

https://dspace.org/

# An independent evaluation of subspace facial recognition algorithms

**Surajpal, Dhiresh Ramchander**

**2008-12-23T07:42:07Z**
http://hdl.handle.net/10539/5906

Downloaded from a DSpace repository. https://dspace.org/

# AN INDEPENDENT EVALUATION OF SUBSPACE FACIAL RECOGNITION ALGORITHMS

A RESEARCH REPORT SUBMITTED TO THE SCHOOL OF ELECTRICAL AND INFORMATION ENGINEERING
FACULTY OF ENGINEERING AND BUILT ENVIRONMENT, UNIVERSITY OF WITWATERSRAND

IN PARTIAL FULFILLMENT OF THE
DEGREE IN MASTER OF SCIENCE IN ENGINEERING

August 2006

By

**Dhiresh Ramchander Surajpal**

## *DECLARATION*

I declare that this is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering at the University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination at any other university.

Signed on this _____ day of _____ 2006.

_____

Dhiresh R. Surajpal

## *EXECUTIVE SUMMARY*

In traversing the diverse field of biometric security and face recognition techniques, this investigation explores a rather rare comparative study of three of the most popular Appearance-based Face Recognition projection classes, these being the methodologies of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA). Both the linear and kernel alternatives are investigated along with the four most widely accepted similarity measures of City Block (L1), Euclidean (L2), Cosine and the Mahalanobis metrics. Although comparisons between these classes can become fairly complex given the different task natures, the algorithm architectures and the distance metrics that must be taken into account, an important aspect of this study is the completely equal working conditions that are provided in order to facilitate fair and proper comparative levels of evaluation. In doing so, one is able to realise an independent study that significantly contributes to prior literary findings, either by verifying previous results, offering further insight into why certain conclusions were made or by providing a better understanding as to why certain claims should be disputed and under which conditions they may hold true. The experimental procedure examines ten algorithms in the categories of expression, illumination, occlusion and temporal delay; the results are then evaluated based on a sequential combination of assessment tools that facilitate both intuitive and statistical decisiveness among the intra and inter-class comparisons. In a bid to boost the overall efficiency and accuracy levels of the identification system, the 'best' categorical algorithms are then incorporated into a hybrid methodology, where the advantageous effects of fusion strategies are considered. This investigation explores the weighted-sum approach, which by fusion at a matching score level, effectively harnesses the complimentary strengths of the component algorithms and in doing so highlights the improved performance levels that can be provided by hybrid implementations. In the process, by firstly exploring previous literature with respect to each other and secondly by relating the important findings of this paper to previous works one is also able to meet the primary objective in providing an amateur with a very insightful understanding of publicly available subspace techniques and their comparable application status within the environment of face recognition.

## *ACKNOWLEDGEMENTS*

In helping me achieve one of the most important milestones in my life, I firstly would like to thank Professor Tshilidzi Marwala of the School of Electrical and Information Engineering, University of the Witwatersrand, for being such a superb and inspirational supervisor and for always encouraging me to deliver my best in all the work I produced.

To the National Research Foundation (NRF) and Denel Aerospace Systems (Kentron), their financial support has had a tremendous impact on fulfilling my Post-graduate education and I would like to extend my heart-felt thanks for their generosity.

And lastly and most importantly, to my parents, for everything they have done for me throughout the years. Their unwavering devotion, sacrifice, encouragement, guidance and inspiration are the reasons for all I have achieved and all I have become. Mum and Dad thank you for the world you've given me.

*"Nothing's impossible……Your heart's passion, your mind's determination and a little hard work is all you need in realising your dreams!"*

# *TABLE OF CONTENTS*


LIST OF FIGURES ..... VII
LIST OF TABLES ..... VIII
LIST OF ABBREVIATIONS ..... IX
LIST OF SYMBOLS ..... XI

**CHAPTER 1 – INTRODUCTION ..... 1**

1.1 INTRODUCTION ..... 1

**CHAPTER 2 – BACKGROUND ..... 8**

2.1 INTRODUCTION ..... 8
2.2 PRINCIPAL COMPONENT ANALYSIS (PCA) ..... 9
2.3 LINEAR DISCRIMINANT ANALYSIS (LDA) ..... 12
2.4 INDEPENDENT COMPONENT ANALYSIS (ICA) ..... 15
*2.4.1 Architecture I: Statistically Independent Basis Images ..... 18*
*2.4.2 Architecture II: Statistically Independent Coefficients (Factorial Face Code) ..... 19*
2.5 KERNEL METHODS ..... 20
*2.5.1 Kernel Principal Component Analysis (KPCA) ..... 22*
*2.5.2 Kernel Discriminant Analysis (KDA) ..... 23*
2.6 LITERARY PERFORMANCE REVIEW ..... 25

**CHAPTER 3 – IMPLEMENTATION ..... 27**

3.1 INTRODUCTION ..... 27
3.2 DATABASE SELECTION ..... 28
3.3 SOFTWARE TOOLS ..... 30
3.4 IMAGE PRE-PROCESSING ..... 31
*3.4.1 Traditional Procedure ..... 31*
*3.4.2 Alternative Techniques ..... 32*
3.5 DIVISION OF GALLERY AND PROBE SETS ..... 33
3.6 SIMULATION METHODOLOGY ..... 35
*3.6.1 Training ..... 35*
*3.6.2 Testing ..... 38*
*3.6.3 Similarity Measures ..... 41*
*3.6.4 Evaluation Tools ..... 42*
*3.6.5 Selection of Subspace Algorithms ..... 47*

**CHAPTER 4 – RESULTS AND ANALYSIS ..... 49**

4.1 INTRODUCTION ..... 49
4.2 EXPRESSION ..... 50
4.3 ILLUMINATION ..... 57
4.4 LOWER OCCLUSION ..... 63
4.5 UPPER OCCLUSION ..... 70
4.6 TEMPORAL DELAY ..... 76
4.7 SUMMARY OF RESULTS ..... 81
4.8 COMPARISON TO PREVIOUS LITERARY WORKS ..... 88

**CHAPTER 5 – HYBRID FORMULATION ..... 91**

5.1 INTRODUCTION ..... 91
5.2 FUSION STRATEGY ..... 94
5.3 HYBRID RESULTS AND COMPARISONS ..... 96

**CHAPTER 6 – CONCLUSION ..... 100**

6.1 CONCLUSION ..... 100

**REFERENCES ..... 103**

***APPENDIX A* – SUPPORT MATERIAL AND GRAPHS ..... 111**

***APPENDIX B* – MATLAB MODULES ..... 114**

***APPENDIX C* – CONFERENCE PAPER ..... 117**

# *LIST OF FIGURES*


FIGURE 1 BLIND SOURCE SEPARATION MODEL ..... 16
FIGURE 2 TWO ARCHITECTURES FOR PERFORMING ICA ON FACES [20] ..... 18
FIGURE 3 ILLUSTRATIVE EXAMPLE OF SUBSPACE BASIS VECTORS [30] ..... 19
FIGURE 4 EXAMPLE OF AN INDIVIDUAL SET FROM THE AR DATABASE ..... 30
FIGURE 5 DEPICTION OF PRE-PROCESSING PROCEDURE ..... 31
FIGURE 6 POSE VARIATION AFTER ALIGNMENT ..... 32
FIGURE 7 AN EXAMPLE OF MASKING ..... 33
FIGURE 8 INDIVIDUAL SAMPLE OF TRAINING AND PROBE IMAGES ..... 34
FIGURE 9 GENERIC TRAINING MODEL ..... 36
FIGURE 10 ENERGY SPECTRUM FOR CLASSES OF PCA, ICA AND LDA ..... 37
FIGURE 11 RANK 1 ACCURACY VS. DIMENSIONALITY ..... 38
FIGURE 12 GENERIC TESTING MODEL ..... 39
FIGURE 13 EIGENFACES ..... 39
FIGURE 14 FISHERFACES ..... 40
FIGURE 15 ICA FACES ..... 40
FIGURE 16 RANK SCORE EXAMPLE ..... 43
FIGURE 17 EXAMPLE OF CMS CURVES WITH 95% CI OVER TEN TESTS ..... 44
FIGURE 18 CMS FOR TWO EXAMPLE ALGORITHMS ..... 45
FIGURE 19 AN EXAMPLE PLOT OF *P* VALUES FOR MCNEMAR'S TEST ..... 46
FIGURE 20 DECISION GRAPH FOR MCNEMAR'S TEST ..... 47
FIGURE 21 CMS CURVES FOR ALL SIMILARITY MEASURES – EXPRESSION: PCA ..... 52
FIGURE 22 CMS CURVES FOR ALL SIMILARITY MEASURES – EXPRESSION: ICA ..... 52
FIGURE 23 CMS CURVES FOR ALL SIMILARITY MEASURES – EXPRESSION: LDA ..... 53
FIGURE 24 MCNEMAR'S TEST FOR THE TOP 3 INTER-CLASS ALGORITHMS - EXPRESSION ..... 56
FIGURE 25 CMS CURVES FOR ALL SIMILARITY MEASURES – ILLUMINATION: PCA ..... 58
FIGURE 26 CMS CURVES FOR ALL SIMILARITY MEASURES – ILLUMINATION: LDA ..... 58
FIGURE 27 CMS CURVES FOR ALL SIMILARITY MEASURES – ILLUMINATION: ICA ..... 59
FIGURE 28 MCNEMAR'S TEST FOR THE TOP 3 INTER-CLASS ALGORITHMS - ILLUMINATION ..... 62
FIGURE 29 CMS CURVES FOR ALL SIMILARITY MEASURES – LOWER OCCLUSION: PCA ..... 65
FIGURE 30 CMS CURVES FOR ALL SIMILARITY MEASURES – LOWER OCCLUSION: ICA ..... 65
FIGURE 31 CMS CURVES FOR ALL SIMILARITY MEASURES – LOWER OCCLUSION: LDA ..... 66
FIGURE 32 MCNEMAR'S TEST FOR THE TOP 3 INTER-CLASS ALGORITHMS – LOWER OCCLUSION ..... 69
FIGURE 33 CMS CURVES FOR ALL SIMILARITY MEASURES – UPPER OCCLUSION: PCA ..... 71
FIGURE 34 CMS CURVES FOR ALL SIMILARITY MEASURES – UPPER OCCLUSION: LDA ..... 72
FIGURE 35 CMS CURVES FOR ALL SIMILARITY MEASURES – UPPER OCCLUSION: ICA ..... 72
FIGURE 36 MCNEMAR'S TEST FOR THE TOP 3 INTER-CLASS ALGORITHMS – UPPER OCCLUSION ..... 75
FIGURE 37 CMS CURVES FOR ALL SIMILARITY MEASURES – TIME DELAY: PCA ..... 77
FIGURE 38 CMS CURVES FOR ALL SIMILARITY MEASURES – TIME DELAY: ICA ..... 77
FIGURE 39 CMS CURVES FOR ALL SIMILARITY MEASURES – TIME DELAY: LDA ..... 78
FIGURE 40 MCNEMAR'S TEST FOR THE TOP 3 INTER-CLASS ALGORITHMS – TIME DELAY ..... 80
FIGURE 41 HYBRID CMS CHARTS ..... 97

# *LIST OF TABLES*

TABLE 1 RANK-1 METRIC RESULTS FOR THE EXPRESSION PROBE SET ..... 51
TABLE 2 RANK-1 METRIC RESULTS FOR THE ILLUMINATION PROBE SET ..... 58
TABLE 3 RANK-1 METRIC RESULTS FOR THE LOWER OCCLUSION PROBE SET ..... 64
TABLE 4 RANK-1 METRIC RESULTS FOR THE UPPER OCCLUSION PROBE SET ..... 71
TABLE 5 RANK-1 METRIC RESULTS FOR THE TIME DELAY PROBE SET ..... 76
TABLE 6 RANK-1 COMPARATIVE HYBRID RESULTS ..... 97

# *LIST OF ABBREVIATIONS*

| | |
|---|---|
| AFGBR | AUTOMATED FACE AND GESTURE-BASED RECOGNITION |
| AFR | AUTOMATED FACE RECOGNITION |
| AVBPA | AUDIO AND VIDEO-BASED PERSON AUTHENTIFICATION |
| ARCH I | ARCHITECTURE I |
| ARCH II | ARCHITECTURE II |
| CASPEAL | CHINESE ACADEMY OF SCIENCES POSE, EXPRESSION, ACCESSORY AND LIGHTING |
| CMS | CUMULATIVE MATCH SCORES |
| CMU PIE | CARNEGIE MELLON UNIVERSITY POSE, ILLUMINATION AND EXPRESSION |
| COS | COSINE |
| EEG | ELECTROENCEPHALOGRAM |
| FAST A1/FA1 | FASTICA ARCHITECTURE I |
| FAST A2/FA2 | FASTICA ARCHITECTURE II |
| FERET | FACE RECOGNITION TECHNOLOGY |
| FR | FACE RECOGNITION |
| FRT | FACE RECOGNITION TECHNIQUES |
| GDA | GENERALISED DISCRIMINANT ANALYSIS |
| HCI | HUMAN COMPUTER INTERACTION |
| ICA | INDEPENDENT COMPONENT ANALYSIS |
| INFO A1/IA1 | INFOMAX ARCHITECTURE I |
| INFO A2/IA2 | INFOMAX ARCHITECTURE II |
| JAFFE | JAPANESE FEMALE FACIAL EXPRESSION |
| KDA | KERNEL DISCRIMINANT ANALYSIS |

| | |
|---|---|
| KLDA | KERNEL LINEAR DISCRIMINANT ANALYSIS |
| KPCA | KERNEL PRINCIPAL COMPONENT ANALYSIS |
| L1 | CITY BLOCK DISTANCE METRIC |
| L2 | EUCLIDEAN DISTANCE METRIC |
| LDA | LINEAR DISCRIMINANT ANALYSIS |
| MAH | MAHALANOBIS |
| PCA | PRINCIPAL COMPONENT ANALYSIS |
| RBF | RADIAL BASIS FUNCTION |
| UV | ULTRA-VIOLET |
| XM2VTS | EXTENDED MULTI-MODAL VERIFICATION FOR TELESERVICES AND SECURITY APPLICATIONS |
| 2D | TWO DIMENSIONAL |
| 3D | THREE DIMENSIONAL |

# *LIST OF SYMBOLS*

| | |
|---|---|
| A | MATRIX OF AVERAGE FACES AND IN SECTION 2.3 THE MIXING MATRIX |
| C | COVARIANCE MATRIX |
| c | NUMBER OF CLASSES IN TRAINING SET |
| D | USER SPECIFIED SUBSPACE DIMENSIONALITY |
| $f_Y$ | PROBABILITY DENSITY FUNCTION OF Y |
| H | ENTROPY FUNCTION WITH THE INFOMAX ALGORITHM |
| J | JACOBIAN |
| K | KERNEL MATRIX |
| M | NUMBER OF FACIAL IMAGES IN A SET |
| N | IMAGE-SPACE DIMENSIONALITY |
| P | MATRIX OF PROBE IMAGES |
| S | MATRIX OF UNKNOWN SOURCE SIGNALS |
| $S_B$ | BETWEEN-CLASS SCATTER MATRIX |
| $S_W$ | WITHIN-CLASS SCATTER MATRIX |
| W | TRANSFORMATION MATRIX |
| X | MATRIX OF INPUT IMAGES |
| Y | SUBSPACE MATRIX OF BASIS VECTORS |
| $\Psi$ | MEAN IMAGE |
| $\Phi$ | AVERAGE FACE VECTOR |
| $\Lambda$ | DIAGONAL MATRIX OF EIGENVALUES |
| $\Omega$ | PROJECTION MATRIX OF PROBE IMAGES |

| | |
|---|---|
| $\Phi$ | NON-LINEAR MAPPING FUNCTION |
| $\alpha$ | KERNEL COEFFICIENTS |
| $\beta$ | REGULARISATION FACTOR |
| $P$(H0) | MCNEMAR'S PROBABILITY VALUE ($P$-VALUE) FOR NULL HYPOTHESIS |

# Chapter 1

# Introduction

## 1.1 Introduction

Rapid technological advancement and the spurning Global Village effect have greatly fuelled research and growth within the areas of security. Digital security has extended itself to become a primary need within everyday life, whereby individuals seek to protect themselves, their assets and other values both at home and in work environments. Consequently the primary role of all security systems has evolved into protecting an entity by preventing or eliminating any and all possible threat factors. Currently one of the more predominant threats lies within the human persona, whereby security systems now serve to establish, maintain and protect the true relationship pertaining to the body and the identity of an individual.

The science of achieving the latter purpose lies in the field of Biometrics whereby various techniques, used to establish the true nature of an identity relationship, have been developed to embody the digital representation of an entity [1]. Biometric methods include identification by means of DNA Sequence Matching, Finger Print and Hand Geometry, Voice, Signature, Iris and Retina Recognition. All these techniques offer an excellent degree of reliability and have been used with great success in areas of forensics and law enforcements. However, the major stumbling block with these approaches lies in the

fact that they require the full knowledge, cooperation or special interaction on the part of the subject. Security systems, particularly in the areas of surveillance cannot afford this luxury and must be as non-intrusive as possible. Face Recognition (FR) has, therefore, become the most versatile and perhaps the most effective non-intrusive biometric that serves to automate the process in which it attempts to assign or confirm the identity of individuals by learning and recognizing facial information. Not only has its non-intrusive nature bolstered popularity, but the fact the face recognition is the primary means by which humans identify each other with little or no effort has inspired research to mimic this human ability in the areas of computer vision and consequently propelled the development of Automated Face Recognition (AFR) systems [1].

A human face is an extremely complex and dynamic object with features that can vary considerably and rapidly over time. Skin coverage offers a non-uniform material that is often difficult to model [2] and that can change quickly in response to the effects of emotion, temperature, reflectance properties and perspiration levels. The face is also highly deformable, creating a large variety and variability within the configurations of facial expression. Another avenue includes time-varying changes by measure of growth, facial hair, effects on the skin due to aging and skin colour changes attributed to ultraviolet exposure. A further complexity is introduced by artefact related changes such as change due to injury and fashion-related issues such as cosmetics, jewellery, hairstyles and garments [2].

Given the vast complexities surrounding modelling the human face, it would be obvious to state that any AFR system would be subject to these dynamics and will therefore inherently be faced with immense difficulties in matching, identifying, verifying and successfully recognising an individual within a given environment. The only factor that would remain constant would be that all systems would be built around the single complexity of change over time. All the major influencing factors can essentially be categorised into five barriers of concern: changes in pose (orientation and geometry),

illumination (lighting), expression (emotion), occlusion (artefacts and garments) and time (temporal) [3,4]. It is quite evident that computer modelling and recognition of the human face is not a trivial task and this is one of the reasons why the computer vision research community has spent such a long time devoted to the area of face recognition.

Over the last two decades, Face Recognition has gained significant attention and has made phenomenal progress in the fields of Image Analysis and Understanding, Pattern Recognition, Human Computer Interaction (HCI), Neuroscience as well as Psychology [5,6]. This is particularly evident over the last several years with the emergence of specific FR conferences such Automated Face and Gesture-Based Recognition (AFGBR) [7] and Audio and Video-Based Person Authentification (AVBPA) [8], systematic empirical evaluation techniques such as Face Recognition Technology (FERET) [9,10] and the Extended Multi-Modal Verification for Teleservices and Security Applications (XM2VTS) protocols [11]. Researchers have spent much time and effort developing and realising robust approaches that would overcome the barriers of face recognition and be favourably comparable to human recognition abilities. This has resulted in the field of FR becoming immensely diverse, with its beginnings in two dimensional recognition using geometric features and template based approaches, extending into appearance based methodologies and branching to model-based techniques such as Elastic Bunch Graphing and Active Appearance Modelling as well as a whole host of other approaches. Recent developments saw the incorporation of three-dimensional depth, texture and shape information, which has seen great success in alleviating many of the difficulties surrounding accuracy and performance of AFR systems.

Face Recognition Techniques (FRT) have, over the years, been optimised to an extent that commercial systems became a very profitable reality. Much of the world's best systems provide solutions for face detection, image registration, and image matching, all in real time [12]. Most of these algorithms find

their niche in sophisticated security systems for governments, corporations and research institutes. Although the details of most of these systems are confidential [12], many of the computer vision systems reported in literature still employ the popular appearance-based paradigm for object recognition [13]. Appearance-based analysis, which is one of the oldest approaches, is still said to give the most promising results [5].

Appearance-based methods belong to the family of holistic techniques, wherein global facial features are employed in recognition. An image of size m x n is represented in terms of optimal pixel information by means of a vector, in an m x n dimensional space that is used to uniquely identify one image/individual from another [3]. In practice, however, these m x n dimensional spaces are far to large to allow robust and fast recognition; this vast dimensionality naturally invokes the curse of dimensionality whereby statistical sampling demands, parameter requirements, ill-posed estimations and computational expense all grow exponentially and consequently contribute to a very poor solution approach [13]. A more efficient, practical and popular solution to resolve this problem is made possible by means of dimensionality reduction techniques known as subspace methodologies [13]. Given that similar data signals (face images) often belong to a manifold of intrinsically low dimension [14], subspace methods attempt to optimally find some transformation W for mapping images from the input space into a lower dimensional feature space. If X, for example, is a matrix of 'known' faces, recognition then becomes a two-step process, wherein both 'known' and 'unknown' faces are projected into the same subspace and then compared by means of a nearest-neighbour distance measure to realise the most likely classification [5,12].

Among the most popular publicly available subspace approaches are the classes of Principal Component Analysis (PCA – Turk and Pentland 1991), Independent Component Analysis (ICA – Lee 1998) and Linear Discriminant Analysis (LDA – Etemad and Chellappa 1997). Originally implemented

in a linear fashion, these methods may differ in the way the basis vectors Y and transformation matrix W are defined, but they share the common approach in which facial representation is extracted, such that $Y=W^T X$ (where $Y$ is the subspace matrix of basis vectors, $W^T$ is the transformation matrix and $X$ is the matrix of input images).

In PCA, the basis vectors are defined as the eigenvectors of the covariance matrix $C = XX^T$. So, the transformation matrix $W_{PCA}$ is composed of the D eigenvectors of C corresponding to the D largest eigenvalues. ICA is said to be a generalization of PCA in the sense that it can separate the high-order statistics of the input faces while PCA deals only with the first and second order statistical dependencies. LDA, on the other-hand, seeks to find a linear transformation by maximizing the between-class variance and minimizing the within-class variance [5]. Researchers have also found that the extrinsic and intrinsic variations in face images are highly complex and inherently non-linear in nature; meaning that the basic linear approaches above may fail to adequately capture these variations [14,15]. This has resulted in the successful developments of kernel-machine approaches in the form of Kernel PCA (Schőlkopf et al 1998), Kernel ICA (Bach and Jordan 2002) and Generalised Discriminant Analysis (GDA - Mika et al 2000) [17,18,19].

When looking at the performance of all these algorithms, it is interesting to note the often 'contradictory' and confusing claims that have been made in the literature. Bartlett et al. [20] and Liu et al. [21], for example, claim that ICA outperforms PCA, while Baek et al. [22] claim that PCA is better. Moghaddam [23] states that there is no significant difference. Beveridge et al. [24] claim that in their tests LDA performed uniformly worse than PCA, Martinez [13] states that LDA is better for some tasks, and Belhumeur et al. [25] and Navarrete et al. [26] claim that LDA outperforms PCA. While all these claims may in fact hold a good degree of truth, one should bear in mind that there were differing control factors surrounding each conclusion i.e. the actual task statement, the subspace distance

metrics, dimensionality retention and the non-standardised database choices etc [5,12]. All these conclusions have contributed to much debate and confusion over the years, particularly for an individual who is new to the field of FR and subspace methodologies and who seeks a good comparative understanding of the available techniques.

Very rarely are all the classes compared in the same investigation and almost never are all of their implementations considered. This research serves to provide a platform of equal conditions upon which the popular appearance-based subspace techniques can be fairly and properly benchmarked. In doing so one hopes to realise an independent comparative study that will greatly contribute to previous work, either by offering further insight into certain conclusions or providing understanding as to why certain claims should be disputed or hold true only under certain conditions. This investigation will compare the appearance-based methodologies of PCA, LDA and ICA in both linear and kernel projections. Also both ICA architectures I and II [12,20] as well as both the InfoMax and FastICA implementations will be reviewed. The four popular and most widely used distance measures of L1 (City Block), L2 (Euclidean), Cosine and Mahalanobis have been chosen as the comparative classification metrics. The performance effects of illumination, expression, occlusion and time variations are provided by the AR Database (Martinez and Benavente 1998 [56]) and will be compared across all the techniques to conclusively yield the 'best' algorithm in each category.

