, the error vector $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a sparse vector of errors, whose elements are zero except for those representing the corrupt or occluded pixel elements of $\boldsymbol{y}$. Given that $\boldsymbol{y}_0 = \boldsymbol{A}\boldsymbol{x}_0$, eq. (5.10) can be re-written as:

$$\boldsymbol{y} = [\boldsymbol{A}, \boldsymbol{I}] \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{e}_0 \end{bmatrix} = \boldsymbol{B}\boldsymbol{w}_0, \tag{5.11}$$

where $\boldsymbol{B} = [\boldsymbol{A}, \boldsymbol{I}] \in \mathbb{R}^{m \times (n+m)}$, which will always make (5.11) underdetermined meaning again, there is no unique solution for $\boldsymbol{w}_0$. It is also given that $\boldsymbol{w}_0 = [\boldsymbol{x}_0, \boldsymbol{e}_0]$ has at most $n_i + \rho m$ nonzeros. Consider the matrix $\boldsymbol{B}$, provided $\boldsymbol{y} = \boldsymbol{B}\tilde{\boldsymbol{w}}$ for some $\tilde{\boldsymbol{w}}$ with less than $m/2$ non-zeros, $\tilde{\boldsymbol{w}}$ is the unique sparsest solution. It can therefore be inferred that this solution is the sparsest if the test image has corrupt pixels $\boldsymbol{e}$ values that represent approximately less than 50 percent of the image given by: $\frac{m-n_i}{2}$.

As above, the task is to try to obtain the sparsest solution $\boldsymbol{w}_0$ by solving the following *robust* ℓ^1-minimization problem for the robust case;

$$(\ell_r^1): \quad \hat{\boldsymbol{w}}_1 = \arg\min \|\boldsymbol{w}\|_1 \quad \text{subject to } \boldsymbol{B}\boldsymbol{w} = \boldsymbol{y} \tag{5.12}$$

following which class assignment (recognition) can take place.

5.3.3 Recognition Based on a Sparse Vector

To complete the recognition process, we must assign a given test image $\boldsymbol{y}$ to a class, i. In an ideal case, the class membership of $\boldsymbol{y}$ should be obvious from the nonzero elements of the sparse vector $\hat{\boldsymbol{x}}_1$ or $\hat{\boldsymbol{w}}_1$ obtained from (5.9) or (5.12) respectively, but due to noise and modelling errors, a small number of nonzero elements associated with different classes may be present in the sparse vector.

As such, recognition is defined by the closest estimation of $\boldsymbol{y}$ by the sparse vectors corresponding to expression class i. Formally, for each expression class, let $\boldsymbol{\delta}_i$: $\mathbb{R}^n \rightarrow \mathbb{R}^n$ be the function that takes only coefficients corresponding to the i-th class from $\boldsymbol{x} \in \mathbb{R}^n$ to make up a new vector $\boldsymbol{\delta}_i(\boldsymbol{x}) \in \mathbb{R}^n$ whose elements are zero except for the members of $\boldsymbol{x}$ relating to the i-th class. Using the estimations from each class, $\boldsymbol{y}$ is recognised as being a member of the object class which minimises the residual between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}_1$, obtained by:

$$\min_i r_i(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{A}\delta_i(\widehat{\boldsymbol{x}}_1)\|_2 \tag{5.13}$$

As mentioned earlier, an overcomplete dictionary is a prerequisite for the performance of sparse representation and the following section discusses the steps taken to achieve this.

5.4 Practically Achieving an Overcomplete Dictionary

As stated above, the optimal performance of sparse representation is dependent on an overcomplete training dictionary $\boldsymbol{A}$ also described as an underdetermined dictionary. This is achieved when the total number of training images n is greater than the number of elements in each image m, i.e. $m < n$. The three methods employed to ensure this dimensionality constraint is met are detailed below.

5.4.1 Down-sampling

The first and likely the most basic approach to ensuring $m < n$ was to use the `imresize` Matlab function to down sample each image to 18×24, making the value of $m = 432$ (elements in each image) which is just less than the number of images in the LUSED dataset.

5.4.2 Contrived Dictionary

As there are 435 images in the main dataset, and the above down sampling results in 432 elements per image, a simple but novel approach for artificially expanding the dataset is adopted. A new *contrived* image is formed by linearly combining two images from the same class weighted with a random scalar. More

clearly, given two images $\boldsymbol{f}$ and $\boldsymbol{g}$ both $\in \mathbb{R}^m$ and belonging to the same expression class, a new *contrived* image $\boldsymbol{h} \in \mathbb{R}^m$ is formed by the following linear combination:

$$\boldsymbol{h} = \boldsymbol{f}\phi + \boldsymbol{g}\psi \tag{5.14}$$

where ϕ and ψ are both random scalars between 0 and 1. The dataset was expanded by around 33 percent, doubling the number of images in each expression class to form a *contrived* subset with a total of 654 images (breakdown below). Figure 5-3 below shows an example of two images weighted and combined to form a third *contrived* image, further examples of which can be seen Figure 5-4. Although the new image $\boldsymbol{h}$ will be correlated with other members of the dictionary however, experiments were able to show that the identity of a given pseudo image $\boldsymbol{h}$ is sufficiently different from each of the two original images from which it was formed $\boldsymbol{f}$ and $\boldsymbol{g}$ so as not to be recognized as a false positive on the basis of identity but similar enough in facial expression to be correctly recognized as belonging to the same class as the original images.