While it may be possible to find the most robust classifier in each category, it is extremely difficult for a single appearance-based classifier to perform well across all categories [27]. It has been shown in [28] that each classifier has a different sensitivity to different changes in facial appearance and consequently offers different performance levels on different categories of images, suggesting that different classifiers could contribute complementary information with regards to the classification task. This investigation will therefore propose a hybrid formulation that combines the 'best' algorithm from

each category, in an effort to realise a combination scheme that will yield improved results across all testing categories. Following Kittler's theoretical framework on combining classifiers [29], the techniques will be combined at the matching score level using the sum rule strategy. The hybrid will of course be compared to component classifiers to provide a better overall understanding into appearance based subspace methodologies and their performance within the face recognition environment.

The proceeding sections of this report are organised as follows: Chapter 2 will explore the background and mathematics surrounding subspace methodologies and the techniques being investigated. Chapter 3 will provide the implementation which includes the database selection, pre-processing and division as well as the training and testing methodologies employed. Chapter 4 will investigate the results and provide the necessary analysis. Chapter 5 will realise the hybrid formulation and effective comparisons. And finally Chapter 6 will conclude on the findings and provide the reader with sound comparative insight into the performance of appearance-based subspace techniques.

# Chapter 2

# Background

## 2.1 Introduction

A two dimensional image, $\Gamma(x,y)$ of size $m$ (rows) by $n$ (columns) pixels can generally be represented by concatenating the raster ordered values to create a vector in an $N$ dimensional *image space* ($R^{N=m \cdot n}$). This image space, however, constitutes a rather high-dimensional space and recognition therein would be deemed computationally infeasible [14, 30]. If, however, an image of an object (say a face) is considered to be a point in the image space, then a set of $M$ facial images can be represented as a set of points (samples of probability distribution) in the same confined *subspace*, referred to as the *face space* [14].

Theoretically it is common to model this subspace as a (possibly disconnected) lower-dimensional *principal manifold*, embedded in a higher dimensional image space [14, 30], wherein the *intrinsic* dimensionality is determined by the number of degrees of freedom within the face space. Gong *et al* [31] has shown that this intrinsic dimensionality, despite the variations in pose, expression and lighting, is very much smaller than that of the image space. The goal of subspace analysis is thus to determine the value of this dimensionality and thereafter extract the *principal modes* of the underlying manifold, while retaining as much information (energy) from the original images as possible [5]. By doing this,

subspace methodologies ensure that computational efficiency and hence the successful viability of face recognition algorithms can be achieved [30].

Appearance-based subspace analysis computes the principal modes, referred to as *basis vectors*, as a function of the image pixel values and provides a means of establishing a relationship between points in the subspace and all the images of the corresponding object (the face). It thus makes it possible to characterise and hence efficiently classify various image subsets [30]. Among the most popular and widely proven subspace analysis tools are the algorithms of PCA, LDA, ICA and more recently their non-linear variations in the form of kernel approaches [32]. Each has been derived based on different statistical viewpoints and each offer different advantages relative to the task statement at hand [12,30]. The proceeding subsections, beginning with the linear variants, will describe each of these classical appearance-based classifiers, offering insight into their historic and mathematical backgrounds within the context of face recognition.

## 2.2 Principal Component Analysis (PCA)

The PCA algorithm, also known as 'Eigenspace Projection' or 'Karhunen – Loeve Decomposition', is perhaps the most widely used unsupervised subspace projection method in face recognition. It seeks an optimal linear least-squares projection that best represents a given data set [12, 33]. Kirby and Sirovich [34] first applied PCA to representing faces, which was later extended by Turk and Pentland to recognising faces in their famous 'Eigenfaces' framework [35]. Since then much work has been contributed toward eigenface recognition including view-based and modular eigenspaces by Moghaddam and Starner (1994), colour eigenfaces by Finlayson *et al* (1996) and more recently Bayesian application in Eigenspaces by Pentland and Moghaddam (1998) [36,37,38].

Recalling that the set of all training images $M$ are stored in matrices of dimension ($m$ x $n$); each of these images are reshaped into column vectors of length $N = mn$. Each vector is then arrayed in a column-wise manner, to form the ($N$ x $M$) dimensional image matrix $X = [x_1,\ldots,x_M]$. The first step in creating the feature space is to subtract the average image from the training set. The mean image $\Psi$ is given by:

$$\psi = \frac{1}{M}\sum_{n=1}^{M} x_n \tag{1}$$

Each face image now differs from the average face by a vector $\Phi$, such that:

$$\phi_i = x_i - \psi \quad for\ i = 1{:}M \tag{2}$$

These vectors are then sequentially arranged into an ($N$ x $M$) dimensional zero-mean (scatter) training matrix $A = [\Phi_1,\ldots, \Phi_M]$. It is this matrix that shall then be subject to PCA [30].

The next step is to determine the projection or transformation matrix $W = [w_1,\ldots,w_D]$, which comprises of the top $D$ eigenvectors that best describes the distribution of the input data in a norm sense i.e. the Euclidean projection error is minimised [30]. $D$ is the desired user-specified subspace dimensionality associated with the eigenvectors capturing the top $D$ energy levels. This is achieved by defining the $N$ dimensional data covariance matrix $C$ as:

$$C = AA^T \tag{3}$$

whereby the principal components are then computed by employing eigen-decomposition to solve

$$W^T CW = \Lambda \tag{4}$$

where $\Lambda$ is the diagonal matrix of eigenvalues and $W$, as previously mentioned, is the matrix of orthonormal eigenvectors [12]. $W$ is a rotational matrix, which rotates the original coordinate system onto the eigenvectors, where the eigenvector associated with the largest eigenvalue defines the axis of maximum variance, the eigenvector associated with the second largest eigenvalue is then the orthogonal axis with the second largest variance and so on and so forth.

Ideally a PCA transformation that retains all the data information is limited to dimensionality of $M$-1, however, the proficiency of PCA lies in the fact that one can eliminate all but the best $D$ ($D \leq M << N$) eigenvectors and associated eigenvalues. This is based on the strong assumption that the mean and variance are sufficient statistics to entirely describe the data. The only zero-mean distribution that follows this, is the Gaussian distribution, meaning that vectors with the largest variances will contain the most important (*principal)* information and the axes of small variance are probably noise, eliminating these axes would therefore improve matching accuracy [12,30].

Each normalised eigenvector has the same dimensionality as the original face images and when rearranged into an image matrix, the image yields the impression of a 'ghost-like' face, which is dubbed the *Eigenface* [30,35]. The projection matrix is used to compute the feature (*basis*) vectors within the matrix $Y = [y_1,\ldots,y_M]$, that defines the new $D$ x $M$ dimensional subspace. This is accomplished by the linear transformation:

$$Y = W^T A \tag{5}$$

Each original training image can thus be formulated by the mean image plus a linear combination of eigenfaces, whereby the specific weighting of each eigenvector is defined by the respective basis vector.

The projected data (feature vectors) are then used to calculate, what is termed the 'class-centroid' i.e. the mean subspace point associated with all the data specific to each class [33]. The resulting class-centroid matrix is now of dimension ($D$ x $c$), where $c$ is the number of subjects/classes in the training set. It is this matrix that defines the *face space* and is used in the final recognition phase. Given a set of $P = [p_1,\ldots,p_P]$ probe images, each image can be classified by rearranging the pixels values in a vector format, similar to the training images, subtracting the mean image $\Psi$ and projecting the resulting matrix into to same face space such that:

$$\Omega = W^T (P - \psi) \tag{6}$$

where Ω comprises the column-wise projection vector of each probe image. Classification is then performed by comparing the nearest-neighbour distance measures between each probe vector and every centroid vector in the face space. Various similarity metrics can be used, the most popular being the measures of City Block, Euclidean, Cosine and Mahalanobis. Class assignment decisions are then obtained by selecting the class whose projection point is closest in measure to the probe image (maximum matching score) and whose distance falls below a predefined face threshold [30]. This class then provides the best possible match that describes the test image and is consequently assigned as the resulting classification; otherwise the image is classified as 'unknown' and can perhaps be added to a future training set [2].

## 2.3 Linear Discriminant Analysis (LDA)

While it is true that the Eigenface approach is successful in selecting a subspace that retains most of the data variation, it does not however, employ any optimality properties that facilitate class discriminality [14,39]. The training data is taken as a whole, wherein projections are made very close together resulting in information defining class membership and identity being mixed or lost all together i.e. PCA makes no use of between-class scatter or how many images may belong to the same individual [2]. The question then arises, how does one utilise label information in creating informative projections that may offer better classification recognition rates?

Linear Discriminant Analysis, also referred to as Fisher's Linear Discriminant (FLD) [40], is one such method that exploits class information to realise differentiation between individual faces but recognises faces of the same class [30]. First developed by R.A. Fisher in 1936 [40], it was not until recently,

1997, that LDA was first applied to facial identification by Belhumeur *et al* in their famous *Fisherface* approach [25]. LDA serves to provide basis vectors that yield the largest mean differences between the desired classes and in doing creates a linear subspace that best discriminates between classes, rather than one that best describes the data [13].

Mathematically, this is achieved by calculating an optimal projection matrix $W_{\mathrm{LDA}}$ such that the Fisher Discriminant Criterion is maximised in the form of:

$$W_{LDA} = \arg\left( \max_W \frac{|W^T S_B W|}{|W^T S_W W|} \right) = [w_1, ..., w_D] \tag{7}$$

The goal is then to maximise the between-class measure given by

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \psi)(\mu_i - \psi)^T \tag{8}$$

and minimise the within-class measure

$$S_W = \sum_{i=1}^{c} \sum_{x_m \in X_i} (x_m - \mu_i)(x_m - \mu_i)^T \tag{9}$$

where $N_{\mathrm{i}}$ is the number of training samples per class $i$ and $\mu_{\mathrm{i}}$ is the mean vector of samples belonging to class $i$ and $X_{\mathrm{i}}$ represents the set of samples belonging to a particular class. $S_B$ represents the scatter of features around the overall mean for all face classes and $S_W$ represents the scatter of features around the mean of each face class [30]. Thereafter it is not difficult to show that the solution to the maximisation problem in (7) lies in the solution of a generalised eigensystem given by:

$$S_B.W = \Lambda.S_W.W \tag{10}$$

where $W = [w_1,\ldots,w_{\mathrm{D}}]$ is the eigenvector (fisherface) matrix and $\Lambda = [\lambda_1,\ldots,\lambda_{\mathrm{D}}]$ are the corresponding eigenvalues of the between and within-class scatter matrices [30]. In directly solving this eigensystem, one can run into problems that may include the fact that:

i) The eigensystem does not have orthogonal eigenvector because $S_W^{-1}.S_B$ is, is general, not symmetric [30].

ii) The matrices of $S_W$ and $S_B$ are usually too large, given the image pixel dimensionality.

iii) There are at most $c - 1$ nonzero generalised eigenvectors meaning that $D$ is limited to an upper bound of $c - 1$ [13] and

iv) Because the rank of $S_W$ is at most $X - c$, and often the number of images in the learning set $X$ is much smaller than the image dimensionality, $S_W$ easily becomes singular and hence non-invertible [13,30] .

In addressing the latter problems, particularly the issue of singularity it was proposed, by Belhumer [25] and Martinez [13], that an intermediate subspace be used before projecting the image into the final LDA subspace. This intermediate space was achieved by employing the PCA algorithm prior to LDA projection. This method known as the *Fisherface* approach [25], found much success in improving recognition results and maintaining the non-singularity of $S_W$ [39,41].

After solving for the eigenvectors and hence the optimal transformation matrix $W_{\mathrm{LDA}}$ is found, projection follows very similar to there PCA algorithm whereby the basis vectors are again formulated by calculating the dot product of the images with each of the eigenvectors such that:

$$Y = W_{LDA}{}^{T}(X - \psi) \tag{11}$$

Recognition again follows the PCA methodology, whereby the probe images are projected into the same face space as the training centroids and classification is achieved by virtue of nearest-neighbour distance metrics.

## 2.4 Independent Component Analysis (ICA)

As described in Section 2.1, PCA's performance relies on one important assumption, the notion that input data always follows the Gaussian probability distribution. When this assumption holds true, the covariance matrix is able to capture all the information describing the zero-mean input variables i.e. PCA is only concerned with second-order statistics that only capture the amplitude spectrum of images [30, 42]. When one empirically observes true signal sources such as speech, natural images and EEG signals etc. one finds that these signals do not follow the above mentioned assumption, but instead have a more general long-tailed distributions that fall under the category of *high kurtosis* or *super-Gaussian* sources [20], meaning that complete representation lies in higher-order statistics.

These higher-order statistics capture the phase spectrum of an image and it is the phase spectrum that is believed to contain the structural information of images that facilitates human perception. With respect to faces and other natural images, PCA's limited sensitivity to only the power spectrum suggests it might not be particularly well suited for facial representation [20]. Independent Component Analysis was developed to minimise both second and higher-order dependencies in an effort to provide a more powerful data representation.

ICA finds its origins in the area of *blind source separation* (BSS) and has been rigorously defined over the years particularly by Jutten and Heralt [43] and Comon [44]. ICA serves to decompose an observed signal into a linear combination of unknown independent signals as opposed to signals that are merely uncorrelated [2,12]. Mathematically, if one denotes $S = [s_1,\ldots,s_D]$ as the matrix of unknown source signals, $X = [x_1,\ldots,x_M]$ as the matrix of observed mixtures and $A$ is the unknown mixing matrix, then the mixing model can be defined as:

$$X = AS \tag{12}$$

If it is assumed that the source signals are independent of each and that the mixing matrix $A$ is invertible, the ICA algorithm then attempts to find $A$ or the separating (transformation) matrix $W$ such that

$$Y = WX = WAS \tag{13}$$

Y is the estimation of the independent signals [12] or basis vectors that under ideal conditions follow the conditions that $W = A^{-1}$ and $Y \approx S$ [32].

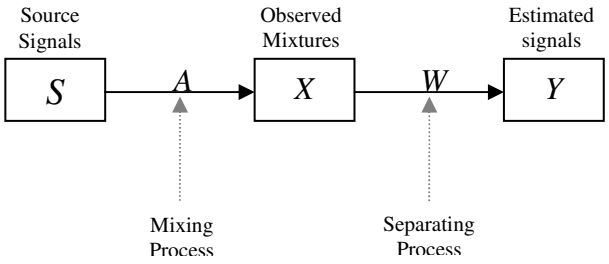



**Figure 1 Blind Source Separation model**

ICA can be viewed as a generalization of PCA, whereby data is not only initially decorrelated but also the constraint of unit variance is implemented. This is realized by a whitening transform [12] that results in the orthogonal signal separation, ICA then separates this whitened data into independent signals. These signals are said to be statistically independent when

$$f_y(y) = \prod_i f_{y_i}(y_i) \tag{14}$$

where $f_y$ is the probability density function of $Y$. Unfortunately there is no closed form expression to find any $W$ that fully satisfies this independence condition. Instead several algorithms have been developed to iteratively approximate $W$ so as to maximise independence by optimising a smooth function whose global optima occurs when the basis vectors in $Y$ are independent [30].

Among these algorithms the most popular and widely proven are the approaches of InfoMax [45], JADE [46] and FastICA [47]. Bell and Sejnowshi's InfoMax algorithm relies on the observation that independence is maximised when entropy H($y$) is maximised by:

$$H(y) \equiv -\int f_y(y) \log f_y(y) dy \tag{15}$$

where gradient ascent is applied on the elements of $W$ so as to maximise H($y$). The JADE algorithm minimises kurtosis of $f_y(y)$, whereby statistical independence is also maximised through the minimisation of a join diagonalisation of fourth order cumulants [12]. The FastICA algorithm is arguably the most efficient and general approach that maximises independence by maximising the function

$$J(y_i) \approx c\left[E\{G(y_i)\} - E\{G(v)\}\right]^2 \tag{16}$$

where $G$ is any non-quadratic function, $c$ is any positive constant and $v$ is a Gaussian random variable [47]. All these algorithms should ideally arrive at the same solution, but practically, the different approximation formulations find different solutions. This has again led to differing literary opinions regarding the performance superiority of these algorithms [12].

Regardless of which algorithm is used, within the context of face recognition there are fundamentally on two ways in which ICA can be applied. Bartlett *et al* [20] first proposed these approaches that became commonly known as *Architecture I* and *Architecture II*, whose basic differences are illustrated and described on the next page.

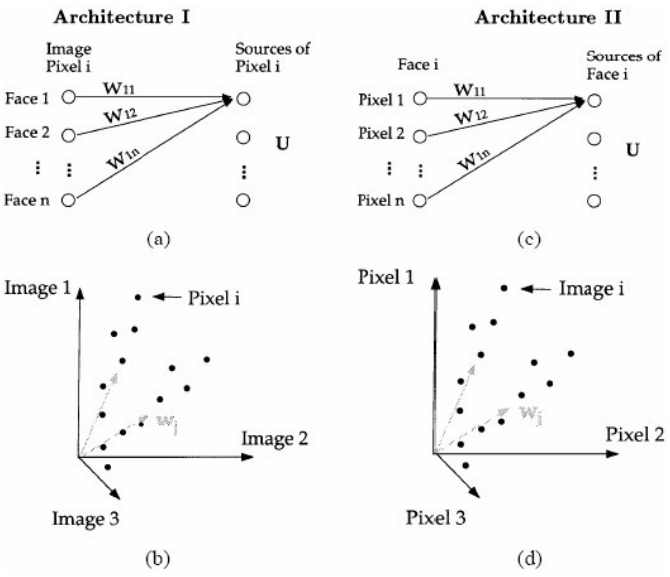



**Figure 2 Two architectures for performing ICA on faces [20]**

(a) Arch I finds statistically independent basis images, whereby source separation yields IC images in rows of *U*. (b) Gray values at pixel location i are plotted for each image. Weight vectors are found in the directions of statistical dependencies *among the pixel locations*. (c) Arch II applies source separation to the pixels and produces column-wise factorial code in *U*. (d) Weight vectors are now found in the direction of statistical dependencies *among the face images*.

### 2.4.1 Architecture I: Statistically Independent Basis Images

In *Arch I*, the input matrix *X* is arranged so that each row vector is a different image [20]. The images are considered to be the random variables and the pixels now become the trials, such that *X* is considered a linear mixture of statistically independent basis images *S*, combined by an unknown matrix *A*. ICA then learns that weight matrix *W* such that the rows (basis images) of $Y = WX$ are as statistically independent as possible [30]. The compressed representation of a face image is now a vector of coefficients that are used for linearly combining the basis images in reconstructing the input image. An example of these basis vectors are illustrated in the second row of figure 3 below.

An important observation is that these basis vectors produce images that are spatially localised i.e. each row of the mixing matrix *W* represents a cluster of pixels that have similar behaviour across all the images [30]. Unlike those of PCA (1st row) and ICA *Arch II* (4th row), where every image feature is

influenced by every pixel so to produce global features, the features of *Arch I* are only influenced by small parts of an image. This attribute has been argued to produce object recognition that is superior to PCA, LDA and ICA *Arch II*, particularly in areas of expression and occlusion [12].

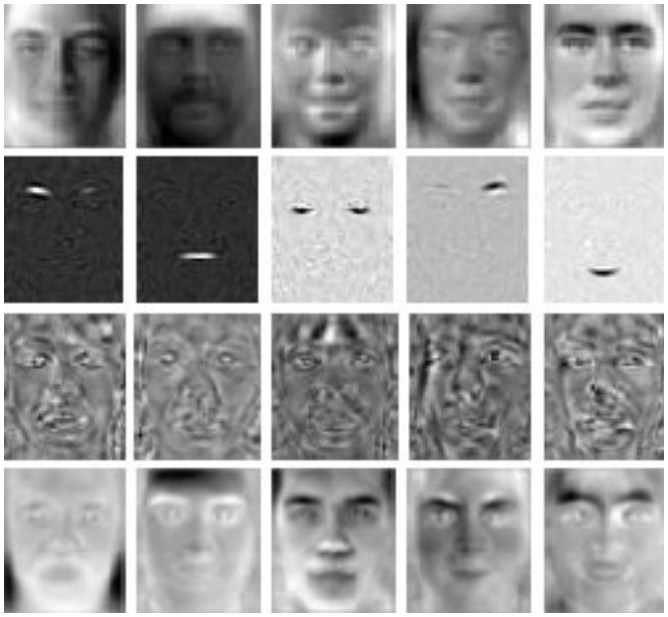


**Figure 3 Illustrative example of subspace basis vectors [30]**
PCA (1st row), ICA Arch1 (2nd row), LDA (3rd row), ICA Arch2 (4th row)

### 2.4.2 Architecture II: Statistically Independent Coefficients (Factorial Face Code)

Although the basis images of *Arch I* are statistically independent, the coefficients that define the subspace are not [12, 30]. *Arch II* was developed with the particular goal of finding independent coefficients for the input data. The images of the input matrix $X$ are now arrayed in a column-wise fashion such that the pixels now become the variables and the images are the observations. The source separation is therefore performed on the pixel values, and each row of $W$ is an image, whereby the columns of $A$, ($A = W^{-1}$), form the basis images that can be rearranged and viewed as shown in the fourth row of figure 5 above. The statistically independent source coefficients in $S$ are then recovered in the columns of $Y$ [12]. In this approach, ICA determines the mixing matrix by attempting to find a

cluster of images that are similar across all pixels and in doing so *Arch II* generates global basis images that are more refined and face-like than the images produced by PCA [30].

In implementing both ICA algorithms it has become common practice to use PCA as a pre-processing step. First applied by Bartlett *et al* [20], the input data is initially projected into a $D$ dimensional subspace to control the number of independent components produced by ICA. In [20], it has been argued that pre-applying PCA enhances ICA-performance by (1) reducing dimensionality and hence computational complexity by minimising the pair-wise dependencies and (2) discarding small trailing eigen-pairs before whitening. PCA performs the data decorrelation and ICA then separates the higher order dependencies [12].

In summary face recognition using ICA can be carried out as follows: the training images are initially projected into an intermediate $D$ dimensional PCA subspace, which is then separated into $D$ independent components (being basis images or source coefficients, depending on the Architecture). These components are then used to formulate the centroid matrix against which projected probe images are compared and classified. Again, this is achieved by virtue of similarity measures, whereby the matching takes places by assigning the class label to the probe image that is closest in nearest-neighbour measure.

## 2.5 Kernel Methods

Linear Subspace techniques have demonstrated much success in numerous computer vision applications, particularly in the area of facial recognition. However, due to their linear nature, particularly in the case of PCA and LDA, they can at best only describe pattern information based on second order statistics [1]. Variations in face images due to viewpoint, illumination and expression

changes have been proven to be highly complex and nonlinear in nature [14, 48] and it has been observed that variations between face images of the same person due to illumination and pose are almost always greater than image variations between the different persons [15]. From a classification viewpoint, linear approaches are therefore said to be suboptimal in terms of accurate data representation. Complete pattern variation is said to be captured within the non-linear relations between neighbouring (three or more) pixels [1, 50]. These relationships are represented in terms of higher order statistics that are crucial in fully representing complex patterns [50].

The concept of ICA, as discussed earlier, was successfully developed to represent higher order dependencies, however, it has been argued that ICA does not provide any significant advantage over PCA and that the linear non-Gaussian seeking approach may not necessarily provide a better representation for face recognition [23, 50]. Many researches believed that a non-linear scheme that still provided computational feasibility would better address this problem. It was in the concept of kernels methods that their solution was realised [50].

First introduced by Aizerman *et al* [51], the 'kernel trick' was used to map the input space to, by means of a nonlinear function expressed as dot products, to a convenient feature space (Hilbert space) in which the input data is nonlinearly related [1]. With the exception of Support Vector Machines [52], kernel methods remained largely unused in the machine learning community. It was not until recently that Schölkopf *et al* [17] extended the classical PCA algorithm to Kernel Principal Component Analysis (KPCA) that was shown to be able to extract non-linear features and in doing so provide better recognition results than PCA [17,50]. KPCA, as with PCA, simply captures variance information and although being nonlinear, it may not necessarily be suitable for discriminatory purposes [25]. Mika *et al* [19] and Baudat and Anouar [53], then proposed Kernel Linear Discriminant Analysis (KLDA), whose results are claimed to be superior to that of PCA, LDA, ICA and KPCA [14].