(a) (b) (c)

Figure 5–3: Example of two subject's images (a) and (b) from the Sad class of LUSED being combined to form a third *pseudo* image (c) also belonging to the Sad class.

(a) (b)

Figure 5–4: Further examples of *pseudo* images from the Disgust (a) and Happy (b) classes.

5.4.3 Feature Extraction

The final method that helped to achieve an overcomplete dictionary was to combine classical feature extraction techniques to sparse representation by inputting into the dictionary the feature representation of a given image which is characteristically lower in dimension than the plain *image space* image. The classical methods applied prior to sparse representation are: PCA, KPCA, FLDA and KFDA details of which can be found in Chapter Four.

5.5 Simulation Setup

In this section, the simulation setup is described in more detail. All the algorithms and simulations using the algorithms detailed in Section 5.3 were run on Matlab R2012b. The hardware was a Viglen desktop computer with the following specification: OS: 64-bit, Windows 7 Enterprise, CPU: Intel(R) Core i5-650, 3.20 GHz, RAM: 4GB. A total of 435 images (each 27×36) from 5 classes of the main LUSED dataset were used for the evaluations using each aforementioned method broken down as follows: DI – 102, FE – 43, HA – 121, SA – 71 and SU – 98. Simulations were also performed on the *down-sampled* subset and also on the 654 images that formed the *contrived* subset of the LUSED dataset detailed in Section 5.4.2 above and broken down as follows:

Expression Class	Original No. of Images	No. of Pseudo Images Added	New Subset Total
Disgust	102	51	153
Fear	43	22	65
Happiness	121	61	182
Sadness	71	36	107
Surprise	98	49	147
Total	**435**	**219**	**654**

Table 5–1: Showing the number of contrived images added to each expression class of the LUSED.

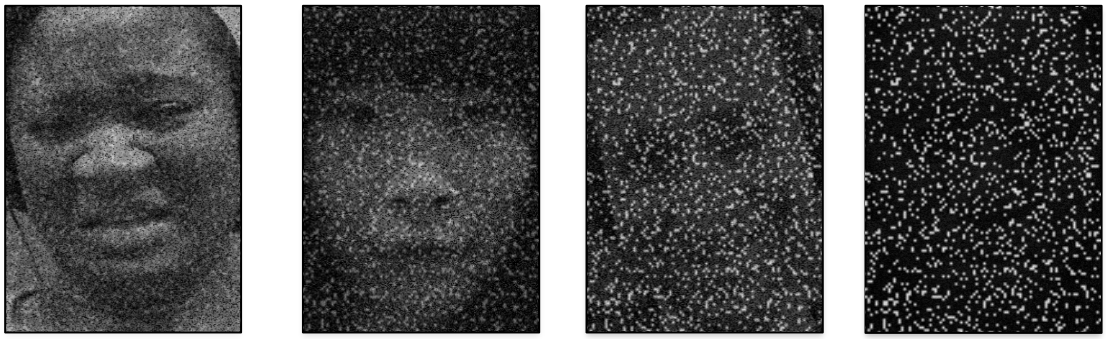
The `linprog` function in Matlab was used for the linear programming optimization required to minimize the ℓ^1-norm while notable functions and simulation setup details relating to the implementation of the classical approaches to FER that were combined with SRC can be found in Chapter Four (Section 4.5). Again, prior to simulations, images were closely cropped around the face and converted to greyscale. Leave-one-out cross validation was also applied again in this chapter's implementation.

In the case of the occluded images, black blocks varying in magnitude up to a maximum of 45 percent of the original image size were placed in three different regions of the face; (1) the *eye* area (arbitrarily on the left or right eye), (2) the *nose* area and (3) the mouth area. In the case of the corrupted images, a percentage of the pixel values are arbitrarily chosen and changed to a similarly distributed (uniform) value between 0 and 255. Examples of images in both categories can be observed in Figures 5-5 and 5-6.

The experiments on corrupt and occluded images were performed using an image size of 36×48 in order to maximize redundancy (as described above in Section 5.3.2) while ensuring the dictionary remained overcomplete. For each of the two cases, duplicate subsets of the main LUSED dataset (one duplicate containing occluded images and the other containing corrupt images) were created in order to allow the algorithm to be trained with *clean* (non-occluded/uncorrupted) images but tested with occluded and corrupt images respectively.

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

Figure 5–5: Examples of images from LUSED showing: (a) image from DI class with block occlusion of the right eye, (b) image from the HA class with block occlusion of the left eye, (c) image from SA class with block occlusion of the nose, and (d) image from SU class with block occlusion of the mouth.



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

Figure 5–6: Examples of corrupted images from the LUSED dataset; (a) test image from DI class with 30% of the pixels corrupt, (b) test image from FE class with 50% of the pixels corrupt, (c) test image from HA class with 70% of the pixels corrupt and (d) test image from SA class with 90% of the pixels corrupt.

5.6 Evaluation Results

This section contains the results of the simulations carried out on the various subsets of the LUSED dataset using the aforementioned methods. The results are presented using *confusion matrices* for both the number of images as well as the percentages. The rows represent the ground truth assigned class and the columns represent the algorithm assigned class, consequently the diagonal of the matrix represents correctly recognised images for each class and the average recognition rate for each simulation is derived from the average of the diagonal.

For ease of presentation, a summary table showing the summary of correctly recognized images rates (both the number of images and the percentage) in each class using the various method combinations is shown for experiments on the main dataset, the contrived subset and the down-sampled subset. Results of the corrupt and occluded images are also presented.

PCA + SRC (No. of Images)

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**73**	8	9	8	4	102
FE	4	**26**	4	3	6	43
HA	13	8	**95**	2	3	121
SA	5	10	7	**46**	3	71
SU	7	9	10	4	**68**	98

Table 5–2: Confusion matrix showing recognition results (number of images) on the main LUSED dataset.

PCA + SRC %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**71.57**	7.84	8.82	7.84	3.92
FE	9.30	**60.47**	9.30	6.98	13.95
HA	10.74	6.61	**78.51**	1.65	2.48
SA	7.04	14.08	9.86	**64.79**	4.23
SU	7.14	9.18	10.20	4.08	**69.39**
Average Recognition Accuracy				**68.94%**	

Table 5–3: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

An average recognition rate of just over 68 percent was recorded when PCA was combined with SRC. The performance from class to class was consistent, with no single class unduly influencing the average. The recognition patterns were also consistent with theory. Expectedly, the HA class had the highest recognition rate (78%) although 10 percent of HA class images we placed in the DI class. The lowest recognition accuracy was the FE class; with the largest proportion of wrongly recognised FE images (13%) were assigned to the SU class while 14 percent of images belonging to the SA class were wrongly assigned to the FE class.

KPCA + SRC (No. of Images)

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**67**	9	11	9	6	102
FE	5	**22**	5	4	7	43
HA	13	9	**91**	2	6	121
SA	5	11	7	**45**	3	71
SU	7	9	10	6	**66**	98

Table 5–4: Confusion matrix showing recognition results (number of images) on the main LUSED dataset.

KPCA + SRC %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**65.69**	8.82	10.78	8.82	5.88
FE	11.63	**51.16**	11.63	9.30	16.28
HA	10.74	7.44	**75.21**	1.65	4.96
SA	7.04	15.49	9.86	**63.38**	4.23
SU	7.14	9.18	10.20	6.12	**67.35**
Average Recognition Accuracy				**64.56%**	

Table 5–5: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

Unexpectedly, the performance of KPCA was slightly lower than PCA when combined with SRC. The class-specific performances were again consistent and evenly distributed except the for the FE class which had the lowest recognition (51%) but was balanced out with above average recognition of the HA class (75%). Another observable trend that is emerging is the high rate of images that belong to the FE class that are wrongly assigned to the SU class.

FLDA + SRC (No. of Images)

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**70**	7	12	6	7	102
FE	4	**23**	6	4	6	43
HA	16	13	**79**	6	7	121
SA	5	9	7	**46**	4	71
SU	9	10	12	5	**62**	98

Table 5–6: Confusion matrix showing recognition results (number of images) on the main LUSED dataset.

FLDA + SRC %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**68.63**	6.86	11.76	5.88	6.86
FE	9.30	**53.49**	13.95	9.30	13.95
HA	13.22	10.74	**65.29**	4.96	5.79
SA	7.04	12.68	9.86	**64.79**	5.63
SU	9.18	10.20	12.24	5.10	**63.27**
Average Recognition Accuracy				**63.09%**	

Table 5–7: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

FLDA+SRC recognized the facial expressions with 63 percent accuracy; in this case DI recorded the best recognition compared to HA in the two previous sets of results. The most wrongly classified images in the DI class were assigned to the HA class – 11 percent while 13 percent of the images in the HA class were wrongly assigned to the DI class (again the largest amount). The FE class again had the lowest recognition accuracy of 53 percent with images wrongly placed in mostly the HA and SU classes (about 14 percent each).

KFDA + SRC (No. of Images)

Class/ Images	DI	FE	HA	SA	SU	Total Images/ Class
DI	**69**	9	10	9	5	102
FE	4	**23**	4	4	8	43
HA	18	13	**79**	4	7	121
SA	5	7	6	**49**	4	71
SU	10	11	13	7	**57**	98

Table 5–8: Confusion matrix showing recognition results (number of images) on the main LUSED dataset.

KFDA + SRC %

Class/ Percentage %	DI	FE	HA	SA	SU
DI	**67.65**	8.82	9.80	8.82	4.90
FE	9.30	**53.49**	9.30	9.30	18.60
HA	14.88	10.74	**65.29**	3.31	5.79
SA	7.04	9.86	8.45	**69.01**	5.63
SU	10.20	11.22	13.27	7.14	**58.16**
Average Recognition Accuracy				**62.72%**	

Table 5–9: Confusion matrix showing recognition results as a percentage of the number of images/class on LUSED.

The combination of KFDA and SRC had the lowest average performance (62%) of the algorithm combinations. In this experiment, the SA class had the best recognition rate (69%) while the FE class recorded the lowest recognition rate (53%) where notably 18 percent of the images were misclassified into the SU class.

Summary of Results - Main LUSED (%)

Method/ Class	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC
DI	71.57	65.69	68.63	67.65
FE	60.47	51.16	53.49	53.49
HA	78.51	75.21	65.29	65.29
SA	64.79	63.38	64.79	69.01
SU	69.39	67.35	63.27	58.16
Average	**68.94**	**64.56**	**63.09**	**62.72**

Table 5–10: Showing the recognition accuracy percentage of each class using the corresponding method as well as the overall average recognition accuracy for each method.