Kernel ICA was also introduced by Bach and Jordan [18] in support of constrained theories they proposed regarding the estimation of Gram matrices in Cholesky Decomposition. The algorithm employs gradient decent optimisation approach and was designed to operate using low-rank results. Its computational demand is said to increase cubically with dimensionality, thereby currently deeming it infeasible for application in face recognition [18] and therefore placing its evaluation beyond the scope of this investigation.

### 2.5.1 Kernel Principal Component Analysis (KPCA)

Given a zero mean, unit variance input matrix $X \in R^N$, PCA attempts to find the projection matrix that captures the maximum variance of a subspace, which is equivalent to performing eigen-decomposition on (4), i.e. $\Lambda W = CW$ for eigenvalues $\Lambda \geq 0$ and eigenvectors $W \in R^N$. In KPCA, each input vector $x$ is projected from the input space to a higher dimensional feature space $R^F$, by using a non-linear mapping function $\boldsymbol{\varphi}$: $R^N \rightarrow R^F$ where $F > N$ [50]. The Eigen problem is now defined as:

$$C^{\varphi} W^{\varphi} = \Lambda W^{\varphi} \tag{17}$$

where $C^{\varphi}$ is the covariance matrix defined by:

$$C^{\varphi} = \frac{1}{M} \sum_{j=1}^{M} \varphi(x_j) \varphi(x_j)^T \tag{18}$$

and all the solutions $W^{\varphi}$ lie in the span of $\boldsymbol{\varphi}(x_1),.., \boldsymbol{\varphi}(x_M)$ There exist coefficients, $\boldsymbol{\alpha}$, such that

$$W^{\varphi} = \sum_{j=1}^{M} \alpha_j \varphi(x_j) \tag{20}$$

Due to the increase in dimensionality, the mapping process is made implicit and economical by use of kernel functions that satisfy Mercer's theorem [14] and allow the ($M$ x $M$) kernel matrix $K$ to be defined as

$$K = k(x_i, x_j) = (\varphi(x_i) \cdot \varphi(x_j)) \quad \text{for } i,j = 1,\ldots,\text{M} \tag{21}$$

The kernel evaluations $k(x_i,x_i)$ correspond to the dot products in the higher dimensional feature space. PCA can thus be formulated without explicit and expensive computation of $\boldsymbol{\varphi}(\text{x})$ [14]. Thus the KPCA eigen problem can be expressed as

$$M\Lambda\alpha = K\alpha \tag{22}$$

where $\boldsymbol{\alpha}$ denotes the coefficient column vector with entries $\alpha_{1,\ldots,}\ \alpha_{\text{M}}$. By diagonalising $K$ one can obtain the eigenvectors $W$ and hence project the vectors in $R^F$ to a lower dimensional subspace. Subsequently the non-linear principal components (Kernel Eigenfaces) of any input vector can be efficiently computed with simple kernel evaluations against the dataset [14]. Letting $\mathbf{x}$ be a test sample, the projection of $\boldsymbol{\varphi}(\mathbf{x})$ onto the eigenvectors $W^{\varphi}$ is given by:

$$Y = W^{\varphi} \cdot \varphi(x) = \sum_{i=1}^{M} \alpha_i k(x, x_i) \tag{23}$$

In other words one can obtain a $D$ – dimensional projection manifold of $\mathbf{x}$, whose individual components are defined by the component vectors of $Y$. Classical PCA is said to be a special case of KPCA, being characterised as a subspace projection employing a first order polynomial kernel [50]. KPCA is considered a generalisation of maximal variance representation and can use different kernel functions such as Gaussian, Polynomial, and Sigmoid etc. to satisfy Mecer's theorem relative to finding optimal projection for a given application.

### 2.5.2 Kernel Discriminant Analysis (KDA)

Similar to the derivation of KPCA, Mika *et al* [19] and Baudat [53] extended the kernel approach to LDA to formulate Kernel Discriminant Analysis (KDA) or as it's more commonly known as Generalised Discriminant Analysis (GDA) [33]. Assuming, again, that the projected samples $\boldsymbol{\varphi}(X)$ are

centred in $R^F$, one can formulate the equations that employ dot products for LDA only. Denoting the between-class and within-class matrices, now, by $S_B^{\varphi}$ and $S_W^{\varphi}$ respectively and applying LDA in kernel space, the generalised eigensystem is now given by:

$$S_B^{\varphi}.W^{\varphi} = \Lambda.S_W^{\varphi}.W^{\varphi} \tag{24}$$

which can be obtained by maximising the Fisher Criterion such that

$$W_{KLDA}^{\varphi} = \arg\left( \max_{W^{\varphi}} \frac{|(W^{\varphi})^T S_B^{\varphi} W^{\varphi}|}{|(W^{\varphi})^T S_W^{\varphi} W^{\varphi}|} \right) = [w_1^{\varphi},...,w_D^{\varphi}] \tag{25}$$

where $W^{\varphi}_{KLDA}$ is the optimum set of generalised eigenvectors corresponding to the $D$ largest generalised eigenvalues. The scatter matrices are now defined as

$$S_B^{\varphi} = \sum_{i=1}^{c} N_i (\mu_i^{\varphi} - \psi^{\varphi})(\mu_i^{\varphi} - \psi^{\varphi})^T \tag{26}$$

$$S_W^{\varphi} = \sum_{i=1}^{c} \sum_{x_m \in X_i} (\varphi(x_m) - \mu_i^{\varphi})(\varphi(x_m) - \mu_i^{\varphi})^T \tag{27}$$

where $\mu^{\varphi}_{\mathrm{i}}$ is now the mean of class $i$ in $R^F$ [50]. Considering a $c$ – class problem, where the $\mathrm{i}^{\mathrm{th}}$ sample of class t and the $\mathrm{j}^{\mathrm{th}}$ sample of class u are $x_{\mathrm{ti}}$ and $x_{\mathrm{uj}}$ respectively, the symmetric kernel matrix can be defined as:

$$K = k(x_{ti}, x_{uj}) = (\varphi(x_{ti}) \cdot \varphi(x_{uj})) \tag{28}$$

One also defines a matrix $Z$ that is an ($M$ x $M$) block diagonal matrix and from the theory of kernel reproduction [53], any solution $W^{\varphi} \in R^F$ must lie in the span of all training samples in $R^F$, i.e.

$$W^{\varphi} = \sum_{l=1}^{c} \sum_{m \in X_i} \alpha_{lm} \varphi(x_{lm}) \tag{29}$$

It then follows that finding the solution for (29) can be found by solving:

$$KZK\alpha = \Lambda KK\alpha \tag{30}$$

And consequently (25) can be rewritten as

$$W_{KLDA}^{\varphi} = \arg\left( \max_{W^{\varphi}} \frac{|\alpha^T KZK\alpha|}{|\alpha^T KK\alpha|} \right) = [w_1^{\varphi}, ..., w_D^{\varphi}] \tag{31}$$

In a similar fashion to KPCA, a test image $\boldsymbol{\varphi}(x)$ can be projected to a lower dimensional space spanned by the eigenvectors $W^{\varphi}$ [53].

$$Y = W^{\varphi} \cdot \varphi(x) = \sum_{i=1}^{c} \sum_{m \in X_i} \alpha_{im} k(x, x_{im}) \tag{32}$$

Another point to note is that the kernel machine can inherently overfit the data and ideally one would want to add in a regularisation factor such that $S_W \rightarrow S_W + \beta \mathrm{I}$ Adding this term ensures that very small eigenvalues are bound away from zero which improves the numerical stability in computing the inverse and the extra term acts as a weight decay or penalty function that favours smaller values of $\alpha$ over larger ones [54].

## 2.6 Literary Performance Review

In assessing all the above algorithms one would like to formulate a comparative assessment of their relative performance within the environment of face recognition. Over the years various groups have reported differing performance results and except for consensus surrounding the importance derived from the specific nature of the task, no straightforward conclusion can be easily drawn with regards to performance superiority.

In comparing PCA and ICA, Bartlett *et al* [20] report that both ICA architectures outperform classical PCA in recognising images containing temporal changes. Draper *et al* [12] examined both the InfoMax and FastICA implementations of ICA *Arch I* and *II* and concluded that the FastICA with *Arch II* yielded better results than PCA pertaining to identity recognition, and InfoMax with *Arch I* performed

much better than PCA in recognising faces with expression changes. Baek *et al* [22], on the other hand, proposed clearly contradicting results, wherein PCA was found to outperform ICA (InfoMax). Moghaddam [23], however, found that there existed no statistical difference between the two algorithms.

Liu *et al* [21] investigated the performance of all three classifiers, PCA, LDA and ICA. It was shown that ICA, when implemented in 'a properly compressed and whitened space', performs significantly better than both the Eigenface and Fisherface approaches. In comparing PCA to LDA, Beveridge [24] claims that LDA performs uniformly worse than PCA, while Belhumer [25] and Navarette *et al* [26] clearly demonstrate LDA's superiority. Martinez [13], however concludes that PCA can outperform LDA but for certain tasks only, particularly when the number of training samples per class is small (one or two images per class) or when the training data non-uniformly sample the underlying distribution.

Delac *et al* [5] also performed an independent study of PCA, LDA and ICA using the popular distance metrics of City Block (L1), Euclidean (L2), Cosine and Mahalanobis. Recognition was categorised into areas of expression, illumination and temporal changes. It was found that no deciding claim could be made with regards to the task of expression. In the category of illumination changes, PCA with L1 outperformed LDA and ICA *Arch I* but only outperformed *Arch II* at higher ranks. ICA *Arch II* with Cos was found to offer the most superior results when it came to temporal changes.

Turning to the kernel implementations of PCA and LDA, Yang [55] showed that both KPCA and KDA provide better data representations and achieve lower error rates than standard PCA, LDA and ICA methods. In comparing the two kernel approaches both Navarette [26] and Pavlou [1] demonstrate that KDA outperforms KPCA for larger training sets.

# Chapter 3

# Implementation

## 3.1 Introduction

As discussed in the previous chapter, there is no straightforward conclusion that can be drawn regarding the overall performance of the three described classes of algorithms. At best it can be stated that each algorithm will perform superiorly for a specific task. Another observation is that there are currently not enough independently conducted surveys of these algorithms, particularly under the same initial conditions i.e. task definition, pre-processing, computational resources etc. Furthermore, never are all the possible implementations considered, let alone with various similarity measures [30].

The work presented in this paper is one of the few independently conducted surveys that will methodically test these popular algorithms by providing a common evaluation platform. In doing so, the initial task requirements and hence testing categories will be defined so that an effective database may be selected or synthesised. All the images will undergo the same pre-processing steps and will be divided into gallery and probe sets that will all be employed in evaluating each algorithm. A single evaluation methodology will be proposed so that all testing is performed in the same simulation environment and the same performance tools are used in effectively establishing results that can comparatively recommend categorised superiority.

## 3.2 Database Selection

Comparing one algorithm to another within the context of face recognition is not a simple task, one needs to firstly refine the nature of the tasks being evaluated. In doing so one considers the most influential factors that will adversely affect recognition rates and uses these factors to categorise the testing criteria upon which each algorithm can be assessed. As mentioned earlier, these factors fall into five main areas of focus [4]:

a) *Viewing angle* – The face being a 3D object being projected onto a 2D image is subject to the effects of camera pose changes i.e. the appearance of the face can change due to projective deformation that leads to stretching and foreshortening of different parts of the face or self and dis-occlusion of parts of the face, therefore making it difficult to recognise individuals from disparate angles. Investigating the functional relationship between viewing angle and recognition therefore becomes a pertinent category.

b) *Illumination* – As with pose variation, illumination variation is inevitable. Ambient lighting, changes greatly within and between daylight periods and among indoor and outdoor environments. Direct lighting sources may also cause shadows and shading that will accentuate or diminish certain facial features. Lighting changes trace back to one of two factors, either the inherent amount of light reflected of the skin or the internal camera controls such as gamma correction and exposure settings. Although these factors are fairly easily overcome by humans they can cause major problems in computer vision and hence illumination effects would form the next category.

c) *Expression* – The face is a non-rigid object, hence expression of emotion, paralinguistic communication and speech acts can produce large variations in facial appearance. Because facial expression affects the apparent geometrical shape and position of facial features, its influence on

recognition may be quite substantial. It is therefore essential to investigate the conditions that may either impair or improve accuracy.

d) *Occlusion* – The face may be occluded by many objects in a scene such as sunglasses or other paraphernalia that may be intentional or unintentional. Individuals, under certain circumstances, are often motivated to diminish recognition efforts by covering portions of their face. And since the goal in many situations is the recognition of non or un-cooperating individuals it is essential to evaluate the quantitative and qualitative effects that occlusion has on recognition.

e) *Time delay* – Faces change over time and whether this is due to changes in hairstyle, makeup, facial hair, muscle tension or aging the effect this has on recognition can become quite pronounced.

Having established the focal points of the investigation, one had to now find, create or synthesise a suitable database that effectively facilitated testing in all or most of the above categories. Within the sector of publicly available databases, there are many popular choices such as the CMU PIE database, Yale, JAFFE, CAS-PEAL, Cohn-Kanade and of course the 'world standard' FERET database [72]. All of these databases offer fantastic galleries, however they either tend to focus extensively on one category or ignore the dept required for complete testing i.e. relatively few samples per class. Before considering any database synthesis, one sort a database that overlapped the desired categories, particularly that of occlusion, and provided an efficient number of class samples to facilitate effective testing. This database was found in the form of the AR Database [56]. Created by A. Martinez and R.Benavente in June 1998 and it is publicly available from the Purdue University, USA. The database contains over 3000 images of 135 individuals (76 males and 60 females. The subjects were recorded over two sessions with a two week interval between shoots. During each session 13 (24-bit-RGB-colour) images per subject under varying conditions, 1 neutral and 3 per variant of expression, illumination and occlusion (with lighting changes). An example set is shown in figure 4.

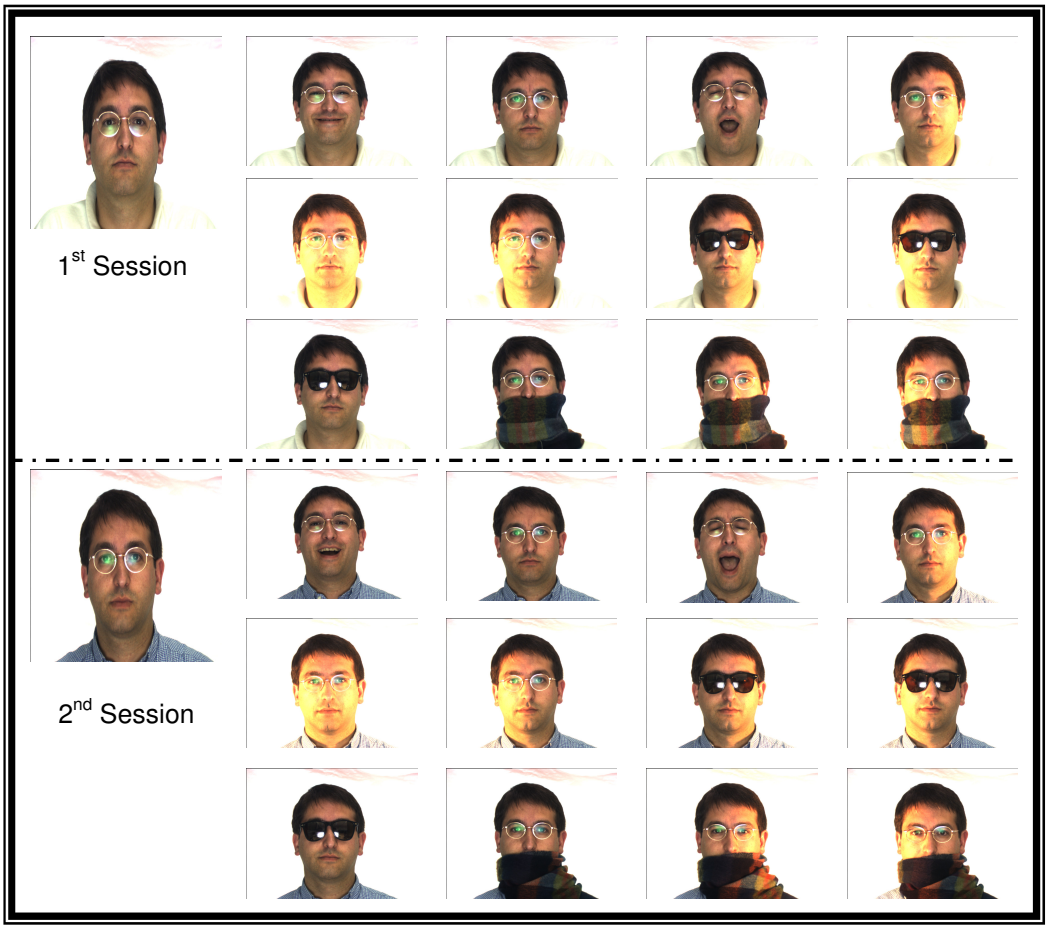



**Figure 4 Example of an individual set from the AR Database**

## 3.3 Software Tools

Much of the implementation was carried out in the Matlab™ (version 7.0) simulation environment, wherein one could easily integrate the mathematics, technical programming and visual analysis required for this comparative assessment. Particular use was made of the Image Processing Toolbox in conjunction with Statistical Pattern Recognition Toolbox [57], the FastICA Toolbox [58] and the ICA Architectural implementations [20]. These toolboxes were used in formulating the system modules used for this investigation, all of which can be explored in *Appendix B* (within the attached CD).

## 3.4 Image Pre-processing

### 3.4.1 Traditional Procedure

The pre-processing stage, depicted in figure 5, was carried out in a very sequential fashion. Every image in the AR database first underwent the *alignment* process, whereby the centres of the eyes and mouth were *detected* by means of a manually annotated procedure. The selected region was then *rotated* and *scaled*, while maintaining feature proportions, such that the centres of the eyes lied on the ocular horizontal line and the mouth was centred in co-ordance with the vertical median.

The selected region of each image was then *cropped* and *resized*, keeping the aspect ratio constant, to obtain face images of dimension (125 x 165 x 3) pixels. For computational efficiency and simplicity all the images were then *greyscaled* to a single plane of pixel values. In an effort to enhance the image intensity information, every image was further subjected to *histogram equalisation* so that the energy spectrum of all the pixel values was equally spread across the entire range. And finally in order to ensure that all the input image vectors are of the same energy, the images were *normalised* and readied for input into the subspace simulation environment.

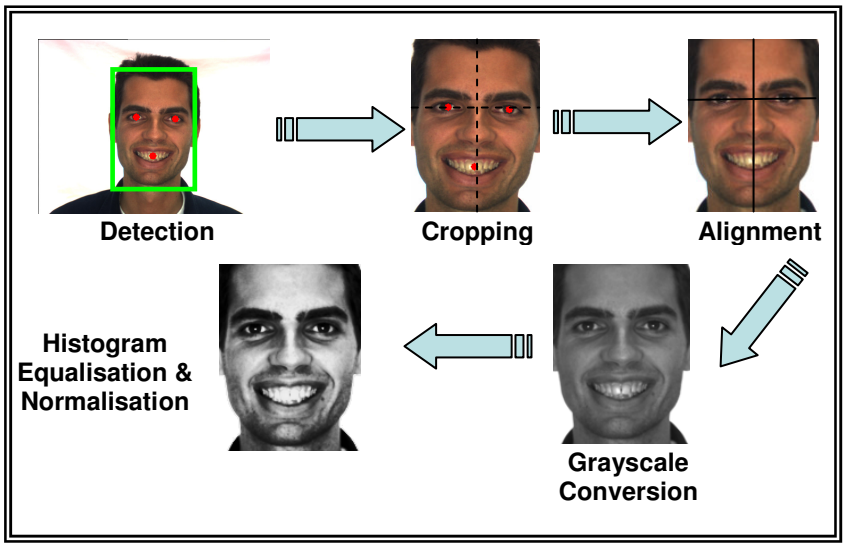



**Figure 5 Depiction of Pre-processing Procedure**

### 3.4.2 Alternative Techniques

While the techniques employed in the pre-processing stage are fairly basic and widely practiced, there also exist a number of other enhancement techniques, some of which were considered and for reasons that follow, were excluded from the implementation process. This section was included for the reader whose investigation task and perhaps pre-processing requirements differ from those of this paper.

The first approach concerns the area of *Localisation and Pose Compensation*. The alignment procedure, while quite effective, only aligns images about the *X*-axis of correction, looking at figure 6 we observe that there still exist images, wherein although the eye and mouth coordinates are aligned, the variation about the *Y* and *Z* axes still provides changes in viewing angle of up to 15°. Whether this variation is due to camera adjustments, inherent human nature to change position during the shoot or minor errors during the manual annotation process, ideally one would like to minimise the effects of pose variation on recognition. There are a number of compensation techniques that are currently being used such as *Warping* [13], *Neural Net Template Matching* and the use of *Depth Information* (3D approaches) etc. In this investigation however, one is not primarily seeking the best accuracy levels and coupled with the fact that the AR Database does not explicitly provide for pose variation, it was decided to leave the images as at the alignment stage and hence make pose variation implicit to the database and hence to the testing phase.

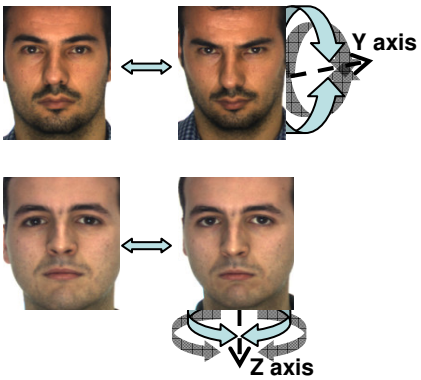



**Figure 6 Pose variation after alignment**

The second technique that was considered is that of *Masking*, whereby all the background information are removed so that only the pixels corresponding to the image face remains, an example is shown in figure 7 below. This technique was not included because it was found that after the images were cropped, over 90% of the image was facial data and hence masking was considered unnecessary.

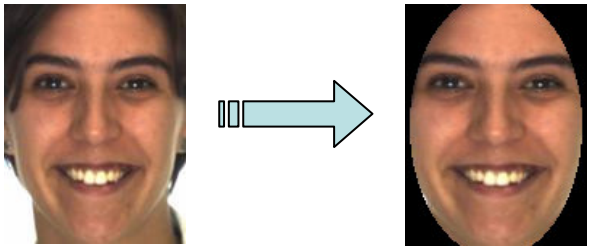


**Figure 7 An example of Masking**

The last area of concern is that of *Illumination Compensation*. There are many techniques that have been developed over the years, the most popular of which include: Illumination Gradient compensation, whereby one formulates image representations that are insensitive to illumination changes; exploitation of 3D information, whereby face models are used to generate images under synthesized lighting conditions and closed loop models that consider the variables of facial shape, texture and illumination dynamics in an effort to predict future representations of an individual under varying lighting conditions [59,60]. But again, this investigation seeks to test the effects of illumination and hence no compensation techniques were invoked.

## 3.5 Division of Gallery and Probe Sets

Upon completion of the pre-processing stage, the database now formulated a baseline system [12] and following the lead of Moon and Phillips [63], the algorithms are to be compared in this context i.e. no face detection or image registration is done, instead testing assumes that all the images (probe and

gallery) have been pre-registered. For every probe image there will be only one class label of that subject, so trials are essentially binomial in nature.

With the goal of having a balanced data set, in terms of number of male and female subjects, 110 subjects (55 males and 55 females) were randomly selected, yielding a database of 2860 images. Since it is desirable to have no overlap between the training and testing images, subject images were divided into as follows:

- The *Training Set* comprised of the neutral expressions of both the 1st and 2nd sessions. This offered all the necessary facial and temporal information and effectively facilitated a realistic investigation of geometrical changes for a small sized training set i.e. expression and occlusion over time given that few training samples may be available (2 x 110 = 220 images for training).
- The *Test Set* was divided into four categories: *Expression, Illumination, Upper Occlusion* and *Lower Occlusion*. Each test set contained 6 images per class/subject (3 per session, 660 images in total per category).