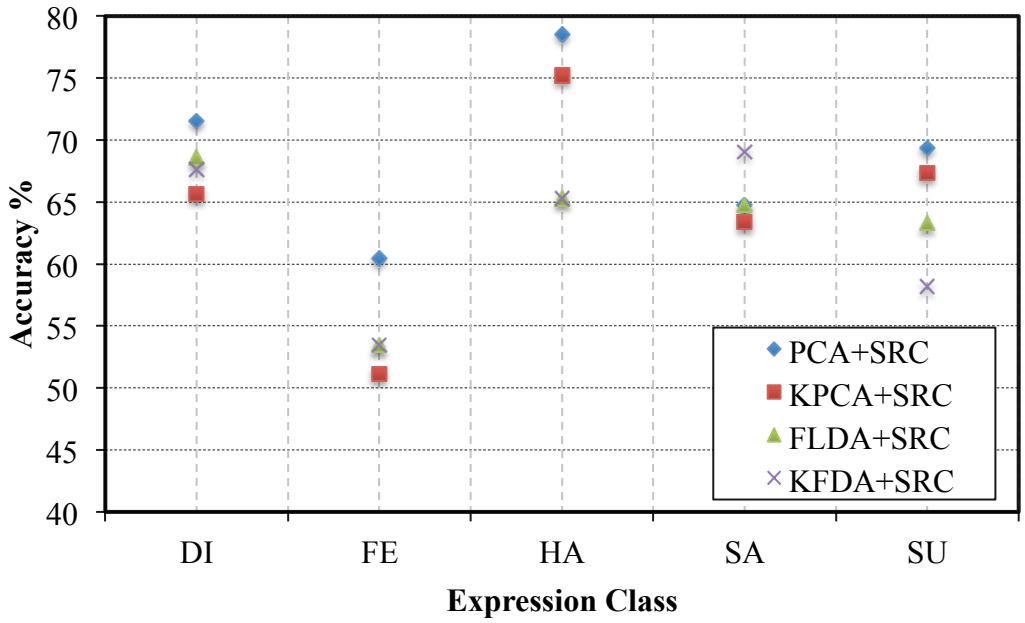


Figure 5–7: Plot of average recognition rates for each class with the use of the different methods to evaluate the **main** LUSED.

The efficacy of the SRC-based method is shown by the above recognition rates, where generally as expected, PCA and KPCA when combined with SRC outperformed the same algorithms when combined with NN by up to 17 percent. Interestingly however, FLDA and KFDA underperformed compared to PCA and KPCA when combined with SRC. This underperformance can occur when the training set is small – sometimes referred to as the Small Sample Size (SSS) problem which is in fact similar to having a dictionary that is not overcomplete.

The kernel variants of PCA and FLDA – KPCA and KFDA should theoretically improve the ability to distinguish between the classes by mapping the data into a (higher dimensional) non-linear space but this has not translated to improved recognition performance of the LUSED dataset which could be due in part to the fewer images in the FE and SA classes. The corresponding lower performances of the kernel variants in those classes are responsible for the lower average recognition rates. KFDA+SRC performed better with the highest recorded average of 70 percent on the recognition of the *contrived* subset (results below), which had a larger number of images.

It can also be hypothesized that the variation in lighting, head position and scale in the LUSED may account in part for the inconsistency in the expected performance order of the algorithms. Other factors such as the choice of features for extraction – PCA and FLDA eigen-vectors were shown not to be important to the performance when sparse representation is employed.

In terms of performance from expression class to class, the performance advantage of PCA+SRC is most clearly observed (Figure 5-7) in the FE class where the difference to the other algorithm combinations is just over five percent. Even though theoretically, FE is a difficult class to recognize but on the LUSED dataset, PCA+SRC is able to better estimate the test image as a function of the dictionary. Consistent in all the algorithms is the fact that the HA and DI classes of LUSED have the best separability from all the other classes while FE and SA were often mixed up and also had the least performance in terms of recognition. This is likely due to respective similarities in the physical manifestation of the HA and DI expressions by the different subjects in LUSED both in terms of appearance and intensity. The similarities of the expressions within each of these classes along with their difference from the expressions in other classes made for better recognition accuracies.

From results across the different algorithm combinations, it can notably be observed that the classes with the lowest recognition accuracies – FE and SA in addition to being theoretical among the more difficult to recognize also have the lowest number of images within the LUSED dataset – 43 and 71 images respectively. Conversely, the classes with best recognition rates from the algorithm combinations – HA and DI are theoretically among the less difficult to recognize but also have the largest number of images in the LUSED – 121 and 102 respectively. The extended contrived is successfully validated by the results below;

Summary of Results - *Contrived* LUSED (No. of Images)

Method/ Class	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC	Total no. of Images
DI	107	101	105	111	**153**
FE	39	33	35	47	**65**
HA	138	137	119	122	**182**
SA	68	71	69	77	**107**
SU	102	99	93	104	**147**

Table 5–11: Showing the number of correctly recognized images in each class using the corresponding method as well as the total number of images in each class.

Summary of Results - *Contrived* LUSED (%)

Method/ Class	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC
DI	69.93	66.01	68.63	72.55
FE	60.00	50.77	53.85	72.31
HA	75.82	75.27	65.38	67.03
SA	63.55	66.36	64.49	71.96
SU	69.39	67.35	63.27	70.75
Average	**67.74**	**65.15**	**63.12**	**70.92**

Table 5–12: Showing the recognition accuracy percentage of each class using the corresponding method as well as the overall average recognition accuracy for each method.

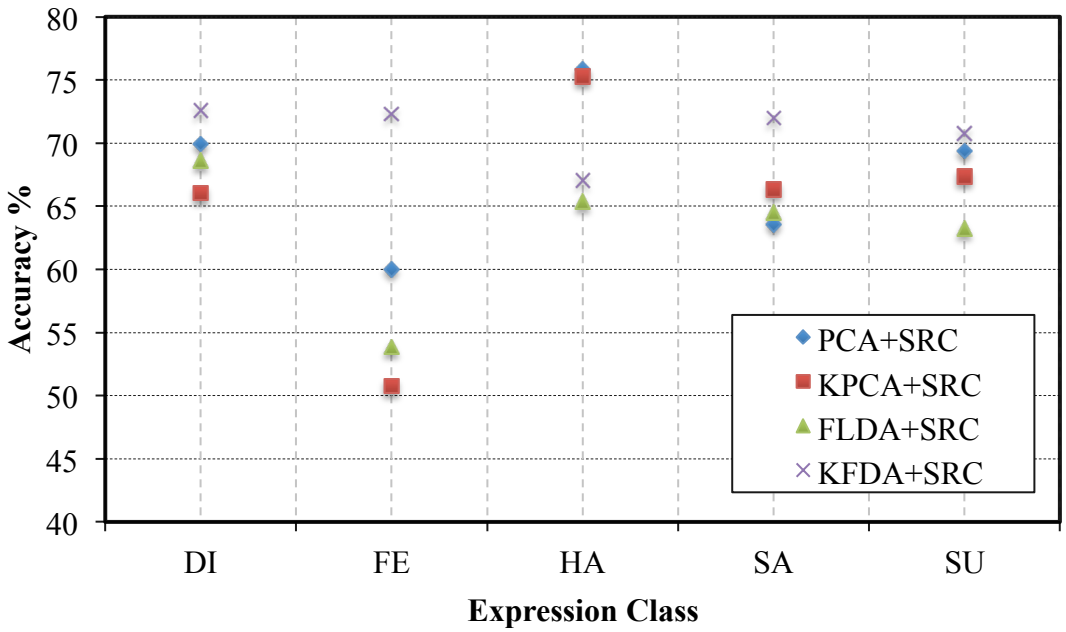


Figure 5–8: Plot of average recognition rates for each class with the use of the different methods to evaluate the **contrived** LUSED subset.

Images from the main subset of LUSED were downsized to 18×24 and the following recognition results were obtained;

Summary of Results – Down-sampled LUSED (No. of Images)

Method/ Class	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC	Total no. of Images
DI	72	66	69	68	**102**
FE	22	20	21	23	**43**
HA	94	93	77	79	**121**
SA	45	45	44	48	**71**
SU	66	65	63	58	**98**

Table 5–13: Showing the number of correctly recognized images in each class using the corresponding method as well as the total number of images in each class.

Summary of Results – Down-sampled LUSED (%)

Method/ Class	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC
DI	70.59	64.71	67.65	66.67
FE	51.16	46.51	48.84	53.49
HA	77.69	76.86	63.64	65.29
SA	63.38	63.38	61.97	67.61
SU	67.35	66.33	64.29	59.18
Average	**66.03**	**63.56**	**61.28**	**62.45**

Table 5–14: Showing the recognition accuracy percentage of each class using the corresponding method as well as the overall average recognition accuracy for each method.

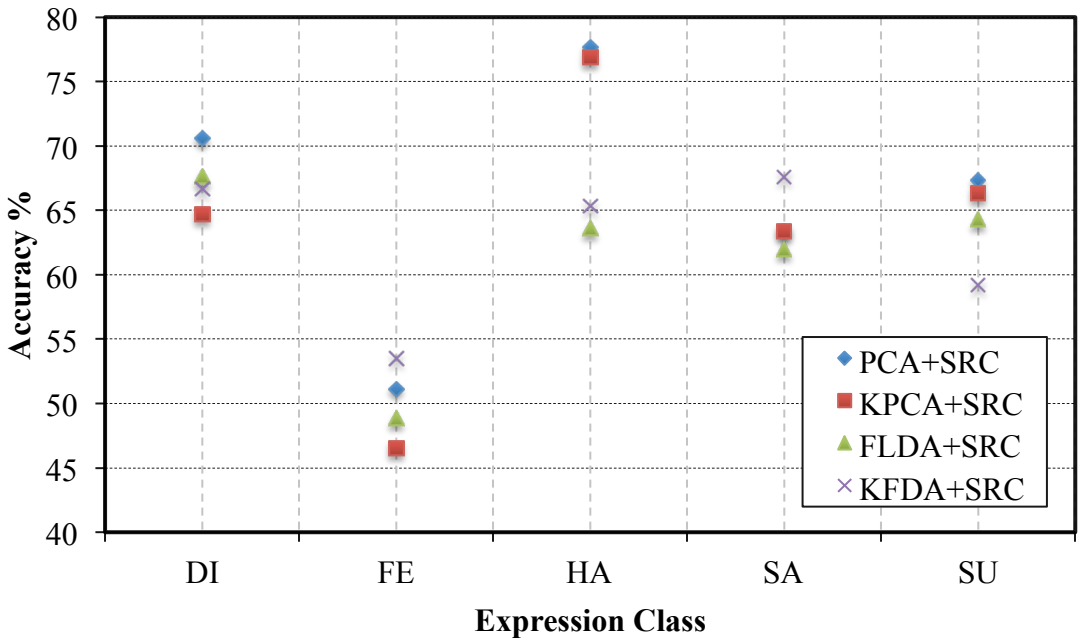


Figure 5–9: Plot of average recognition rates for each class with the use of the different methods to evaluate the **down-sampled** LUSED.