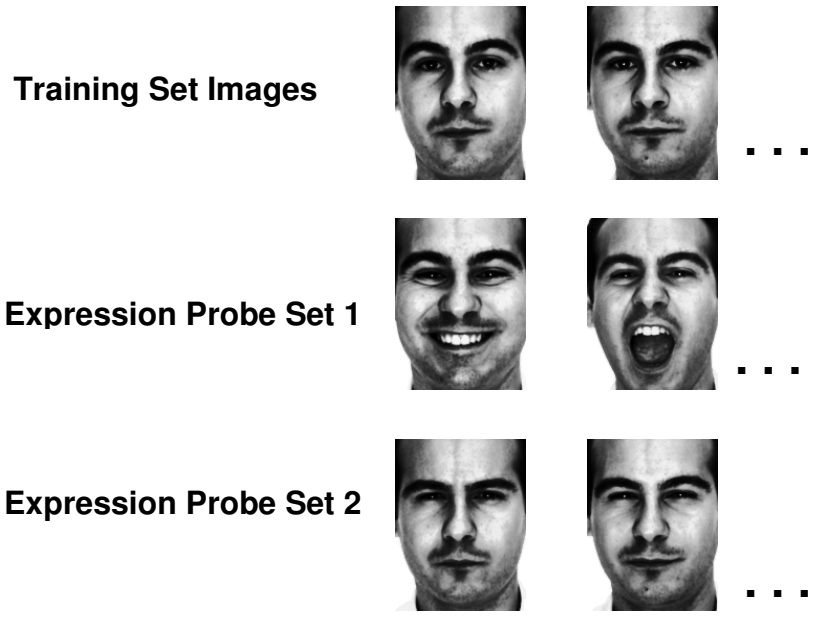



**Figure 8 Individual sample of Training and Probe Images**

- In order to achieve a good degree of confidence in the results, each probe set was further subdivided into 10 probe sets comprising 2 random images per subject (220 images per probe set). The results will then be presented as an average over the 10 test sets.

Although temporal changes are inherently considered in each category due to the session 1 and 2 images, it was also decided to directly compare the effects of time only. A further *Time* probe set was created, using only the neutral images i.e. images from the 2nd session are tested against the neutral images from the 1st session. Of course this task could be extended into various combinations, however this was left for future work.

## 3.6 Simulation Methodology

Subspace evaluation is essentially divided into two sections, the *training phase* where feature spaces for each algorithm are created and stored and the *testing phase* where probe images are projected into the same subspaces allowing a comparative assessment of classification performance to be formulated.

### 3.6.1 Training

All the images of the *Training Set* are rearranged into a column-wise format and sequentially stacked into a (20625 x $m$) input matrix, where $m$ = 220. The mean image is then subtracted from every column of the input matrix, leaving only the variances that are then projected into the subspace defined by each algorithm. The projections yield the matrix ($m$ x $Dim$, where $Dim$ is the user defined subspace dimension) of feature vectors that are then used to calculate the centroid matrix (2 images per class

label), which is in turn stored in the projection database. This process is repeated for each subspace technique being evaluated, thereby creating a projection database comprising of 10 subspaces.

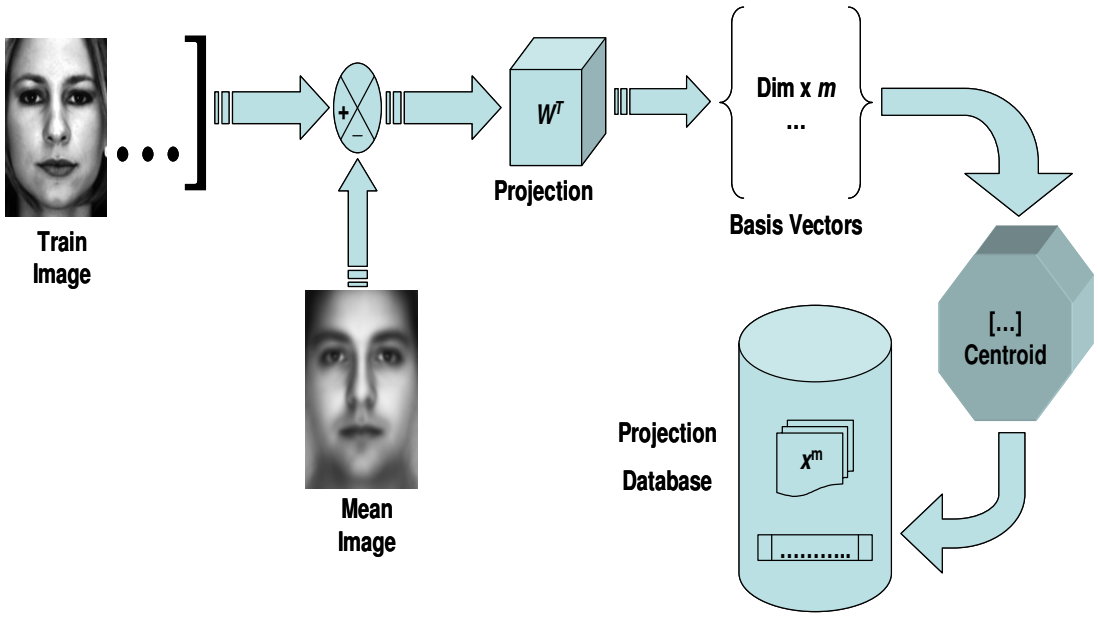



**Figure 9 Generic Training Model**

Before any projection or testing could begin however, one needed to obtain a *region of dimensional optimality* for each algorithm i.e. subspace dimensionality selection. Within this paper this region has been defined as a '*Region of constant differentiability, between the similarity measures, which offers comparable accuracies amongst the different algorithms while simultaneously contributing toward effective algorithmic generalisation and computational efficiency*'.

Currently there exists no closed form expression that explicitly determines this region and hence this choice was always left to individual discrimination; this is perhaps a significant contributing reason as to why such contradicting performance claims are found in face recognition literature. In order to create a comparable platform of individual works, one needs to define or follow an easily adoptable standard. One standard in particular is the FERET heuristic that suggests selecting 40% of the dimensionality would retain approximately 96% of the energy spectrum (signal information) [5,12,32]. The upper limit of subspace dimensionality is of course predefined by either the number of training images less one

($M$ – 1) in the case of PCA and ICA variants or the number of classes less one ($C$ -1) for LDA algorithms [32,50].

While the above heuristic may hold true, particularly for the FERET face database, it was decided to validate this statement with respect to images obtained from the AR Database. This was firstly accomplished by obtaining the energy/eigenvalue percentage contribution of each dimension and secondly selecting smaller random test sets, one image per subject, for each of the prior mentioned categories. These sets were then tested against the *Training Set* to yield rank 1 accuracies across the dimensionality spectrum. Having the energy spectrum, one could match the 96% contribution to the corresponding dimensionality and then relate this value to the accuracy performance of each algorithm and hence establishing the region of optimality. The results are shown in figure 10.

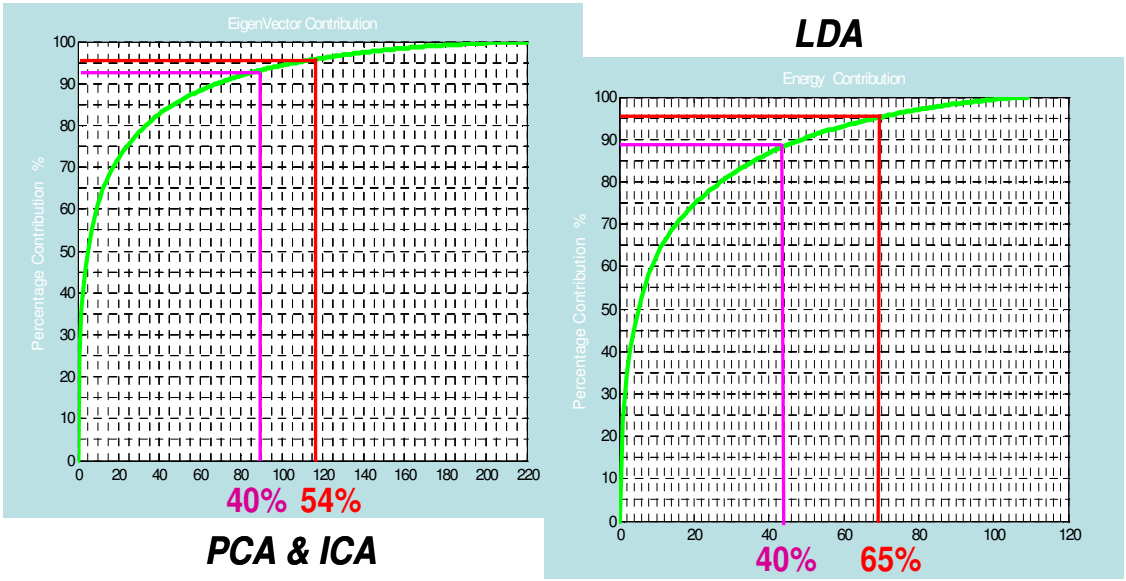



**Figure 10 Energy Spectrum for classes of PCA, ICA and LDA**

Looking at figure 10 above, clearly 40% of dimensionality does not equivocate to the 96% energy spectrum mark. Pertaining to the classes of PCA and ICA, this dimensional contribution is approximately 54% and lies around 65% when employing the LDA algorithm. Looking at a sample of rank 1 accuracy graphs in figure 11, one observes that the region of optimality lies between the

dimensions of 116 – 119 for PCA and ICA and between 72 – 74 for LDA. This region clearly supports computational efficiency, while simultaneously providing comparable accuracies that clearly do not belong to the, perhaps poorly generalised, region of highest accuracy.

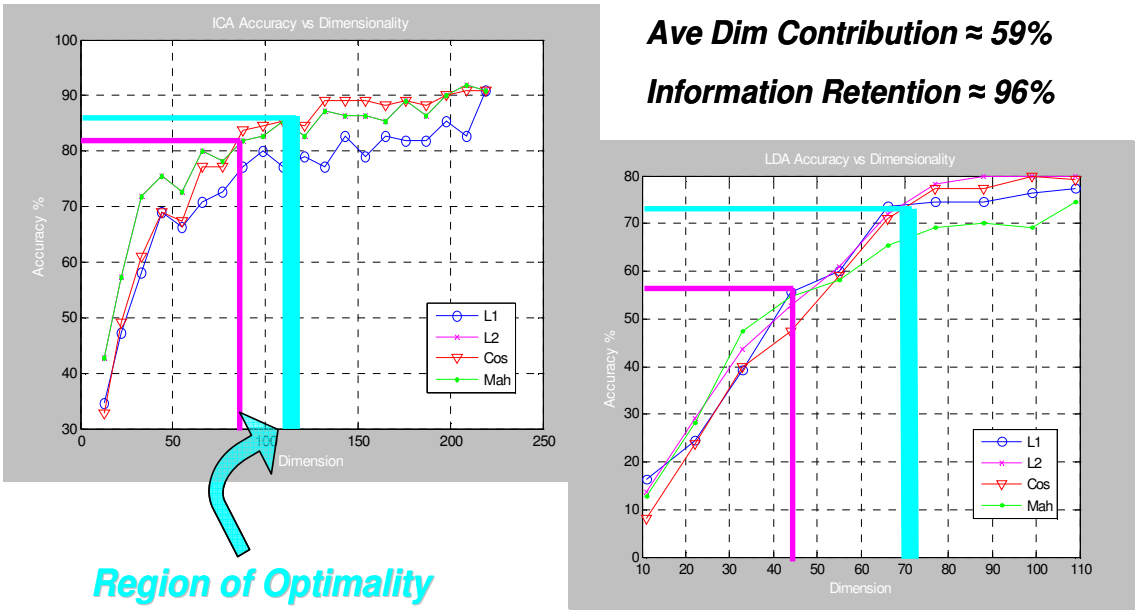



**Figure 11 Rank 1 Accuracy vs. Dimensionality**

Having established the optimum dimensionality for each subspace class (Linear and Kernel) i.e. $D$ = 116 for variants of PCA and ICA and $D$ = 72 for LDA algorithms, one could now formulate the projection subspaces for each algorithm. The top five component 'faces' for each algorithm are shown in figure 13 – 15.

### 3.6.2 Testing

As was done with the training images, the testing images are also arrayed into a column-wise matrix, where the mean training image is subtracted from each column. Using the transformation matrix obtained during training, the zero-meaned probe matrix ($P$) is projected into the subspace for each respective algorithm. The resulting basis vectors are then compared, using the nearest neighbour

approach, to the centroid vectors of class. Each probe image is finally classified with a class label, the results of which are put forth to the evaluation phase.

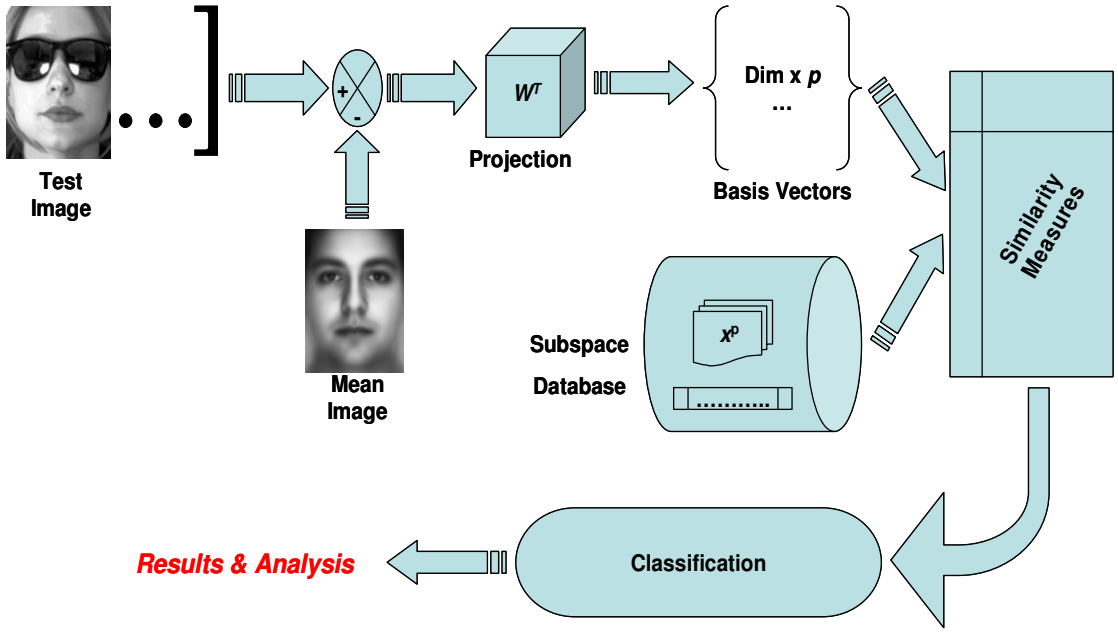



**Figure 12 Generic Testing Model**



**Figure 13 Eigenfaces**

**Rows 1 – 3 are Eigenfaces of Linear PCA, KPCA (rbf) and KPCA (poly) respectively.**

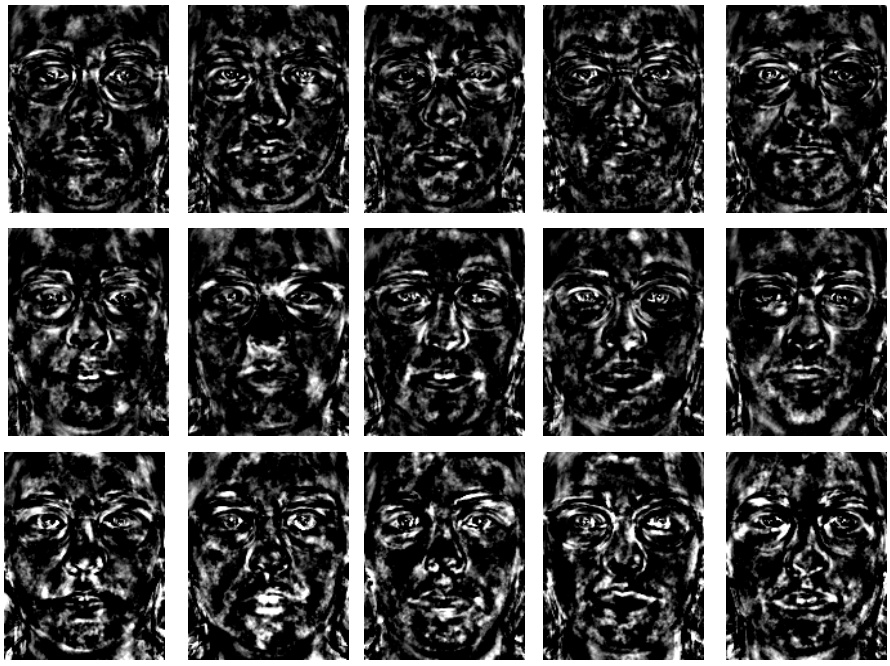


**Figure 14 Fisherfaces**
**Rows 1 – 3 are Fisherfaces of Linear LDA, KDA(rbf) and KDA (poly) respectively.**



**Figure 15 ICA faces**
Rows 1 – 4 are ICA faces of InfoMax *Arch I*, InfoMax *Arch II*, FastICA *Arch I* and FastICA *Arch II* respectively.

### 3.6.3 Similarity Measures

According to Duda [62], pattern classification is a one-to-one class assignment belonging to a set of pre-defined element classes (training images). This assignment will be based on the cost-effective, powerful non-parametric concept of *nearest neighbour classifier*, which in turn is based on the minimisation of a metric or distance measure between testing data (probe images) and each class centroid [33].

Distance metrics evaluate the difference between two vectors, $p$ (probe image) and $c$ (target image/centroid). There are many distance measures that are commonly used in engineering practice, with some metrics known to outperform others depending of course on the task, the data and the complementing algorithm being used. Comparative evaluation is based on the selection of four widely accepted distance metrics, namely *City Block* ($L1$), *Euclidean* ($L2$), *Cosine* (Cos) and *Mahalanobis* (Mah). These measures are defined as follows:

1. *City Block*

$$L1(p,c) = \sum_{i=1}^{P} |p_i - c_i| \tag{33}$$

2. *Euclidean*

$$L2(p,c) = \sqrt{\sum_{i=1}^{P} (p_i - c_i)^2} \tag{34}$$

3. *Cosine*

$$Cos(p,c) = 1 - \frac{P^T C}{|P| \cdot |C|} \tag{35}$$

4. *Mahalanobis*

$$Mah(p,c) = (p-c)^T \Sigma^{-1} (p-c) \tag{36}$$

where $\Sigma$ is the data covariance matrix defined by Bishop [63] as:

$$\Sigma = E\left[(p-c)(p-c)^T\right] \tag{37}$$

### 3.6.4 Evaluation Tools

The comparative assessment will be evaluated based on the combinatorial results of three successive tools. Firstly the binomial cumulative probability of correct class assignment will be presented in traditional tabular format. This will be followed by the FERET testing protocol using *Cumulative Match Scores* (CMS) curves also known as rank score [33]. The notion of rank score is based on the probability of guessing the data assignment correctly, such that a rank score equal to one (rank-1 score) is the probability of making the correct class assignment and accordingly a rank-2 score corresponds to the probability of making a correct class assignment in the top two, most likely, decisions. A rank-*N* score then, of course, refers to the probability of correct assignment within the top *N* candidates. Extending this concept, a rank-*c* score defined on *c* classes of data is equal to 1, as all potential classes are included as candidates. Additionally, a classifier with a 100% classification accuracy has a rank-1 score equal to 1, therefore, the rank score value, expressed as a function of the rank or class index, is a monotonically cumulative function increasing from 0 to 1 [63]. It is common to observe that an algorithm with superior rank-1 results may not yield the best performance at higher ranks i.e. due to various influencing factors when an image is taken, the classification solution may not always lie within the top one result but instead lie within the top *N* results, in other words, among how many nearest classes is the correct one. The CMS protocol thus answers the question of which algorithm offers better performance throughout the rank spectrum. CMS curves, tables and bar charts displaying percentages were traditionally the only means of representing comparative results, whereby any conclusions thus far were based solely on intuitive analysis i.e. which ever curve appeared higher on the CMS plot was the 'better' algorithm. Clearly a difference in performance of 1% or 2% could be attributed to pure chance, while 10% or more is probably not.

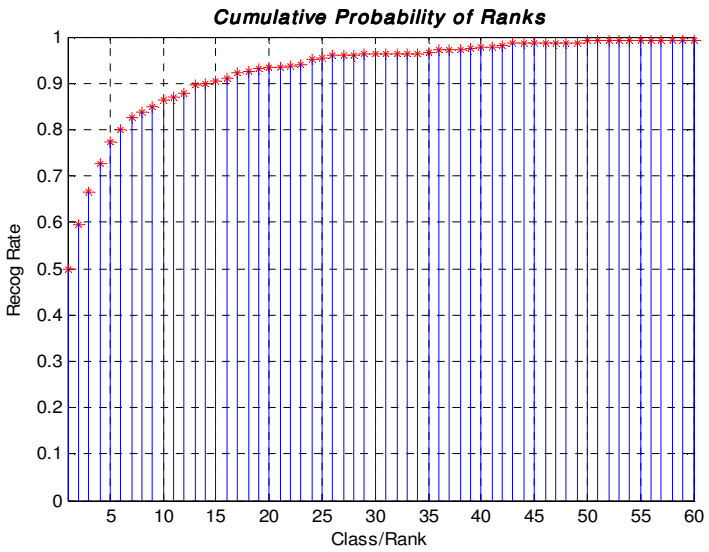



**Figure 16 Rank Score Example**

It was only when Yambor *et al* [64] mentioned the need to employ deeper analysis tools, did any further protocol develop. Up until that stage, as Givens *et al* [65] pointed out, it was common for large amounts of performance data to be collected and then not apply any statistical tools in determining the comparative performance levels of different algorithms, this was surprising since the only cost demand was computational time [66].

Beveridge *et al* [24] expanded the original FERET protocol by introducing statistics such as Recognition Rate Distributions and Confidence Intervals (CI) into the CMS curves. They also introduced hypothesis testing such as the $z$ –test [66], which was later replaced by more superior tests that were used in comparing the PCA and ICA face recognition algorithms in [12], [20] and [66]. One of the more precise and particularly well suited hypothesis techniques is defined by *McNemar's Test* [66].

McNemar's testing protocol is specially suited for paired data such as ones in face recognition i.e. data generated in testing two algorithms on the same probe set of images. McNemar's offers statistical insight pertaining to *what point does the difference in performance results actually become significant.*

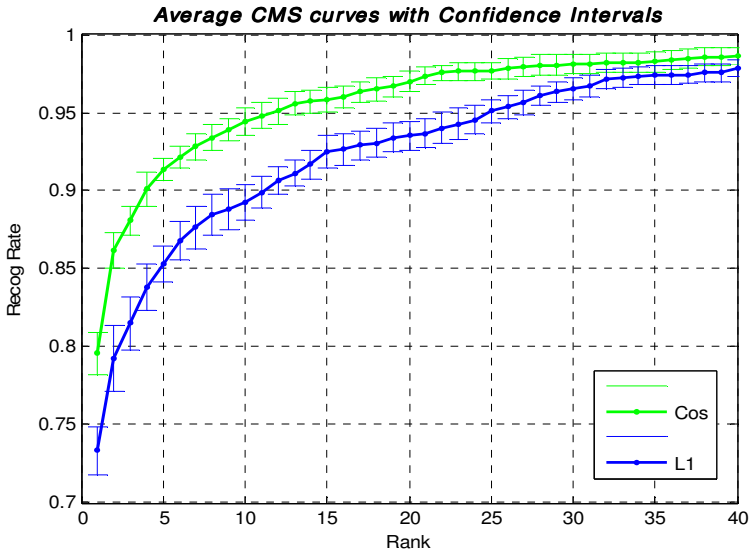



**Figure 17 Example of CMS curves with 95% CI over ten tests**

McNemar's Hypotheses are formulated as follows:

**Alternative Hypothesis (H1)** – When Algorithm 1 and 2 differ on a particular image, Algorithm 1 is more likely to correctly classify it.