While the literature already asserts that the same recognition rates obtained when using raw (full sized) images can be obtained using images down-sampled (up to a factor of 2) [15], the recognition rates of the down-sampled subset are slightly lower than those of the main dataset when feature extraction is applied prior to SRC. The lower recognition rate is due to the loss of significant data (with dimensionality reduction being performed on down-sampled images) yet interestingly; the dip in performance is not proportionally great.

Summary of LUSED Subsets Results (%)

Method/ Subset	PCA+ SRC	KPCA+ SRC	FLDA+ SRC	KFDA+ SRC
Main	68.94	64.56	63.09	62.72
Contrived	67.74	65.15	63.12	70.92
Down-sampled	66.03	63.56	61.28	62.45

Table 5–15: Showing the recognition accuracy of each class using the corresponding method as well as the overall average recognition accuracy for each method.

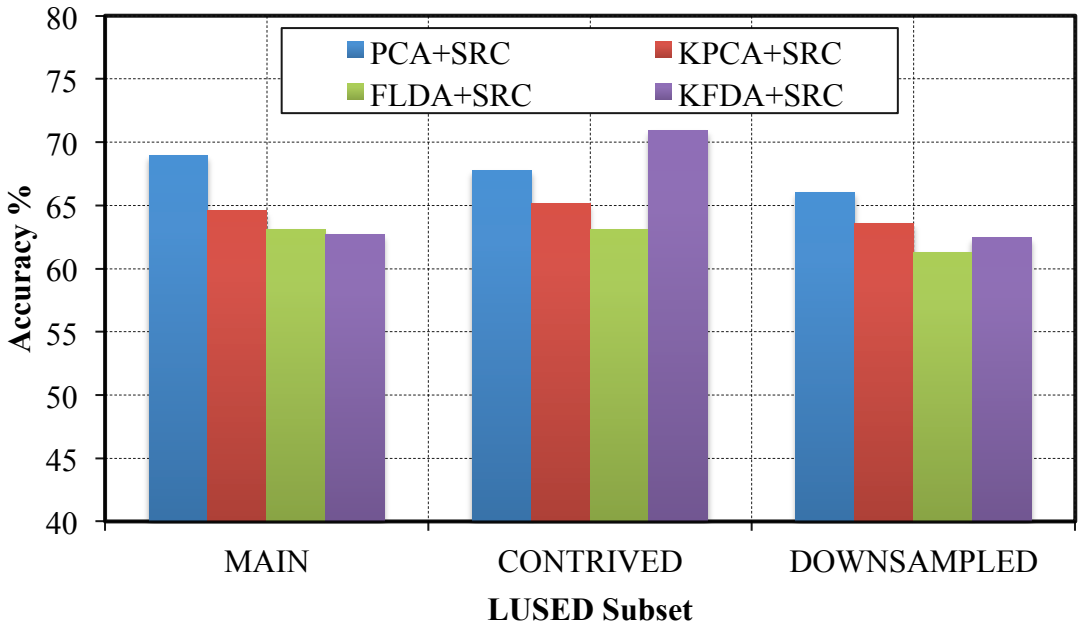


Figure 5–10: Plot of average recognition accuracy (across all classes) using the different method combinations on the three different subsets of LUSED that were

evaluated.

Although the best recorded average results of 68.94% (using PCA+SRC on the main subset) and 70.92% (using KFDA+SRC on the *contrived* subset) are still shy of the recognition rates in the nineties recorded by the papers (on posed databases) reviewed in Section 5.2, they represent the best known recognition rates in the recognition of spontaneous (natural) facial expressions.

The results of the experiments performed on the corrupt and occluded images using PCA+SRC are presented below;

Results – Robustness to Corruption (%)

Corruption Level	0%	30%	50%	70%	90%
Recognition Accuracy	68.94%	68.94%	68.94%	62.53%	4.83%

Table 5–16: Showing the level of corruption and corresponding recognition accuracy.

Results – Robustness to Occlusion (%)

Area of Occlusion	Eye (left)	Eye (right)	Nose	Mouth
Recognition Accuracy	68.94%	68.94%	68.94%	59.77%

Table 5–17: Showing recognition accuracy and corresponding location of up to 30% block occlusion.

For corrupt images, PCA+SRC using equation (5.12) successfully recovered images with up to around 60% of corrupt pixels achieving the same recognition rates (68.94%) as when *clean* images were used for testing. This is despite the fact that images with 50 percent corruption and above would ordinarily not be recognizable with the naked eye as can be observed in Figure 5-6.

Where up to 30% of the image was occluded by a black block in the eye (left/right) and nose region, PCA+SRC was again able to successfully recover images and achieve the same recognition rate as un-occluded images using the same method. However interestingly, images with occlusions in the mouth region of the face were recovered, recognition accuracy reduced by about 9%. This suggests that of the three

facial regions tested with occlusion, the mouth area is the most relevant for facial expression recognition as relates to LUSED.

5.7 Summary

Typically, applications of SRC to the FER problem as reviewed in Section 5.2 have been limited to the use of posed facial expressions for which recognition rates in the high nineties (percent) have been recorded. However the spontaneity inherent in natural facial expression databases play to the strengths of the sparse representation technique particularly in terms of robustness.