**Null Hypothesis (H0)** – When Algorithm 1 and 2 differ on a particular image, none is more likely to correctly classify it.

The standard testing protocol is expanded so that the number of successes obtained by both algorithms, for every probe image, is recorded. This is accomplished by introducing a new variable (of the same size as the number of probe images $p$) which will contain a 1 in $p^{th}$ place if the algorithm correctly classifies the $p^{th}$ image and a 0 if not. Thus four possible outcomes will hold true for a given $p^{th}$ test image:

*SS* – Both algorithms successfully classify it

*SF* – Algorithm 1 is successful and Algorithm 2 is not

*FS* – Algorithm 2 is successful and Algorithm 1 is not

*FF* – Both algorithms fail to classify the image

For a given probe set, the number of times each outcome is observed is recorded. McNemar's first discards the outcomes in which algorithms behave the same i.e. *SS* and *FF*. Next, let $n_{SF}$ denote the number of times *SF* is observed and $n_{FS}$ the number of times *FS* is observed. The probability of the null hypothesis ($p$(**H0**)) is then formulated as such:

$$p(H0) \leq \sum_{i=0}^{n_{FS}} \frac{n!}{i!(n-i)!} \cdot \left(\frac{1}{2}\right)^n \tag{38}$$

where $n = n_{SF} + n_{FS}$. This is the probability or $p$-value for rejecting **H0** in favour of **H1**. The standard $p$-value of 0.05 (5%) cut-off is used in to reject **H0**. Most researchers have limited this test to rank-1 results until Delac *et al* proposed its extension across all points in the CMS curve [66]. It proved to be a good measure of an algorithm's performance across all ranks, thus giving one a sound statistical means of comparing two CMS curves.

In keeping the FERET CMS curve as the main tool for comparisons, this investigation will additionally employ McNemar's hypotheses testing in an evaluating when the difference between two curves becomes significant and thus allowing one to confidently recommend when and under what circumstances is the performance of one algorithm in fact superior to another.

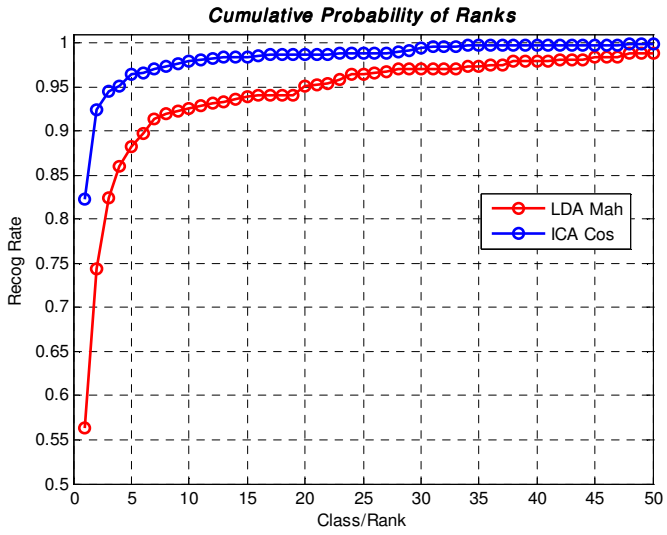



**Figure 18 CMS for two example algorithms**

Using an example to illustrate, figure 18 above shows two CMS curves, LDA with the Mahalanobis measure and ICA *Arch II* with the Cosine distance metric; both curves represent the different algorithmic results for a given probe set. Just by visibly inspecting the CMS plots one could conclude that the ICA algorithm is always superior. This is a very straight forward conclusion and exactly the kind that was used in many previous comparisons.

However applying McNemar's test, shows us this conclusion is in fact incorrect. Recording *SF* and *FS* outcomes for each probe image at each rank, $p(\mathbf{H0})$ for the first fifty ranks are shown in figure 19, where it can be seen that the probability of rejecting **H0** in favour of **H1** rapidly declines after rank 8 and is virtually zero from rank 25 onward.

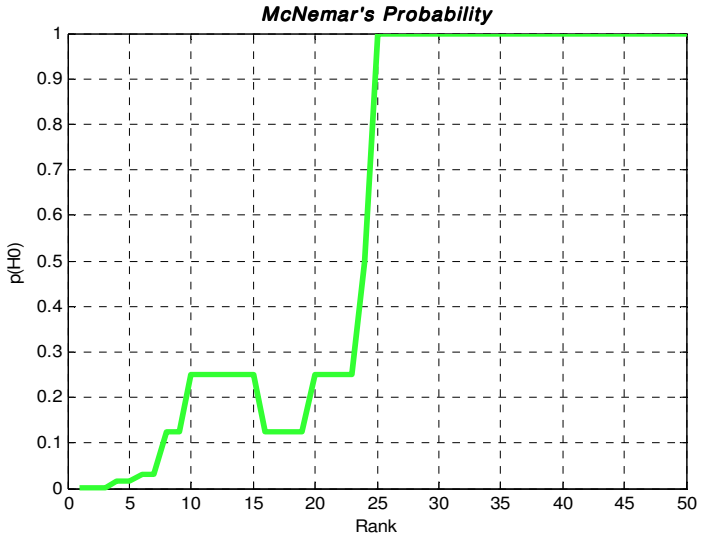



**Figure 19 An example plot of *p* values for McNemar's Test**

In fact we can only state, with more than 95% confidence, that the ICA algorithm is significantly superior to LDA only up to rank 7 and not across all the ranks as previously stated. This is depicted in the decision graph of figure 20, where the value -1 means Algorithm 1 is significantly better, the value 1 means Algorithm 2 is better and 0 means there is no significant difference between the two.

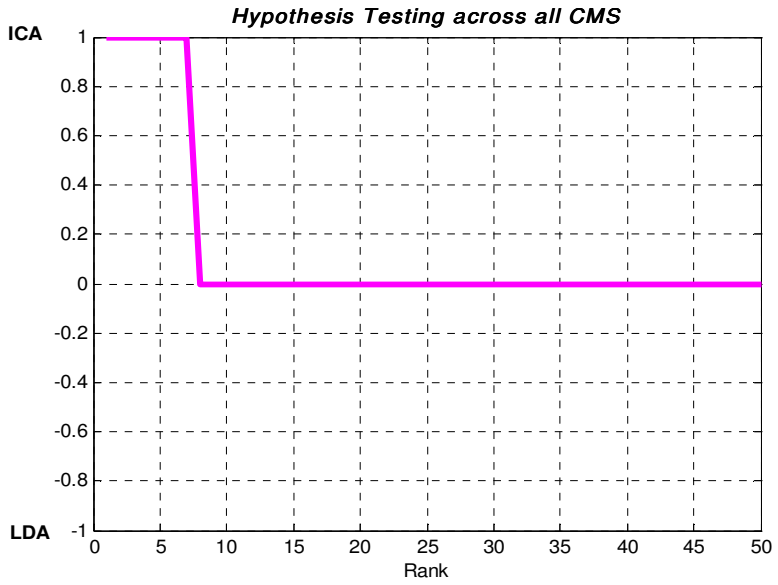



**Figure 20 Decision graph for McNemar's Test**

McNemar's test offers important statistical insights, which may considerably influence any previous and future comparative conclusions. In conjunction with the rank-1 results and the CMS curves, McNemar's will be used to obtain decisive clarity in evaluating the 'best' algorithms in each testing category.

### 3.6.5 Selection of Subspace Algorithms

As mentioned earlier, each class of subspace algorithms, PCA, LDA and ICA, will be comparatively assessed using a common evaluation platform. With regard to the classes of PCA and LDA, both linear and non-linear variants are investigated. Bearing in mind the still, significant image space, $N = 20\,625$, one would like to boost computational efficiency as effectively as possible. Upon initial investigation, in terms of computational resource and time demand, the classical algorithms of PCA and LDA were found to handle this image space very poorly. Further research revealed works by Baudat [53] and Franc [68], which proved that Kernel implementations employing linear functions (1$^{st}$ degree

polynomials) can efficiently handle larger input dimensionality and are equivalent to the classical implementations of PCA and LDA. It was therefore decided, in an effort to significantly reduce the computational load, that the kernel implementations of KPCA and KDA (GDA) will be used to compare both the linear and non-linear projections variants.

Among the non-linear kernel functions, there are many types, the most popular being the Gaussian (RBF) kernel, the Polynomial kernel and the Sigmoidal function. All these functions were investigated and it was found that the Gaussian and second degree Polynomial kernels, empirically achieved the lowest error rates. The final results for the classes of PCA and LDA will hence be presented in three variants: linear, RBF and $2^{nd}$ degree polynomial. Looking at the class of ICA, it has already been mentioned that the Kernel implementation is not feasible for large input spaces; the results will therefore be presented demonstrating the algorithms of InfoMax and FastICA in combination with both Architectures I and II.

# Chapter 4

# Results and Analysis

## 4.1 Introduction

The proceeding results are presented in subsections corresponding to the evaluation categories of Expression, Illumination, Lower Facial Occlusion, Upper Facial Occlusion and Time Delay. In each subsection the results, for every similarity measure averaged over 10 trials, will be presented in two ways:

- Tabulated rank-1 accuracies
- CMS curves across the first 60 ranks

These results will then be evaluated using the CMS curves and McNemar's Testing Protocol, where the best intra-class results will initially be compared in order to determine the 'best' algorithm variant and associated distance metric for each subspace class. The resulting 'best' algorithms will then be compared, inter-class, against each other to reveal the algorithm that offers most superior performance in each respective evaluation category. The final results will then be assessed in the context of the task nature and previous literary findings, thereby yielding meanings comparative insights that will be taken forward to the Hybrid formulation, proposed in Section 5. (Additional support material of individual

CMS curves with Stem Plots and Confidence Interval measures can be found in *Appendix A* contained within the attached CD)

## 4.2 Expression

Using the described methodology, the *Expression* probe set was evaluated first. One interesting observation when answering the question of which algorithm performs better, evident in all the categories, is the discrepancy in some cases between the best rank-1 results and the best CMS results. This brings into question any analyses that are done employing only the CMS curves of those projection metrics that yielded the best rank-1 performance. The best rank-1 performance may not always produce the best results at higher ranks and it is for this reason that each result table presented, will be divided into two sections; the left potion will provide the rank-1 metric accuracies and the right-hand side will show the best CMS metric curve pertaining to each projection method. Only the best CMS projection-metric combinations will be selected for the intra-class and then the inter-class comparisons.

All the CMS curves associated with the expression probe set are shown in figures 21 – 23 below. In the classes of LDA and ICA, there were no discrepancies between the rank-1 and CMS results. However, with every variant of the PCA class, *Linear, RBF* and *Polynomial*, the best rank-1 results produced by the L1 similarity measure did not produce superior results throughout the rank spectrum; this was superseded by the Euclidean (L2) measure and subsequent Cosine performance. Any doubts in metric superiority were put to rest in further analysis by subjecting the metrics in question to McNemar's Testing. The selected projection-metric combinations, shown under the *Highest Curve* section in each table, are therefore both intuitively and statistically superior.

Having established the best projection-metric combinations for each algorithm, this investigation will serve to provide insight into two aspects:

1) The best categorical projection-metric combination *within* each class of subspace projection (Intra-class) and
2) The best categorical projection-metric combination *between* each class of subspace projection (Inter-class)

In addressing the first objective, the top selections are evaluated against each other to reveal the best intra-class projection-metric combination. This comparison is again done using McNemar's protocol. The graphical results and $p(H0)$ values can be found in *Appendix A*.

| | Rank 1 Expression Results | | | | CMS Results | |
|---|---|---|---|---|---|---|
| | ***L1*** | ***L2*** | ***Cos*** | ***Mah*** | ***Highest Curve*** | ***Same as rank-1*** |
| ***PCA*** | | | | | | |
| *Linear* | ***79.41%*** | 77.18% | 76.73% | 69.82% | L2 | No |
| *RBF* | ***80.18%*** | 77.14% | 78.51% | 72.46% | L2 | No |
| *Poly* | ***79.73%*** | 77.32% | 77.55% | 77.46% | L2 | No |
| ***LDA*** | | | | | | |
| *Linear* | 76.05% | 76.05% | ***76.18%*** | 76.09% | Cos | Yes |
| *RBF* | 72.32% | 73.63% | 71.32% | ***83.43%*** | Mah | Yes |
| *Poly* | 82.18% | 81.77% | 81.96% | ***83.25%*** | Mah | Yes |
| ***ICA*** | | | | | | |
| *Info A1* | 71.27% | 70.36% | ***76.86%*** | 73.41% | Cos | Yes |
| *Info A2* | 69.18% | 70.77% | ***78.91%*** | 71.73% | Cos | Yes |
| *Fast A1* | 73.36% | 71.96% | ***79.59%*** | 71.96% | Cos | Yes |
| *Fast A2* | 70.14% | 72.46% | ***80.05%*** | 72.46% | Cos | Yes |

**Table 1 Rank 1 Metric Results for the Expression Probe Set**

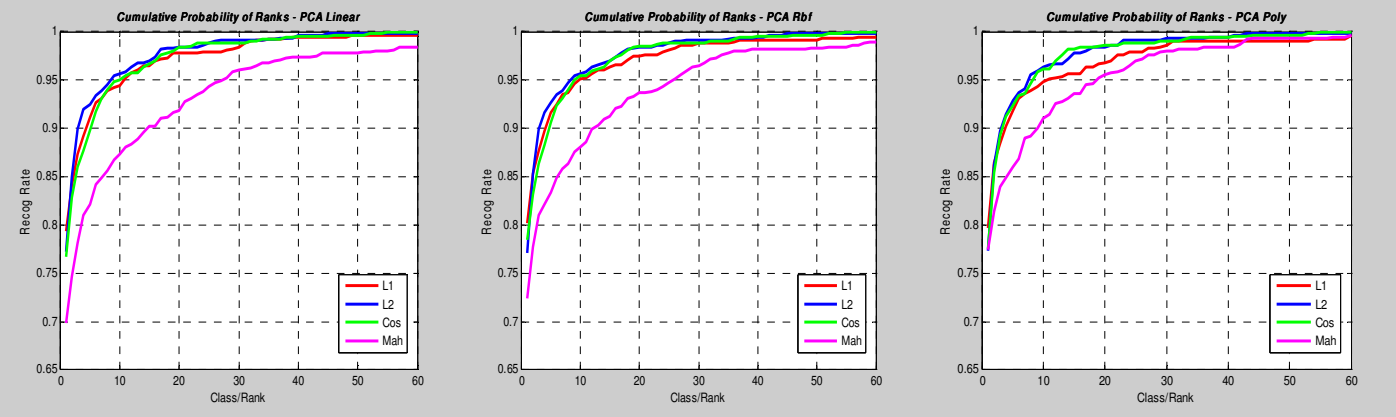



**Figure 21 CMS curves for all similarity measures – Expression: PCA**

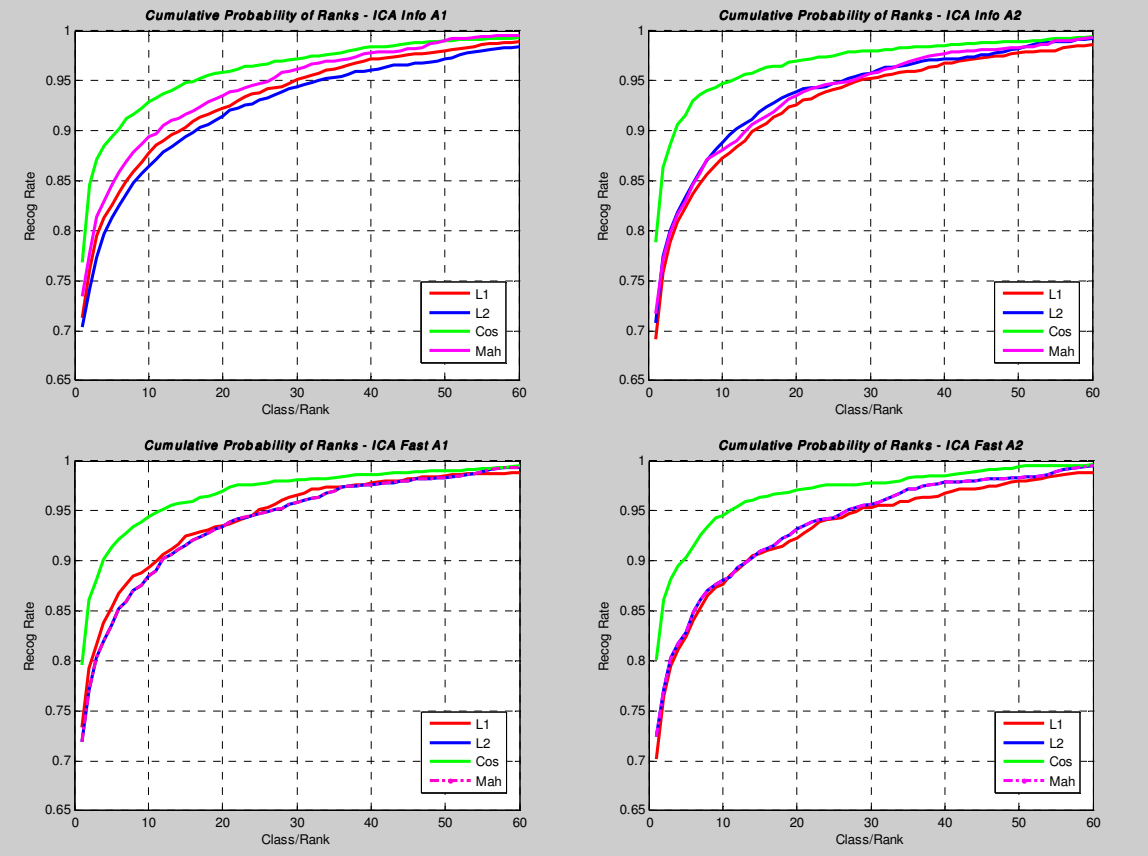



**Figure 22 CMS curves for all similarity measures – Expression: ICA**

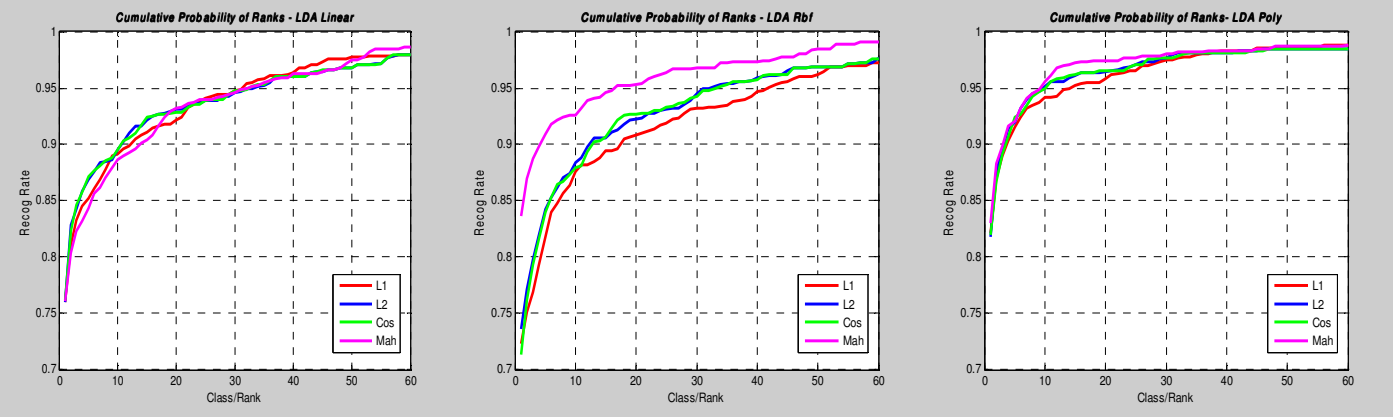



**Figure 23 CMS curves for all similarity measures – Expression: LDA**

Within the class of **PCA** the results were as follows:

*Linear vs RBF*: The CMS curves are virtually identical, except for a few ranks where the linear curve shows marginal superiority. This is supported by McNemar's probability showing a between 50% and 70% chance of better performance at rank 1 and 3. Of course, this is not sufficient as one requires at least 95% confidence in declaring superiority at any rank. Overall there is no significant statistical difference between the two algorithms.

*Linear vs Polynomial*: Statistically there exists no difference between the two algorithms. The only superiority is suggested by the Polynomial CMS curve that displays marginally better results for the first 25 ranks.

*RBF vs Polynomial*: Again there is no significant difference between these variants, except the slightly better performance of the CMS Polynomial curve over the first 30 ranks.

Statistically there exists no difference between any of these PCA variants. Bearing in mind the CMS curve is still the primary means of evaluation and that a selection needs to be made for the second objective to be realised, intuition would suggest the *Polynomial – L2* combination be selected as the 'best' algorithm for inter-class evaluation.

In the class of **LDA**:

*Linear vs RBF*: The CMS curves are clearly dominated by the RBF results, suggesting superiority throughout the spectrum. McNemar's however, reveals that this is only significantly true for the first 2 ranks and thereafter there exist no statistical difference.

*Linear vs Polynomial*: Similar to the RBF results, the Polynomial curve shows a better performance throughout CMS plot, but statistically superiority exists only for the first two ranks.

*RBF vs Polynomial*: Statistically there is absolutely no significant difference between these two variants. The only visible superiority lies in the marginally better performance of the Polynomial CMS curve across the first 40 ranks. Intuitively, the *Polynomial – Mah* algorithm will hence be selected as the 'best' performance measure

**ICA** variants:

*Info A1 vs Info A2*: Statistically the algorithms offer similar performance, however rank 1 McNemar's probability and CMS plots suggest the InfoMax Arch II algorithm to offer better performance.

*Info A1 vs Fast A1*: Only the CMS curves suggest Fast ICA Arch I superiority in early ranks, but McNemar's offers no significant difference between the two variants.

*Info A1 vs Fast A2*: The results of the FastICA Arch II algorithm mimic those found earlier in the InfoMax Arch II Algorithm.

*Info A2 vs Fast A1*: Statistically there is again no significant difference, however the CMS curves suggest the marginally better performance by the InfoMax Arch II variant.

*Info A2 vs Fast A2*: Both algorithms offer exactly the same statistical performance, the only differing factor is the small rank-1 performance margin of the FastICA implementation.

*Fast A1 vs Fast A2*: Again there is no significant difference between these approaches, except the CMS suggestion of Fast A2's marginal improved performance.

Statistically there is no difference between any of the ICA variants. Selection for the 'best' algorithm is, therefore, again left to user intuition. The CMS curves suggest that the best results are provided by Architecture II's implementation in both the InfoMax and FastICA algorithms. The FastICA variant offers a marginal rank-1 superiority and is computational less costly. It is for these reasons that *FastICA Arch II* will be selected for inter-class evaluation

Having selected the 'best' class projection-metric combinations, and noting that this selection was based mainly upon CMS intuition and not significant statistical difference, the inter-class evaluation can now be performed. The results of McNemar's evaluation are shown in figure 24.

Looking first at column 1 in figure 24, the classes of PCA and LDA are evaluated. The CMS curves reflect the clear 'superiority' of PCA in later ranks, which, given the small sample of training data, can be expected. McNemar's however, demonstrates that the only significant *difference* between the two algorithms lies *at rank-1*, where LDA offers greater performance. Considering that PCA superiority at later ranks is less than 2% and that LDA rank-1 performance is better by more 5%, one is undoubtedly inclined to select LDA (*Poly + Mah*) as the better performing algorithm.

Examining the second column, where the performance of PCA is assessed against the class of ICA. The CMS chart again suggests PCA as the better algorithm from rank 3 onward, however the results from McNemar's analysis nullifies this suggestion, proving that the only major *difference* in performance occurs *at rank-1*, where ICA (*FA2 + Cos*) is superior.

LDA and ICA were then pitted against each other and as can be expected, given their similar performance against PCA, there is *no statistical difference* between the two algorithms. As can be observed in column 3, the only difference lies in the CMS chart where LDA presumptuously offers 'better' recognition results for the first twenty ranks.

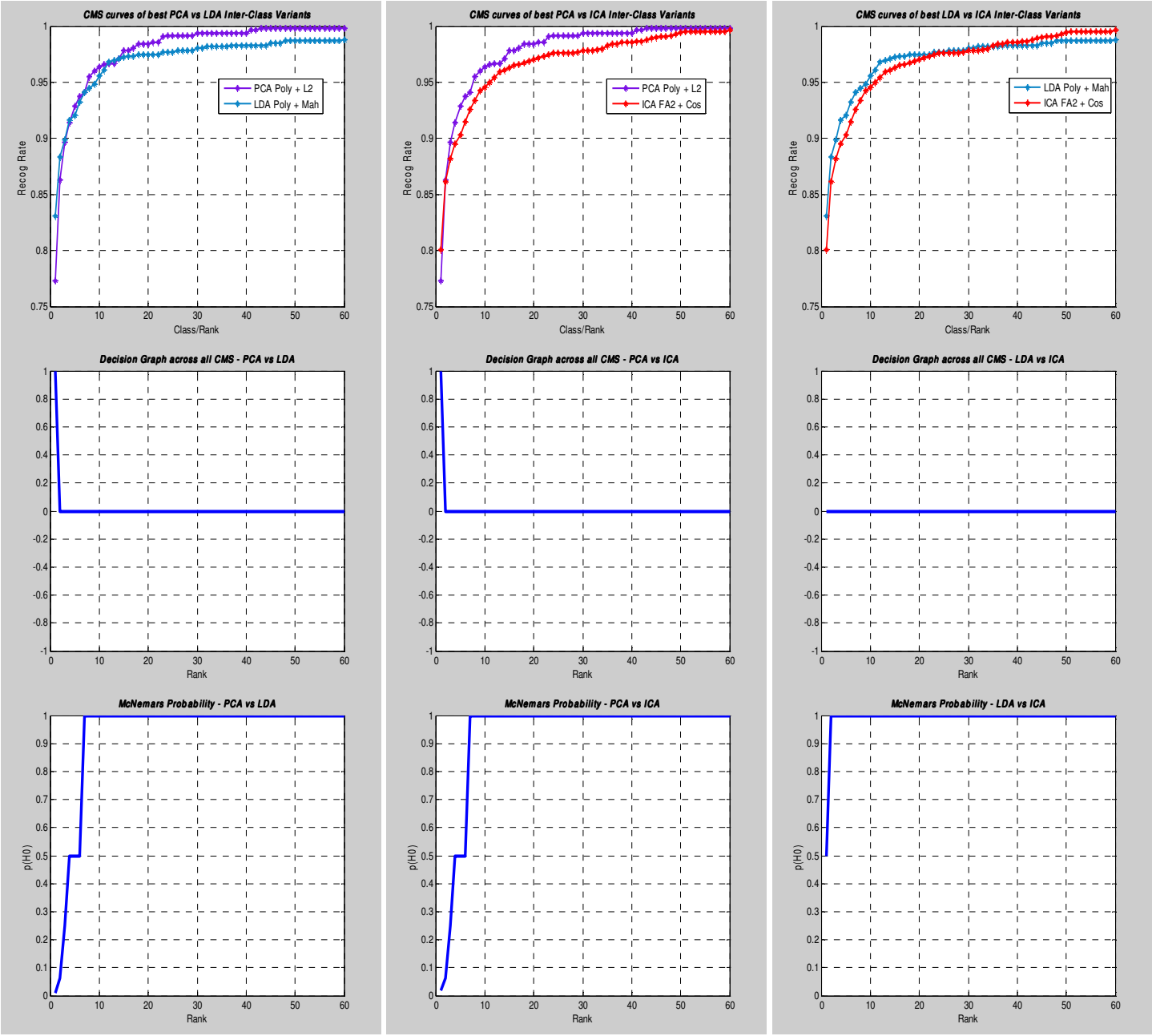



**Figure 24 McNemar's Test for the top 3 inter-class algorithms - Expression**
**Column 1: PCA Poly(L2) vs LDA Poly(Mah), Column 2: PCA Poly(L2) vs ICA FA2(Cos), Column 3: LDA Poly(Mah) vs ICA FA2(Cos)**

In summary, there exists only a small amount of statistical difference between any of the subspace classes. The intra-class evaluations suggest that one could use any variant of the PCA, ICA and LDA approaches, with the exception of the Linear LDA, and achieve similar classification results for expression probes. The inter-class evaluations then show that ICA and LDA offer only rank-1 superiority over PCA and hence one could employ either the ICA or LDA class and obtain similar performance levels.

A final point to consider is the similarity measures that combine best with each class for the task of evaluating expression changes. With PCA it was found that the City Block – *L1* metric offered the best rank-1 accuracies, and the Euclidean – *L2* measure offered better results across the entire rank spectrum. In the LDA class the Mahalanobis – *Mah* distance metric offered the most superior results across the spectrum and lastly the Cosine – *Cos* measure was found, without a doubt, to offer the best results in the class of ICA.

## 4.3 Illumination

Turning to the category of changes in *Illumination*, the rank-1 and CMS results are again presented in tabular form in table 2 below. Only one discrepancy case is observed, in the LDA – *Linear* algorithm, where the highest rank-1 result (L2) does not coincide with the best overall metric, given by the Mahalanobis CMS curve. All the CMS curves for each projection-metric combination associated with the illumination probe set are found in figures 25 – 27. And again the intra-class CMS plots, Decision Graphs and McNemar's Probability graphs can be found in *Appendix A*.

| Rank 1 Illumination Results | | | | | CMS Results | |
|---|---|---|---|---|---|---|
| | *L1* | *L2* | *Cos* | *Mah* | *Highest Curve* | *Same as rank-1* |
| ***PCA*** | | | | | | |
| *Linear* | ***55.82%*** | 53.46% | 51.91% | 49.82% | L1 | Yes |
| *RBF* | ***55.82%*** | 51.91% | 50.64% | 53.73% | L1 | Yes |
| *Poly* | ***57.09%*** | 52.86% | 52.46% | 55.52% | L1 | Yes |
| ***LDA*** | | | | | | |
| *Linear* | 69.73% | **71.91%** | 70.00% | 66.73% | Mah | No |
| *RBF* | 45.46% | 48.73% | 45.46% | ***56.27%*** | Mah | Yes |
| *Poly* | 71.18% | 72.50% | 70.86% | ***75.18%*** | Mah | Yes |
| ***ICA*** | | | | | | |
| *Info A1* | 65.23% | 68.91% | ***83.36%*** | 75.00% | Cos | Yes |
| *Info A2* | 68.27% | 72.18% | ***83.09%*** | 73.96% | Cos | Yes |
| *Fast A1* | 71.46% | 74.59% | ***84.14%*** | 74.59% | Cos | Yes |
| *Fast A2* | 66.09% | 71.59% | ***82.23%*** | 71.59% | Cos | Yes |

**Table 2 Rank 1 Metric Results for the Illumination Probe Set**

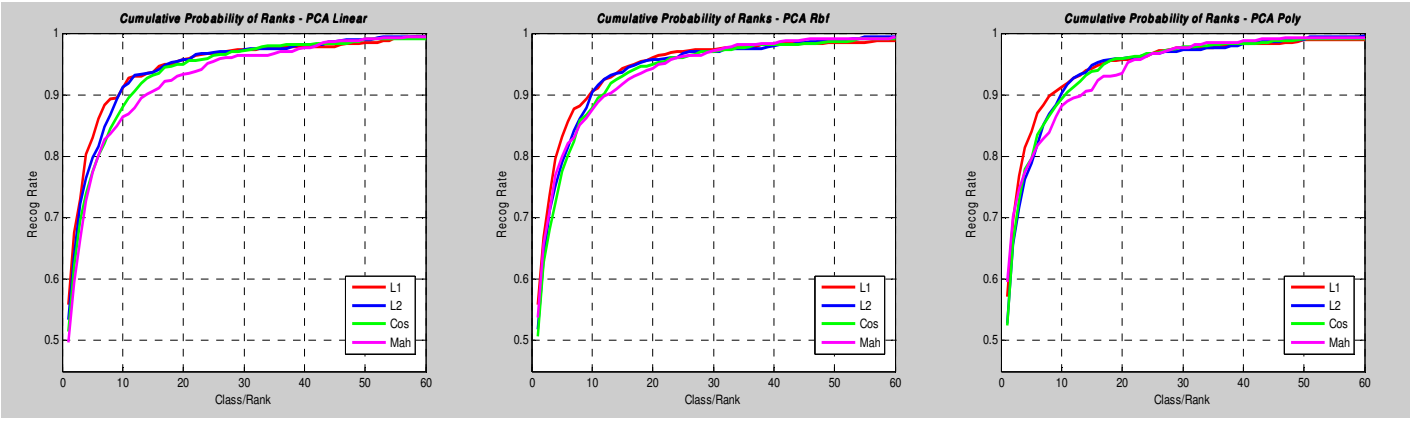



**Figure 25 CMS curves for all similarity measures – Illumination: PCA**

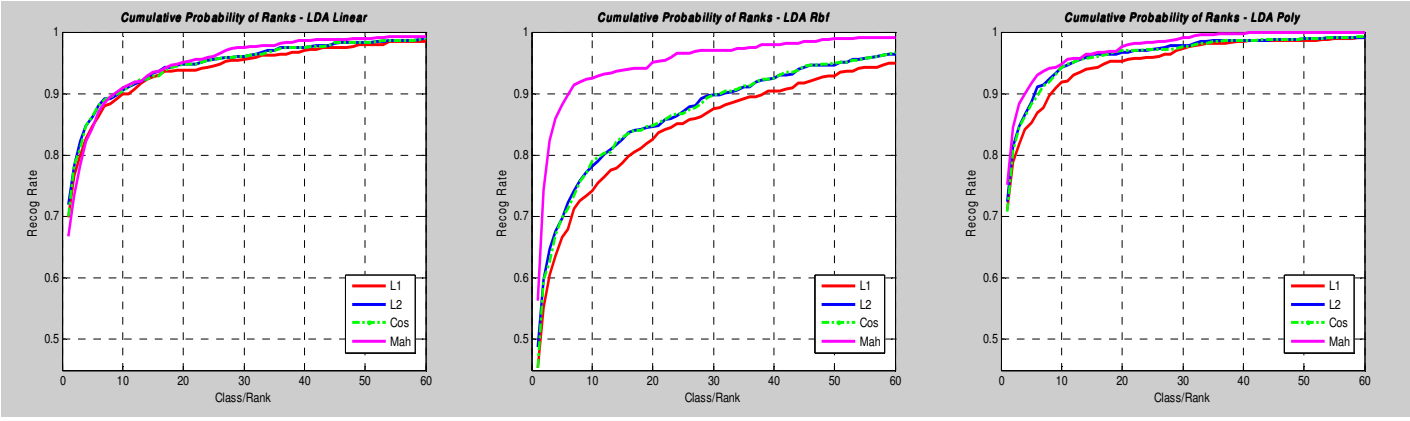



**Figure 26 CMS curves for all similarity measures – Illumination: LDA**

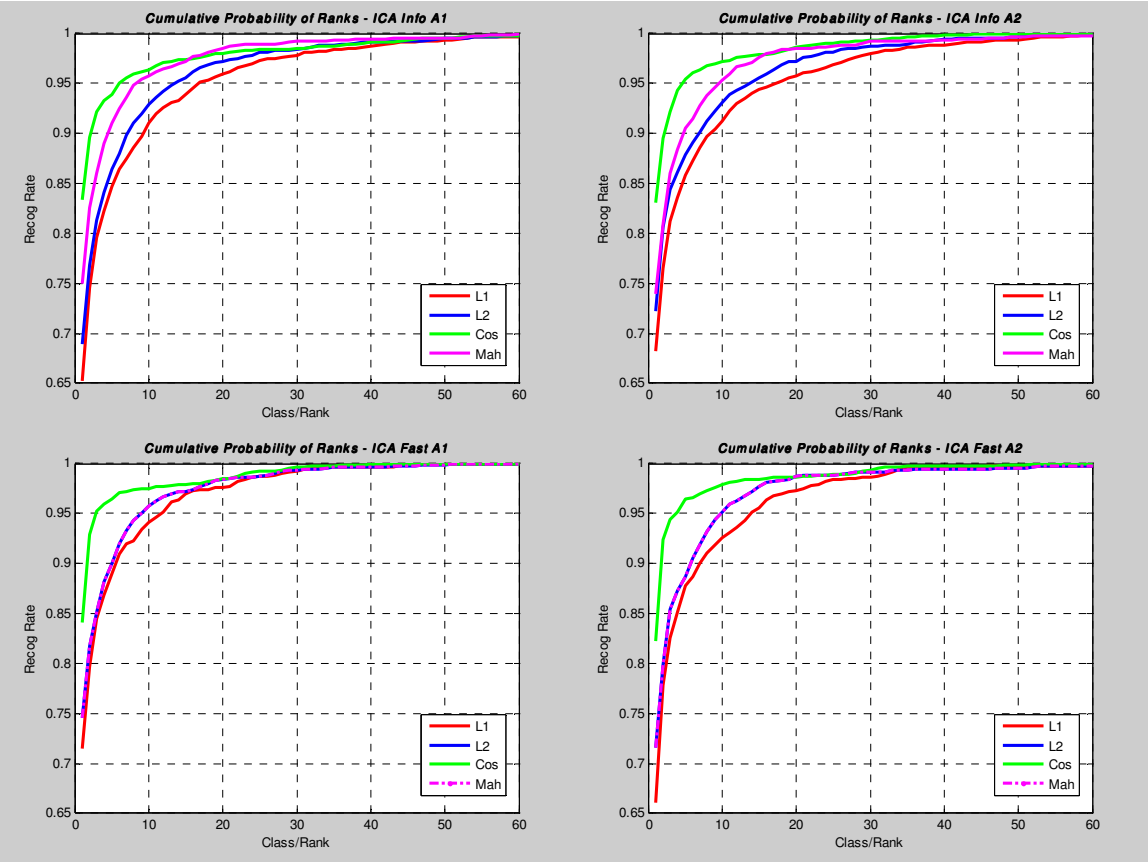



**Figure 27 CMS curves for all similarity measures – Illumination: ICA**

Intra-class testing yielded the following:

**PCA**

*Linear vs RBF*: Statistically there is no sound difference between the two algorithms. The *p*(H0) values fluctuate somewhat during the first twenty ranks, suggesting some confidence in the linear approach in early ranks. This is confirmed by the CMS chart, where the linear algorithm edges marginally forward in certain early ranks.

*Linear vs Polynomial*: The *polynomial approach* offers a distinctively higher performance in the first two ranks, thereafter offers no statistical advantage over its linear counterpart.

*RBF vs Polynomial*: As one would expect, the *polynomial approach* is statistically superior over the first two ranks. Intuition based on the lower *p* values and marginally higher CMS curve for early ranks would suggest performance order in the form of Polynomial, followed by Linear and lastly the RBF variant.

**LDA**

*Linear vs RBF*: An interesting comparison that revealed rank-1 superiority belonged to the linear algorithm and in rank-2 and 3 the *RBF variant* was found to be statistically superior. Following the CMS curves, the RBF curve outperformed the linear curve by approximately 2% in each of the first fifteen ranks. This result again highlights the fact that rank-1 superiority may not yield the better performing algorithm.

*Linear vs Polynomial*: The *polynomial approach* is by far the more statistically superior curve, offering significantly higher performance from rank 1-7 and again from rank 10-13 and over 90% confidence in hypothesis superiority for the first 20 ranks. This performance is almost identically echoed in the CMS chart.

*RBF vs Polynomial*: The *polynomial approach* again offers higher levels of performance, however its superiority is statistically only limited to the first two ranks, with a 90% confidence for the first five ranks.

**ICA**

*IA1 vs IA2*: There exists no statistical difference between these two ICA variants, the CMS plot does however, hint toward InfoMax Arch II's marginal superiority over majority of the spectrum.

*IA1 vs FA1*: McNemar's suggests no statistical difference, however the CMS chart and over 75% confidence *p* values does intuitively suggest FastICA Arch I's superiority in early ranks.

*IA1 vs FA2*: Again there is no significant statistical difference in the results, but similar to *FA1*'s performance, FastICA Arch II does hint at better performance levels during early ranks.

*IA2 vs FA1*: Once again the statistical performance of the two algorithms is the same. Only intuitive analysis support *FA1*'s superiority in the first two ranks, this is evident in both the CMS plot and the almost 90% confidence level found by McNemar's probability.

*IA2 vs FA2*: There is no strong performance difference between these two. *FA2*'s performance is however, as mentioned earlier similar to *FA1* above.

*FA1 vs FA2*: As expected, there is virtually no difference between these subspace implementations, except for *FA1*'s marginally better performance, evident in the early ranks of the CMS curve.

Although *no statistical difference exists between any of the ICA variants* and all of them would yield similar performance levels, the FastICA Arch I algorithm is intuitively chosen for representing the ICA class in the inter-class assessments. Figure 28 displays the results for the inter-class McNemar evaluations. Considering the rank-1 results, there are large differences ( > 10%) in the performance accuracies of each class and one would could expect the inter-class result to be the same i.e. ICA being top performer, followed by LDA and lastly PCA.

As was done in the previous section, the classes of PCA and LDA were evaluated first, the results of which are shown in column 1. The CMS chart would suggest that the LDA algorithm is without a doubt the better projection method throughout the rank spectrum. McNemar's test offers some support to this suggestion, but bringing the clarity that distinguishes at exactly what point is LDA superiority in

fact significant. The Decision Graph demonstrates that LDA is clearly superior for the first 6-7 ranks only. A further observation is that the *p*-value chart reveals that probability of H0 existing is only 7% for ranks 9-13. One could therefore state with good confidence that LDA (*Poly* + *Mah*) offers distinctively higher performance levels than PCA for the first 13 ranks.

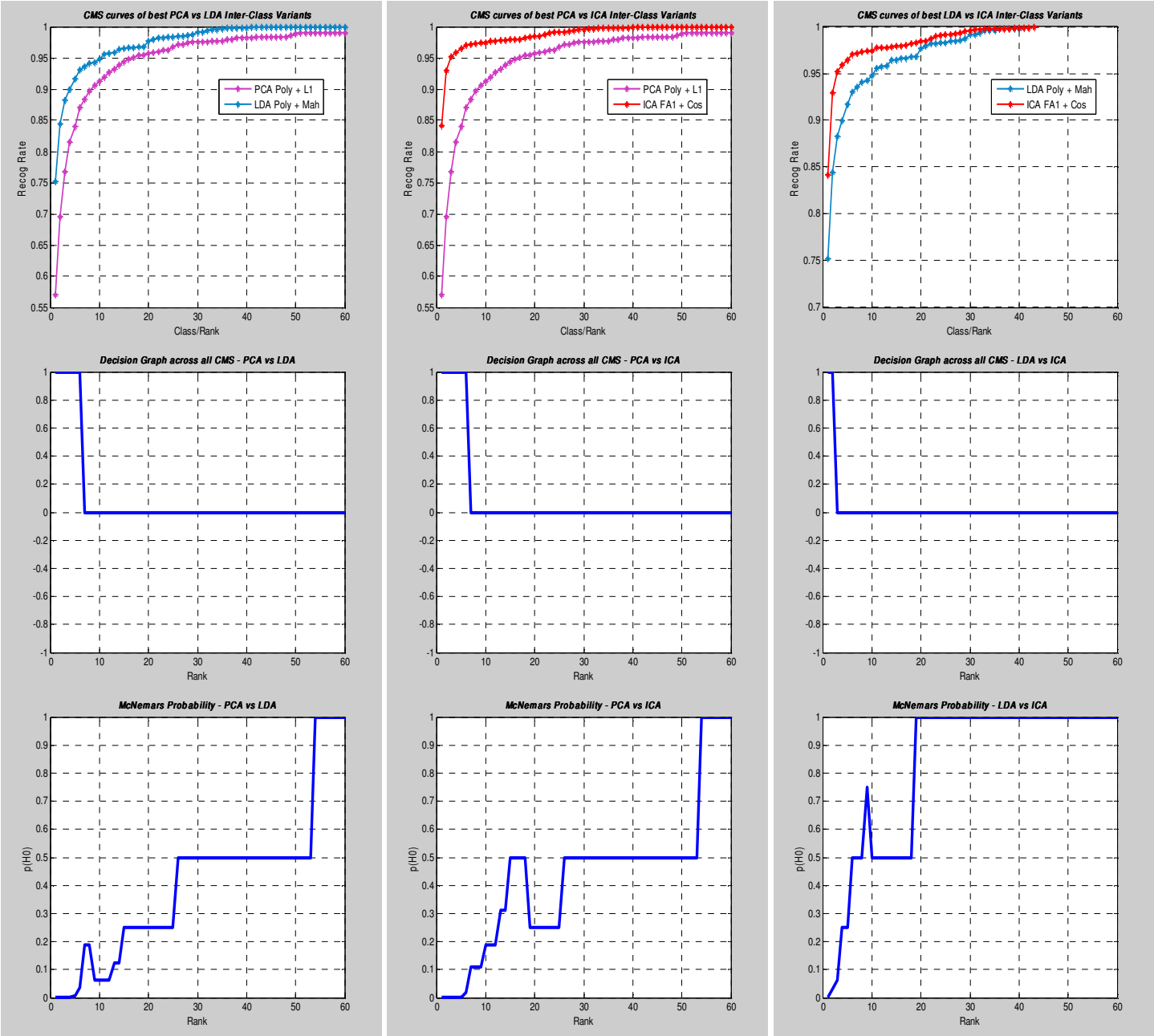



**Figure 28 McNemar's Test for the top 3 inter-class algorithms - Illumination**
Column 1: PCA Poly(L1) vs LDA Poly(Mah), Column 2: PCA Poly(L1) vs ICA FA1(Cos), Column 3: LDA Poly(Mah) vs ICA FA1(Cos)

Moving to column 2, where the classes of PCA and ICA are examined. As expected ICA (*FA1 + Cos*) is clearly superior, but as was the case with LDA, this superiority is exaggerated in the CMS chart and only holds statistically true for the first 6-7 ranks, with a 90% confidence probability for the first ten ranks.

Lastly the classes of LDA and ICA are evaluated in column 3. As intuition would suggest ICA is the better performing algorithm. Its superiority however, is significantly defined in the first 2-3 ranks only and its performance is somewhat distinguishable only until rank-8

In summary, the intra-class evaluations have revealed that in the category of illumination the polynomial algorithms in combination with the City-Block ($L1$) measure for PCA and the Mahalanobis (*Mah*) metric for LDA produced the best results in early ranks. Pertaining to the class of ICA, similarly to the case in the preceding section, there was again no statistical difference between any of the ICA variants and the choice of 'best' algorithm was left to user-intuition.

Comparing the inter-class assessments, PCA's low accuracy levels suggests that the class does not cope well with variations in lighting and illumination changes which are better handled by the classes of LDA and ICA. ICA by far offered the best performing projection-metric combination, albeit only for early ranks, but its recognition levels were still far superior to both the PCA and LDA algorithms.

## 4.4 Lower Occlusion

The category of occlusion, particularly lower, is perhaps the hardest test for any recognition algorithm. With possibly large parts of the face being obscured and no implicit availability of model simulation or 3D registration techniques, it would be realistic to expect a possibly poor classification performance

from any appearance-based subspace methodology. This is clearly evident in the low accuracy levels, shown in the table below, achieved by all the subspace classes.

The PCA class achieved a rank-1 error rate greater than 92%, which may lead one to consider that the few correct classifications could be attributed to nothing more than database coincidence. And surprisingly the Mahalanobis metric, which is rarely recommended for use with PCA, produced the best performance across the rank spectrum for all the PCA variants.

The LDA class, with the exception of the linear variant, produced very similar results to PCA. Performance discrepancies between the rank-1 and CMS results are again noticed with the RBF and Polynomial variants, where, as can be seen in the CMS charts in figure 31, the Mahalanobis measure clearly is the superior metric throughout the ranks.