In this chapter, (1) a detailed review of existing research on the use of SRC for FER is carried out. (2) The currently relevant SRC-based methods are used in the evaluation of the new dataset (LUSED) which was presented in Chapter Three, additionally robustness to occlusion and corruption were experimentally investigated. (3) A simple but new approach for the expansion of the dictionary was introduced and validated. These contributions represent an up-to-date comparison benchmark to complement the classical approaches employed in Chapter Four.

The following chapter concludes the thesis and offers suggestions for future research direction in the area facial expression recognition.

References

[1] J. Wright and A. Yang, "Robust face recognition via sparse representation," *Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[2] S. Zhang, X. Zhao, and B. Lei, "Facial expression recognition using sparse representation," *Wseas Trans. Syst.*, vol. 11, no. 8, pp. 440–452, 2012.

[3] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold based sparse representation for robust expression recognition without neutral subtraction," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2136–2143, 2011.

[4] D. L. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[6] D. L. Donoho, "For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution," *Comm. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, 2006.

[7] J. Mairal, F. R. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Learning discriminative dictionaries for local image analysis," in *Computer Vision and Pattern Recognition, 2008. CVPR IEEE Conference on.*, 2008.

[8] X. Mei and H. Ling, "Robust visual tracking using l1 minimization," *Comput. Vision, 2009 IEEE 12th Int. Conf. on. IEEE*, no. ICCV, pp. 1436–1443, 2009.

[9] H. Li and F. Liu, "Image denoising via sparse and redundant representations over learned dictionaries in wavelet domain," *Proc. 5th Int. Conf. Image Graph. ICIG 2009*, vol. 15, no. 12, pp. 754–758, 2010.

[10] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[11] M. Aharon, M. Elad, and A. Bruckstein, "K - SVD : An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[12] X. L. X. Li, T. J. T. Jia, H. Z. H. Zhang, and V. Tech, "Expression-insensitive 3D face recognition using sparse representation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*,

pp. 2575–2582, 2009.

[13] Z.-W. Wang, M.-W. Huang, and Z.-L. Ying, "The performance study of facial expression recognition via sparse representation," *Int. Conf. Mach. Learn. Cybern. (ICMLC 2010)*, vol. 2, pp. 824–827, 2010.

[14] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization," *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work.*, pp. 32–39, 2010.

[15] S. Cotter, "Sparse representation for accurate classification of corrupted and occluded facial expressions," *Int. Conf. Acoust. Speech Signal Process.*, pp. 838–841, 2010.

[16] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 336–342, 2011.

[17] R. El-Sayed, a El Kholy, and M. El-Nahas, "Robust facial expression recognition via sparse representation and multiple gabor filters," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 4, no. 3, pp. 82–87, 2013.

Chapter Six

6.1 Conclusions

Although most humans are able to recognize facial expressions with little or no conscious effort; reliable, accurate and robust recognition of facial expressions by machines continues to be a challenge. This thesis presented work that was aimed at supporting research in automatic facial expression through the development and evaluation of a new *culturally diverse* and balanced representative training/testing dataset that is now freely available to the research community and can be accessed at [http://www.lboro.ac.uk/departments/eese/research/communications/lused.html].

This chapter will summarize the thesis chapter-by-chapter and highlight the notable contributions and conclusions drawn therein particularly with respect to meeting the aims and objectives. In order to concisely articulate these conclusions, the corresponding steps taken to meet each objective follow the objectives initially set out in Chapter One;

Chapter Two

Chapter Two laid the foundation for the schemes applied in the main contribution Chapters (Three – Five). Noteworthy aspects of pattern recognition systems were introduced as well as some of the possible challenges. Applications of FER systems in different sectors were also discussed.

Chapter Three

- "*Objective One*: To design an experiment that will facilitate the spontaneous display of natural facial expressions that can be captured discretely. The goal is to do this in a semi-controlled environment such that the variability from subject to subject can be reduced without compromising the spontaneity of the expressions being elicited."

For this objective, an experimental set-up was designed for the collection of LUSED where participants watched stimuli video under the guise of participating in a pseudo experiment in order for their facial affective reactions to the video to be discretely captured. The stimuli video was selected and arranged to maximise the potential for eliciting a response. The experiments took place in a purposefully set-up multimedia lab to ensure the uniformity of variables such as lighting. A pilot experiment was used to measure the effectiveness of the stimuli video as well as the experimental process from start to finish, including the potential of participants to discern the true purpose of the experiments.

- "*Objective Two*: To recruit a diverse and balanced set of participants without incentives."

Participants of the experiment were primarily staff and students of Loughborough University and were recruited through mailing list requests, advertisements for participants on digital billboards around campus and word of mouth. None of the subjects were given any incentives or compensation for participation to ensure there was no artificial inducement to take part.

- "*Objective Three*: To establish the ground truth labels for the elicited facial expressions."

Following processing, each image was labelled using information from the debrief notes combined with timing information and the use of an independent judging panel to ensure the accuracy of the labels.

From Chapter Three, an interesting observation relating to the ethnicity of the participants leads to the submission that the intensity and animation of expressions in Asian and Black subjects was greater than in the case of Caucasian and Mixed subjects. There has been research on the universal nature of facial expressions (discussed in Chapters One and Three) and the findings in this thesis are consistent with literature in that ethnic origin has little or no effect on the types of expressions displayed however there is no known literature that specifically looks into the intensity and animation of facial expressions as it relates to the ethnicity of the subject.

Chapter Four

- "*Objective Four*: To review and analyse historic methods for facial expression recognition."

To meet this objective, existing literature on the use of baseline algorithms; particularly PCA and FLDA combined with a NN-based classifier for FER was reviewed.

- "*Objective Five*: To establish baseline results for the recognition of expressions from the newly created database using historic methods."

The new dataset (LUSED) was validated by the use of PCA, KPCA, FLDA and KFDA for feature extraction each combined with a NN classifier. The highest recognition rate of **65%** was recorded using the combination of FLDA and NN. From experimental verification, it can also be asserted that the best eigen-vectors for spontaneous FER are those that correspond to the eigen-values: 5 – 35 inclusive when sorted from largest to least. Furthermore in this chapter, a model for FER in the encrypted domain is proposed to mitigate privacy concerns. This model is based on the FLDA and NN combination and results obtained show that recognition can occur in the encrypted domain with the same level of accuracy obtainable in the plain domain (65%).

- "*Objective Six*: Comparative analysis of the baseline results obtained from the new database and other existing databases."

The performance of the evaluation on LUSED dataset using PCA, KPCA, FLDA and KFDA combined with NN was measured against the performance of another spontaneous dataset – NVIE database. Using the same algorithm combinations, the LUSED dataset out performs the NVIE dataset in three of the four combinations.

Chapter Five

- "*Objective Seven*: To review and analyse up-to-date techniques for facial expression recognition."

A review of literature was performed with a view to identifying more recent signal processing techniques that have been applied to the FER problem.

- "*Objective Eight*: To establish benchmark results for the robust recognition of expressions from the newly created database using up-to-date techniques."

Historic pattern recognition schemes were combined with the more recent signal processing method based on sparse representation. Schemes to ensure an overcomplete dictionary were applied including a simple but novel approach to form new pseudo images. PCA, KPCA, FLDA and KFDA combined with SRC and the PCA+SRC produced the best recognition rate of **68%** from the main database. A recognition rate of **70%** was also recorded on the subset of pseudo images. Additional simulations were carried out to test for robustness to corruption and occlusion; the system could recognize images with up to *50% corruption* while images with up to *30% block occlusions* could be recognized.

6.2 Future Work

As alluded to in previous chapters, there is still a need for continued research in the area of FER. In addition to continued research on new methods and algorithms, the following are areas relating to the database that can be further investigated.

Experiments that will allow the creation of a new dataset by capturing spontaneous facial expressions *in the wild* will be beneficial to FER research. This

refers to capturing data in an uncontrolled setting such as via webcams or from cameras placed in public places. If this is to be done without the knowledge of the subjects, the immediately apparent challenges with this are three fold; (1) the ethical privacy issues related to the collection of images without permission; (2) ensuring a low level of variance from environmental conditions such as lighting and head orientation; (3) finally, once the images are collected, establishing the ground truth labels for each image may be more challenging without the benefit of feedback from the subjects. Notwithstanding, if these challenges can be effectively mitigated, it will be of added value to FER research, as the images will form an even better representation of real-life conditions in terms of training the algorithms.

The *intensity and level of animation* with which different ethnicities display facial expressions is another area where more research can be undertaken. An understanding of the effects of ethnicity or lack thereof can be particularly useful when working with continuous images (videos) to better interpret facial expressions.

Another variable that will be interesting to study is how a person's *mood or life experiences* may affect affective displays. One of the less emotive participants in the LUSED experiments reported that they had seen many of the stimuli material previously and as such were bored. A deeper understanding of subjects' emotive predisposition will be useful in the design of experiments to build more representative datasets.

Another emerging area for future work in FER is the study of *micro expressions*. While basic facial expressions are more obvious in appearance, subtle facial reactions sometimes displayed involuntarily also hold valuable information about affective states. For example, if someone was trying to suppress their facial reaction to a stimuli event, they may give away a micro expression. The ability for machines to recognize these micro expressions will in a sense surpass the capability of most humans.

Algorithmically, an emerging research area worthy of further investigation for naturalistic facial expression recognition is the use of *deep learning* for classification which has already shown good potential in the recognition of posed facial expressions.

Appendix

A. LUSED Experimental Procedure

The following is the step-by-step procedure for volunteer participants of the LUSED experiments.

Step	Activity	Time (Approx.)
Step One	Participant arrives at multimedia lab and is seated on a lazy chair given 'experiment information sheet' containing information about pseudo experiment – memory test and asked to sign a form consenting to participation.	6 minutes
Step Two	Participant is invited to seat in front of computer screen and watch stimuli video while being recorded and observed in real time by moderator.	19 minutes
Step Three	Participant is told true purpose of the experiment and given further details of the real (while still being recorded).	5 minutes
Step Four	Participant is debriefed to get feedback about the emotions experienced in the different sectors of the stimuli video.	10 minutes
Step Five	Permission of participant is sought to be included in the database and a 'release form' is signed.	2 minutes

Table A–1: Showing activity procedure for LUSED experiment participants.

B. LUSED Experimental Setup

Figure A–1: Image showing frontal view of experiment area within the multimedia lab (directly ahead is the participants position and on the left is the moderators position).

Figure A–2: Side view of experiment area showing participants position.

Figure A–3: View from experiment moderator's position showing mock participant.

Figure A–4: Side view showing mock participant seated with arms on the armrest wearing blood pressure monitor. Participant is watching the stimuli video and displaying affective reactions despite knowing the purpose of the video.

Figure A–5: Blood pressure monitor worn by participants to discourage arm movement and obstruction of the face – participants were told the monitor reads their pulse as part of the observations of the pseudo experiment.

Figure A–6: Close up of mock participant's arm position on the armrest.