ICA offers, perhaps, the most promising performance with the Cosine metric, achieving results that are more than five times those produced by PCA and LDA.

| | Rank 1 Lower Occlusion Results | | | | CMS Results | |
|---|---|---|---|---|---|---|
| | *L1* | *L2* | *Cos* | *Mah* | *Highest Curve* | *Same as rank-1* |
| ***PCA*** | | | | | | |
| *Linear* | 2.91% | 3.96% | 3.77% | ***7.23%*** | Mah | Yes |
| *RBF* | 3.36% | 4.14% | 3.82% | ***5.68%*** | Mah | Yes |
| *Poly* | 3.50% | 3.96% | ***4.14%*** | ***4.14%*** | Mah | Yes |
| ***LDA*** | | | | | | |
| *Linear* | 14.96% | 14.96% | ***16.46%*** | 9.77% | Cos | Yes |
| *RBF* | 3.46% | 3.36% | ***4.46%*** | 4.36% | Mah | No |
| *Poly* | ***6.14%*** | 4.59% | 3.46% | 5.55% | Mah | No |
| ***ICA*** | | | | | | |
| *Info A1* | 12.18% | 12.46% | ***28.05%*** | 13.91% | Cos | Yes |
| *Info A2* | 10.55% | 11.00% | ***22.36%*** | 10.55% | Cos | Yes |
| *Fast A1* | 9.50% | 10.18% | ***27.86%*** | 10.18% | Cos | Yes |
| *Fast A2* | 11.55% | 12.46% | ***28.27%*** | 12.46% | Cos | Yes |

**Table 3 Rank 1 Metric Results for the Lower Occlusion Probe Set**

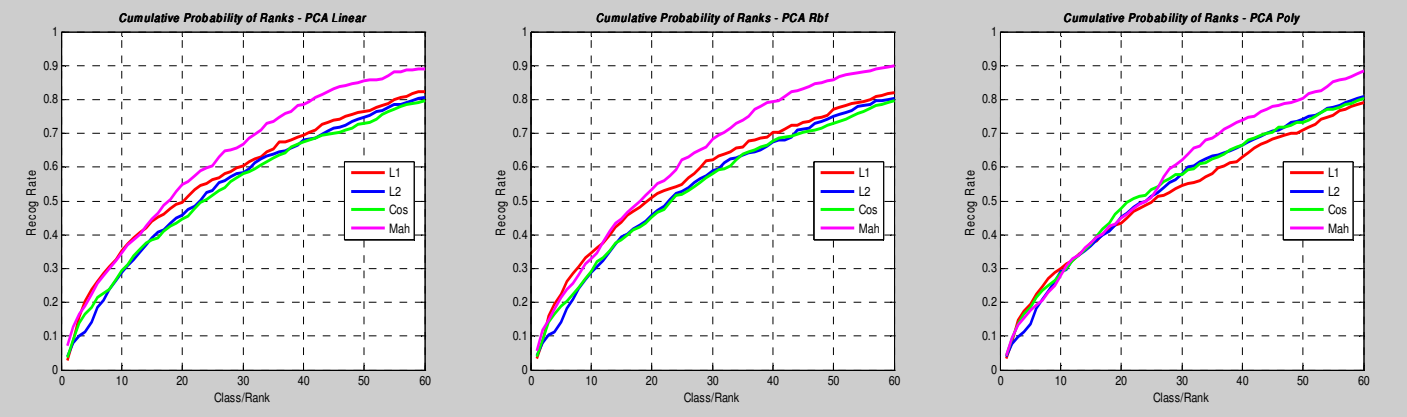



**Figure 29 CMS curves for all similarity measures – Lower Occlusion: PCA**

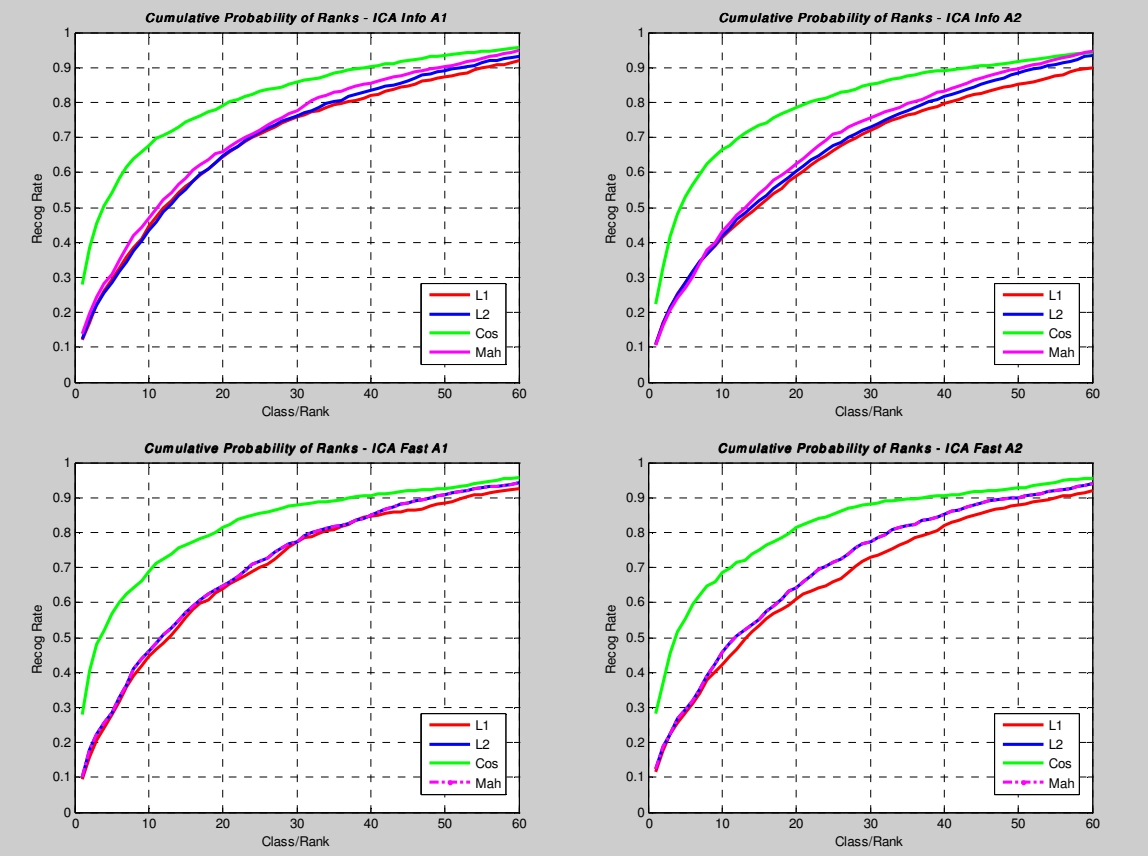



**Figure 30 CMS curves for all similarity measures – Lower Occlusion: ICA**

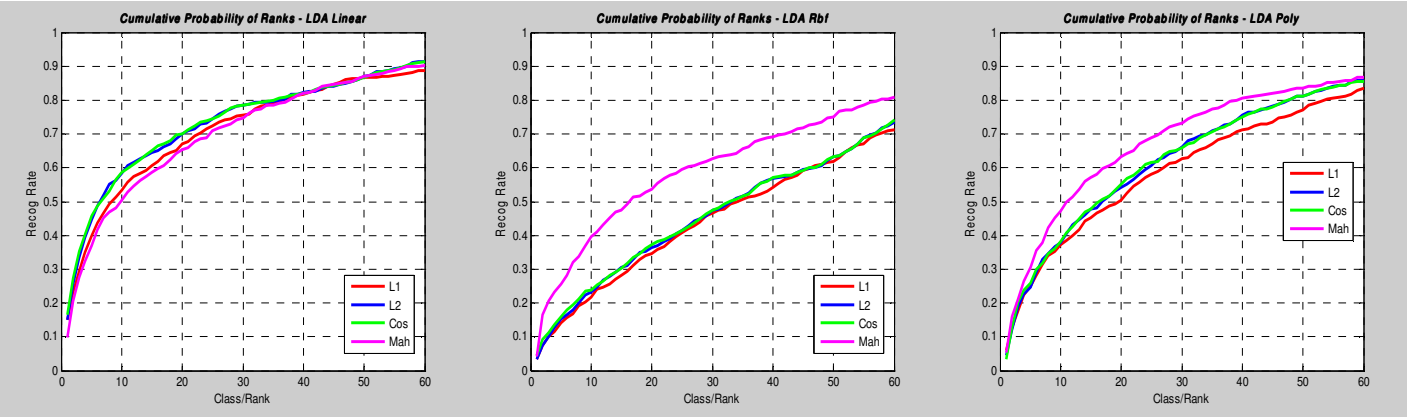



**Figure 31 CMS curves for all similarity measures – Lower Occlusion: LDA**

The intra-class evaluations revealed the following:

**PCA**

*Linear vs RBF*: The CMS chart indicates that the *linear variant* performs marginally better for the first twenty ranks and thereafter the RBF approach offers higher accuracies. This is supported by McNemar's probability that varies between 0% and 68% throughout the spectrum, with a clearly significant distinction at rank 16 where the linear approach is superior and between ranks 35-41 where RBF is the better algorithm choice.

*Linear vs Polynomial*: The *linear variant* is without a doubt the better intuitive and statistical choice, showing superiority with greater than 93% confidence for the first 35 ranks and occasionally for later ranks thereafter.

*RBF vs Polynomial*: As expected the *RBF approach* yields better results across the spectrum, with the only exception of no statistical difference at rank 1,3, 8 and 34

For the purpose of inter-class selection, one has a choice of both Linear and RBF variants, both offer similar performance. However the need for better early rank performance sees the linear approach being selected for the next assessment phase.

**LDA**

*Linear vs RBF*: There is no intuitive of statistical question the *linear approach* performs better across every rank.

*Linear vs Polynomial*: The *linear variant* is again the better choice, with 100% confidence for the first ten ranks, dropping to about 90% at rank-30, and thereafter showing no statistical difference. The CMS chart does however, display a more than marginal superiority across the ranks.

*RBF vs Polynomial*: The *polynomial approach* is without a doubt the algorithm of choice, showing clear statistical superiority from rank-6 onward.

**ICA**

*IA1 vs IA2*: The CMS chart promotes *IA1*'s performance across the spectrum, however, McNemar's specifies that its superiority is only clearly significant at ranks 1-3, 10-14 and 46-57. The *p*-values are always $< \approx 30\%$, suggesting that *IA1* is definitely the better choice.

*IA1 vs FA1*: Despite the CMS plots indicating that *FA1* being the better choice, statistically there is virtually no difference between the two approaches, albeit that the low *p*-values is in support of *FA1*.

*IA1 vs FA2*: InfoMax *Arch I*, shows confidence levels exceeding 90% between rank-5 to 15 and is clearly the better algorithm from rank 10-14. However, *FA2* does show very promising results in later ranks.

*IA2 vs FA1*: As one would expect, given both algorithms performance against *IA1*, *FA1* shows better accuracy levels on the CMS curves, has low *p*-values and is statistically superior at ranks 1-3, 12-14 and 42-47.

*IA2 vs FA2*: FastICA *Arch II*, by far outperforms *IA2*, having superiority at ranks 1and 2 and over 90% statistical confidence between ranks 20-50.

*FA1 vs FA2*: Statistically there is no differentiating decision separating these two algorithms, however early rank (2,7 and 8) confidence levels of 93% and the better CMS curve would suggest that *FA1* yields a better early rank performance levels.

In the class of ICA, it would appear that the implementation of *Architecture I*, offers the 'best' results with both the InfoMax and FastICA approaches. Given that the InfoMax approach is computationally more demanding, the FastICA approach is chosen for inter-class representation.

Looking at figure 32 below, column 1 again assesses the performance of PCA against LDA. Considering the CMS chart alone, the first 30 ranks show LDA's superior performance, with accuracy levels always exceeding PCA's by more than 10% at every rank. McNemar's simply provides statistical confirmation of this, LDA is significantly better for the first 43 ranks.

Column 2 provides the graphical analysis for PCA and ICA, where following the rank-1 results earlier, ICA's recognition rate surpasses by more than 20% in the first 25 ranks. Naturally ICA would be the algorithm of choice between the two. McNemar's again, simply confirms this notion throughout the rank spectrum.

The last column analyses the always popular choices of LDA and ICA, but again McNemar's completely and undoubtedly supports ICA's superiority and this gives us conclusive evidence that in the category of lower occlusion ICA (*FA1 + Cos*) is in fact the best algorithm to apply.

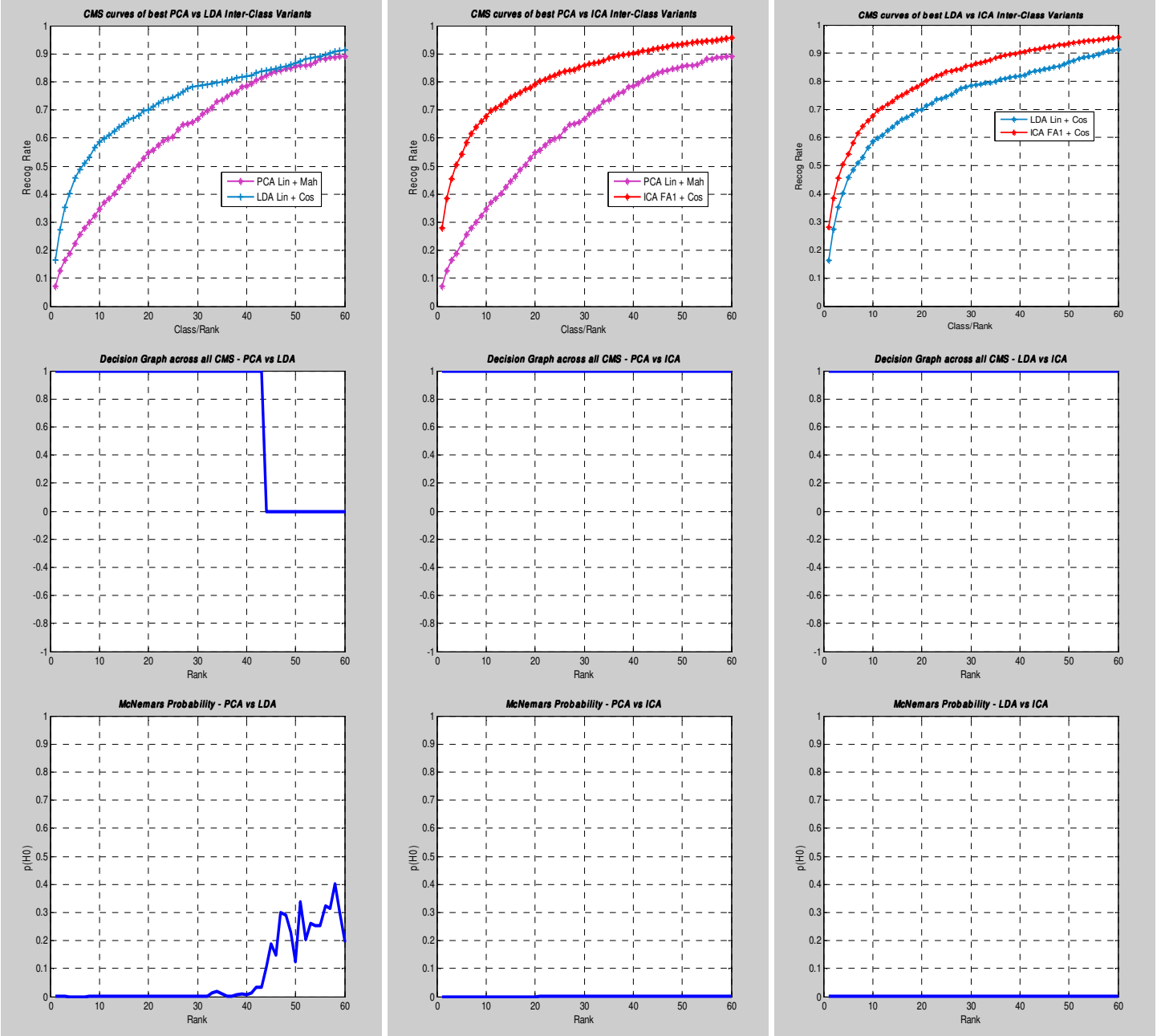



**Figure 32 McNemar's Test for the top 3 inter-class algorithms – Lower Occlusion**
Column 1: PCA Lin(Mah) vs LDA Lin(Cos), Column 2: PCA Lin(Mah) vs ICA FA1(Cos), Column 3: LDA Lin(Cos) vs ICA FA1(Cos)

In summary the intra-class analyses unveiled that in the classes of PCA and LDA, it was surprisingly not the non-linear approaches that offered the better performance, but instead the linear approaches that proved more beneficial. With ICA, it is apparently the spatial localisation of *Architecture I* that offers the advantage in dealing with occluded images. Given that global features would drastically be affected by 'missing' information, it could be expected that local feature comparisons would win the day. This is echoed in the results of the inter-class evaluation, where ICA was undoubtedly the better classed approach.

On a whole, the class of ICA both *Arch I* and *II* in conjunction with the Cosine (*Cos*) metric, performed comparatively well, suggesting that independent component application does offer significant advantages over the other classes. Looking at all results in their entirety, one can clearly see that appearance based methodologies are extremely sensitive to occluded imagery, particularly lower facial occlusion.

## 4.5 Upper Occlusion

The results in table 4 represent the rank-1 and CMS assessments under the test condition of upper facial obscurement i.e. sunglasses concealing between 20% and 30% of the face information. The recognition rates are expectedly low, however, they do show much improvement in comparison to the previous section of lower occlusion. This better performance could be attributed to the fact that the algorithms simply have more classification information available or/and that they are less sensitive to variations in upper facial occlusion.

The only projection-metric combination discrepancies are observed in linear and RBF variants of PCA. L1, which we've seen is a regular metric contender with PCA, offers the best rank-1 results, however overall performance is claimed by the Euclidean measure (L2). One also observes that in the classes of LDA and ICA, the metrics of Mahalanobis and Cosine respectively, once again provide the best accuracy levels.

| | Rank 1 Upper Occlusion Results | | | | CMS Results | |
|---|---|---|---|---|---|---|
| | *L1* | *L2* | *Cos* | *Mah* | *Highest Curve* | *Same as rank-1* |
| ***PCA*** | | | | | | |
| *Linear* | ***31.27%*** | 31.00% | 27.91% | 20.55% | L2 | No |
| *RBF* | ***30.91%*** | 30.46% | 28.50% | 20.96% | L2 | No |
| *Poly* | 31.46% | ***32.59%*** | 28.96% | 29.82% | L2 | Yes |
| ***LDA*** | | | | | | |
| *Linear* | 18.00% | 17.41% | 17.50% | ***18.77%*** | Mah | Yes |
| *RBF* | 14.00% | 18.36% | 15.28% | ***22.55%*** | Mah | Yes |
| *Poly* | 22.91% | 21.36% | 21.32% | ***27.91%*** | Mah | Yes |
| ***ICA*** | | | | | | |
| *Info A1* | 26.09% | 26.64% | ***40.86%*** | 33.32% | Cos | Yes |
| *Info A2* | 26.59% | 29.05% | ***46.73%*** | 29.18% | Cos | Yes |
| *Fast A1* | 34.36% | 31.36% | ***51.59%*** | 31.36% | Cos | Yes |
| *Fast A2* | 24.59% | 27.09% | ***49.64%*** | 27.09% | Cos | Yes |

**Table 4 Rank 1 Metric Results for the Upper Occlusion Probe Set**

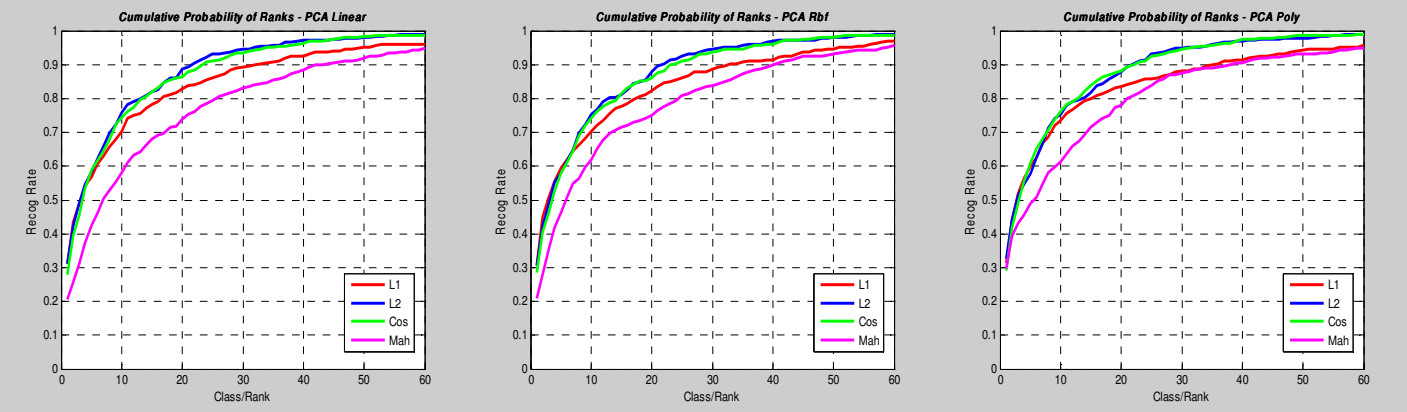



**Figure 33 CMS curves for all similarity measures – Upper Occlusion: PCA**

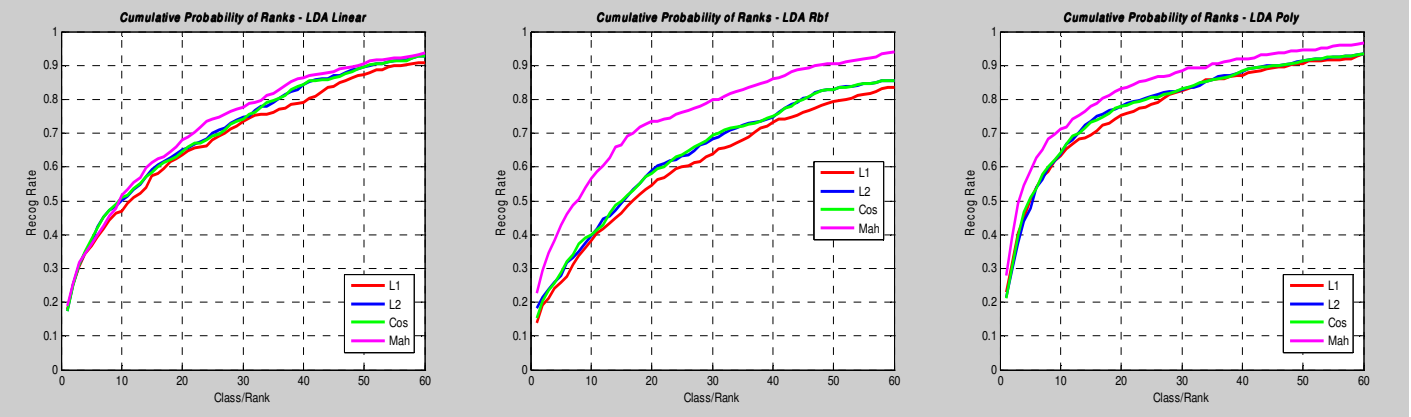



**Figure 34 CMS curves for all similarity measures – Upper Occlusion: LDA**

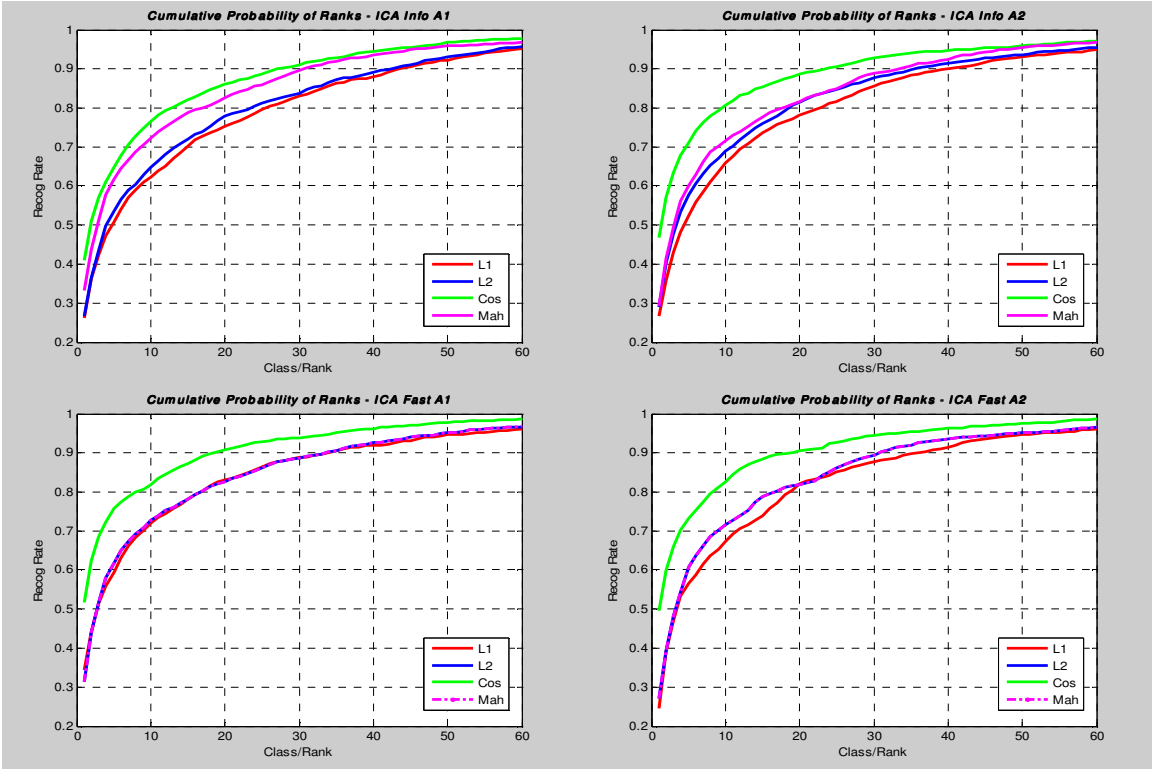



**Figure 35 CMS curves for all similarity measures – Upper Occlusion: ICA**

The intra-class comparisons yielded the following observations:

**PCA**

*Linear vs RBF*: Statistically there is no difference between these algorithms. The early rank CMS plots and low *p*-values do, however, recommend that the linear variant would offer a better performance.

*Linear vs Polynomial*: The statistical and CMS performance of both approaches are almost identical after rank-10, but for ranks 1-10 the *polynomial variant* is the better choice with confidence levels above 80% and clear statistical superiority at rank 2 and 3.

*RBF vs Polynomial*: Expectantly the *polynomial algorithm* is the better choice with an undoubtedly better statistical performance at ranks 2-6.

**LDA**

*Linear vs RBF*: Having confidence levels exceeding 75% for most of the rank spectrum and having distinct statistical support at ranks 5,6,9,18,46 and 47, the *RBF approach* clearly provides higher performance levels.

*Linear vs Polynomial*: The *polynomial variant* offers unanimous superiority, the CMS chart, Decision Graph and McNemar's *p*-values, throughout every rank in the spectrum except for the last two, where there is no statistical difference

*RBF vs Polynomial*: The *polynomial* results are not as superior as was the case against the linear variant, it does however, show a distinctly better statistical performance from ranks 2-48.

**ICA**

*IA1 vs IA2*: The CMS chart suggests *IA2*'s clear superiority for the first 30 ranks, McNemar's however points out that this higher performance is only statistically valid for the first 17 ranks (excluding rank 8 and 9) and thereafter no statistical difference exists between the two variants.

*IA1 vs FA1*: The CMS curve of *FA1* appears more than marginally above *IA1* throughout the rank spectrum. McNemar's, however, clarifies this significant difference to only the first 17 ranks.

*IA1 vs FA2*: *FA2* is the better algorithm, holding statistical significance for ranks 1-16 and 22-25.

*IA2 vs FA1*: The CMS chart suggests *FA1* as the better performing algorithm across the ranks, however the 95% and above confidence levels are only valid at ranks 2-6, 8 and 9.

*IA2 vs FA2*: Not as one would expect, but *FA2*'s superior performance is statistically limited only to rank 7 and 8.

*FA1 vs FA2*: Except for the first 8 ranks, the CMS curves are identical. Statistical evaluation reveals that at ranks 2-7 *FA1* has above 90% confidence in superiority and is convincingly separable at ranks 2 and 3.

Turning to the inter-class assessments depicted in figure 36 below, one finds that PCA, for the first time outperforms LDA. Shown clearly in column 1, McNemar's *p*-values are below 10% for most of the rank spectrum, occasionally etching up towards the 20% mark during the very early and very late ranks.

Column 2 shows the results of the PCA, ICA evaluation, and as one would expect given their respective rank-1 performances, ICA is the better algorithm, however, statistically this is only valid for the first 17 ranks.

In column 3, as one can see, LDA does not provide much competition against ICA, which is clearly the dominant class for the first 45 ranks.

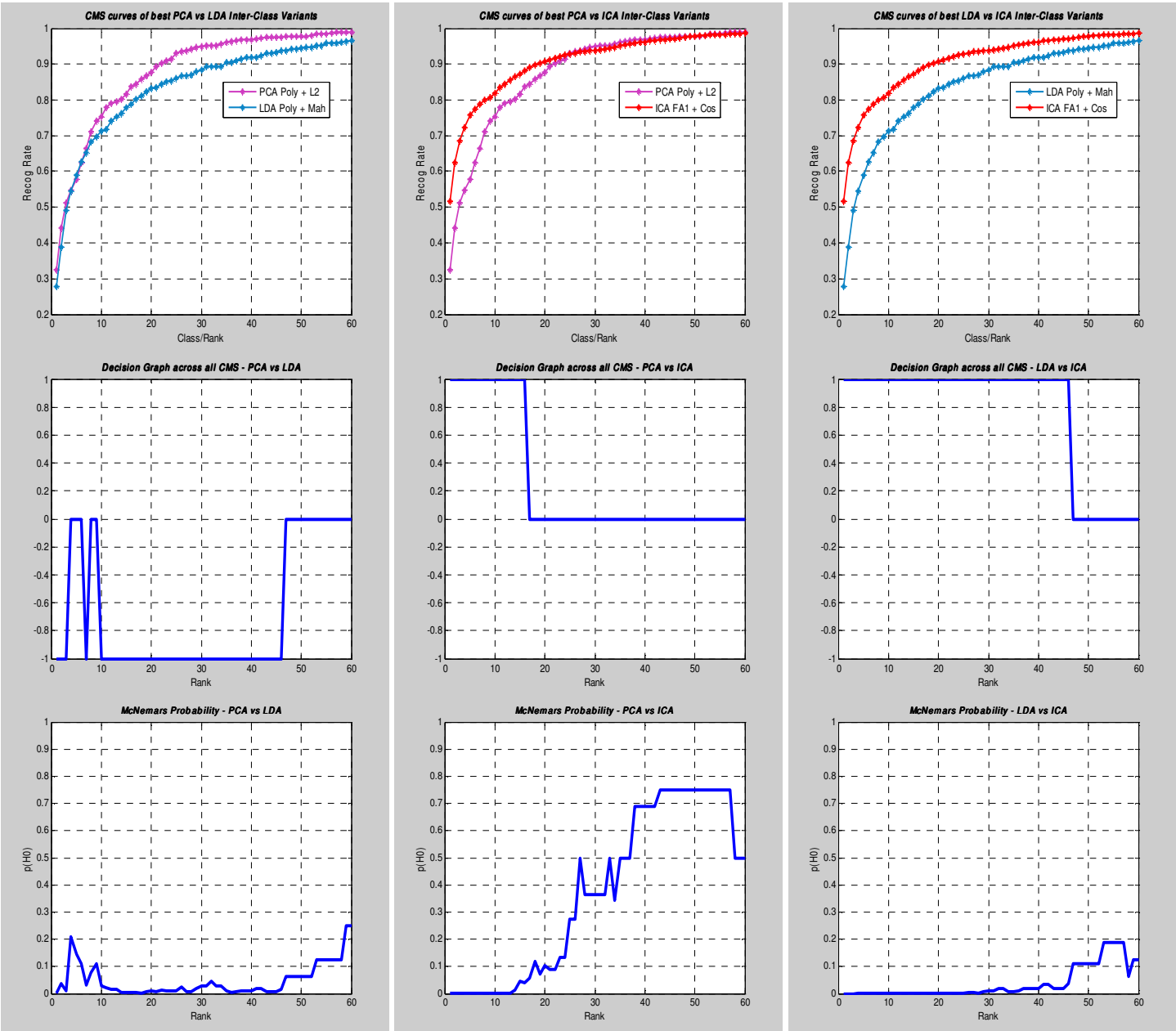



**Figure 36 McNemar's Test for the top 3 inter-class algorithms – Upper Occlusion**
**Column 1: PCA Poly(L2) vs LDA Poly(Mah), Column 2: PCA Poly(L2) vs ICA FA1(Cos), Column 3: LDA Poly(Mah) vs ICA FA1(Cos)**

In summary the intra-class results have revealed that in the classes of PCA and LDA, the polynomial approaches are best suited for upper occlusion. And in the class of ICA, it was once again *Architecture I*, in conjunction was the FastICA implementation that performed the best both within the class and between classes. In this evaluation category one could rank class performance from best to worst as ICA, followed by PCA and lastly LDA.

## 4.6 Temporal Delay

In the category of temporal variation, the evaluation criterion is very much similar to simple face identification, only the probe images are taken a period of time later. The neutral images (no expression, lighting or illumination inferences) of the first session are used to train the algorithms and the neutral images of the second session (taken two-three weeks later) are used as the testing set. Table 5 below presents the results.

| | Rank 1 Time Delay Results | | | | CMS Results | |
|---|---|---|---|---|---|---|
| | *L1* | *L2* | *Cos* | *Mah* | *Highest Curve* | *Same as rank-1* |
| ***PCA*** | | | | | | |
| *Linear* | ***84.55%*** | 74.55% | 74.55% | ***84.55%*** | L1 | Yes |
| *RBF* | 85.46% | 74.55% | 74.55% | ***86.36%*** | L1 | No |
| *Poly* | 78.18% | 74.55% | 73.64% | ***80%*** | L1 | No |
| ***LDA*** | | | | | | |
| *Linear* | 86.36% | ***90.00%*** | 89.09% | 87.27% | L1 | No |
| *RBF* | ***90.91%*** | 89.09% | ***90.91%*** | 90.00% | Mah | No |
| *Poly* | 90.00% | 88.18% | 86.36% | ***90.91%*** | Mah | Yes |
| ***ICA*** | | | | | | |
| *Info A1* | 86.73% | 82.36% | ***89.27%*** | 84.18% | Cos | Yes |
| *Info A2* | 82.55% | 85.27% | ***88.73%*** | 85.64% | Cos | Yes |
| *Fast A1* | 87.27% | 85.46% | ***90.00%*** | 85.46% | Cos | Yes |
| *Fast A2* | 82.73% | 84.55% | ***90.00%*** | 84.55% | Cos | Yes |

**Table 5 Rank 1 Metric Results for the Time Delay Probe Set**

Overall the rank-1 results seem fairly pleasing, with the LDA class yielding the best cross-metric results among all the classes. Discrepancies were again observed in both PCA and LDA classes. In the case of PCA the Mahalanobis measure offered the best rank-1 results, however, the City-Block (L1) metric yielded the best overall accuracy level across the rank spectrum.

In the LDA class, rank-1 results were somewhat scattered amongst the metrics but the CMS curves were provided by the L1 measure for the linear variant and the Mahalanobis measures for the RBF and

polynomial approaches. Looking at the ICA results, all the metrics performed rather well with accuracies above 80%, but once again the cosine metric combined the best with each projection variant.

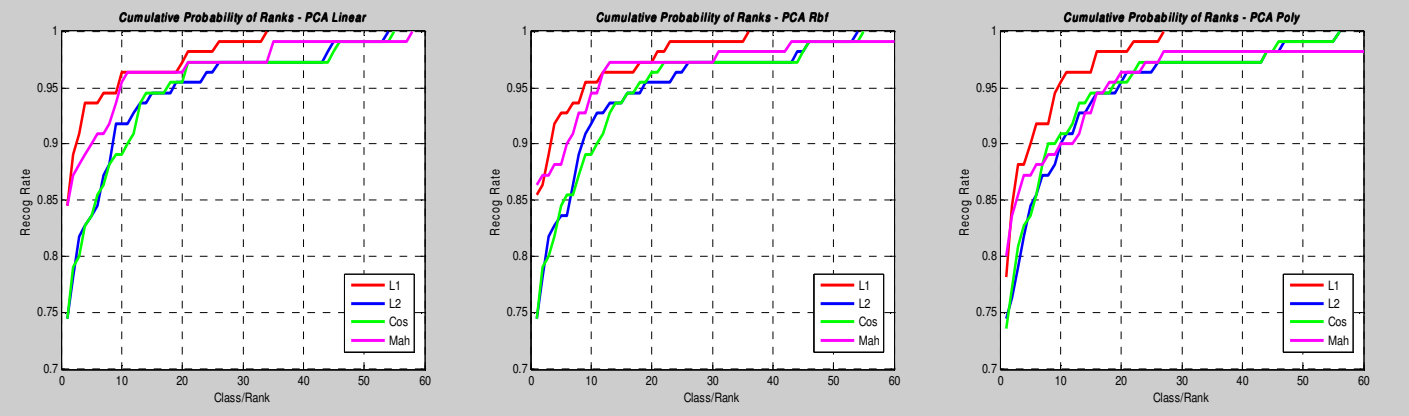



**Figure 37 CMS curves for all similarity measures – Time Delay: PCA**

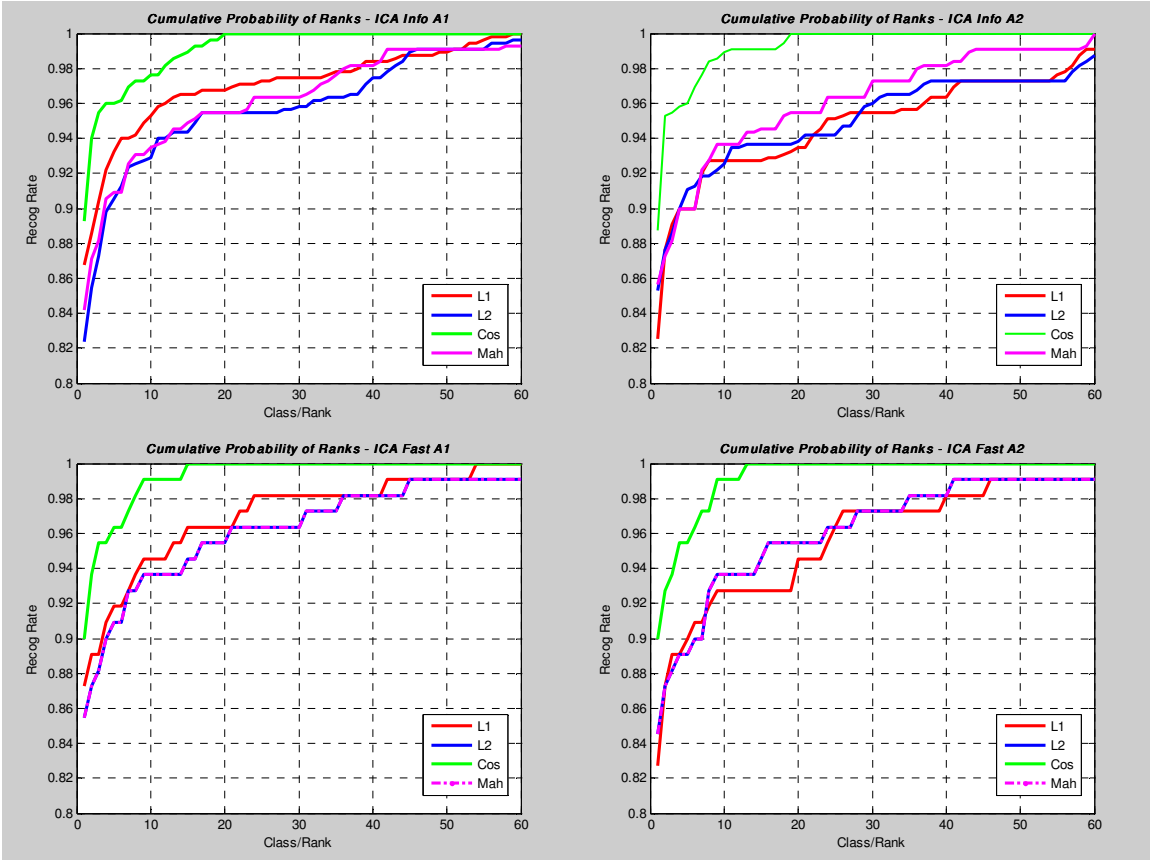



**Figure 38 CMS curves for all similarity measures – Time Delay: ICA**

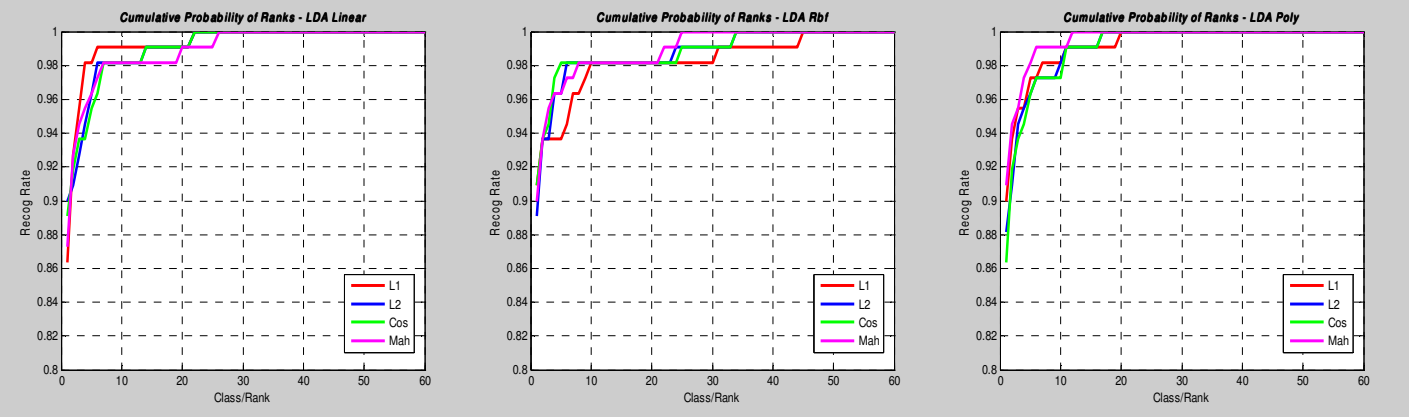



**Figure 39 CMS curves for all similarity measures – Time Delay: LDA**

The intra-class evaluations revealed the following:

**PCA**

*Linear vs RBF*: Statistically there is no difference between these two approaches, but the CMS curves and low *p*-values do suggest a better early rank performance by the linear variant.

*Linear vs Polynomial*: The *polynomial alternative* is definitely the early rank favourite, having a superior CMS curve and clear statistical significance at ranks 1,5 and 6.

*RBF vs Polynomial*: Except for the first two ranks, where the *polynomial algorithm* is the better choice, there is no statistical difference between these approaches

**LDA**

*Linear vs RBF*: There is no significant difference between these two alternatives, the RBF approach though, does offer better rank-1 recognition rates.

*Linear vs Polynomial*: Again both these approaches are statistically identical. The polynomial variant only offers a better rank-1 and 2 CMS performance.

*RBF vs Polynomial*: The only difference between these two implementations lies in the CMS curves, where the polynomial technique offers marginal superiority for the first 21 ranks.

**ICA**

*IA1 vs IA2*: There is no statistical difference observed, the only small advantage is offered by *IA2* in the first ten ranks of the CMS chart.

*IA1 vs FA1*: Both algorithms are statistically the same, *FA1*, like IA2, only offers marginal early rank CMS superiority.

*IA1 vs FA2*: There is no statistical or intuitive advantage offered by either algorithm

*IA2 vs FA1*: Again no difference between these variants, except for *FA1*'s marginally better rank-1 performance.

*IA2 vs FA2*: No statistical difference, except *IA2*'s early rank CMS advantage.

*FA1 vs FA2*: No statically different characteristics are observed, but *FA1* does appear with a small rank advantage in early ranks.

There is no statistical differences between any of the ICA variants and while it is true that any algorithm would perform equally well regarding temporal changes, inter-class selection is left to intuitive observation, where *FA1* seems the most promising approach to embrace.

Performing the inter-class evaluations, firstly between the classes of PCA and LDA we find that the CMS charts indicates LDA supremacy for the first twenty ranks, however, McNemar's shows us that this is only significantly true for the nine ranks, and thereafter slowly reaches the levels of equitable performance.

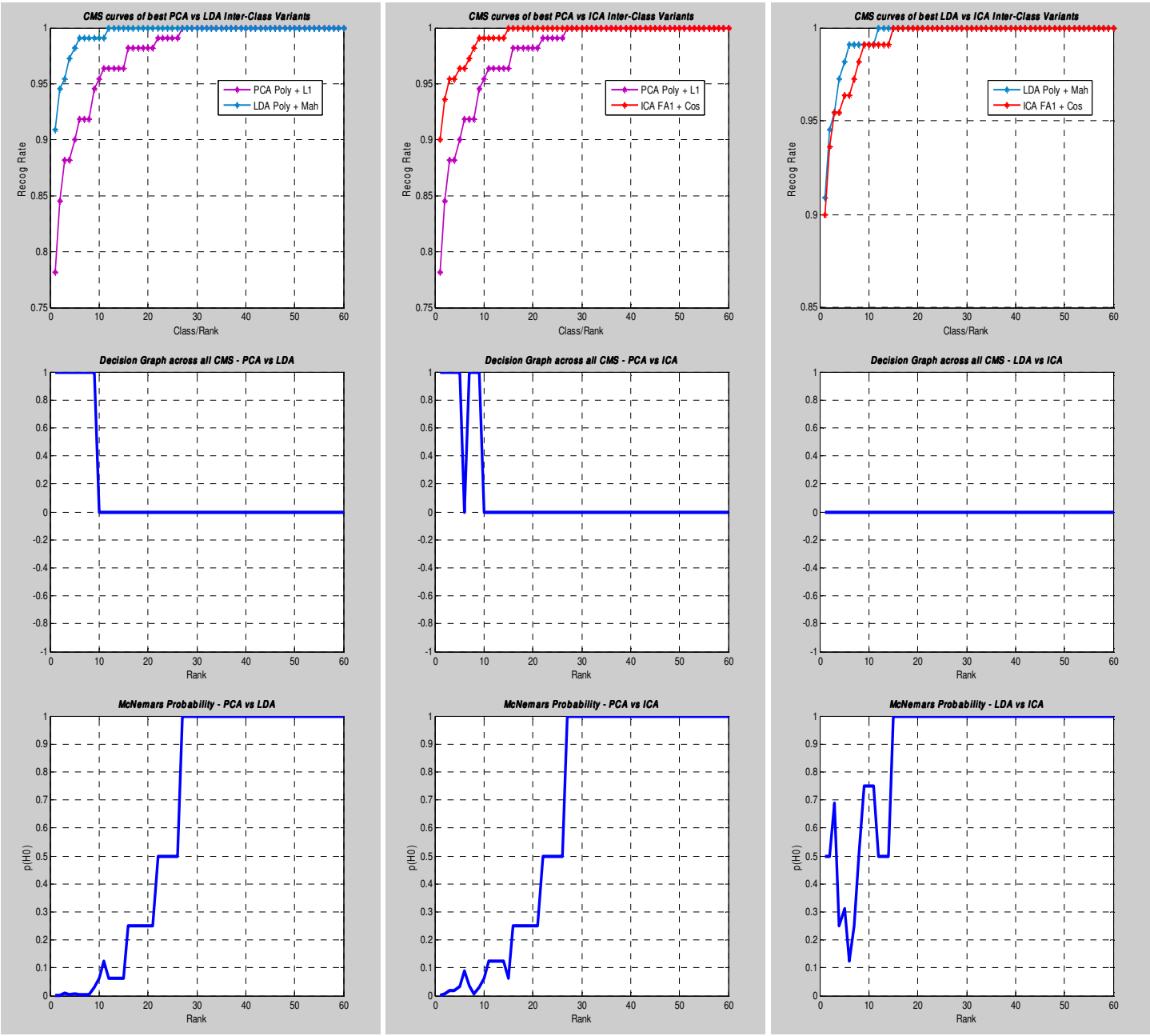



**Figure 40 McNemar's Test for the top 3 inter-class algorithms – Time Delay**
Column 1: PCA Poly(L1) vs LDA Poly(Mah), Column 2: PCA Poly(L1) vs ICA FA1(Cos), Column 3: LDA Poly(Mah) vs ICA FA1(Cos)

Turning to column two in figure 40 above, one finds ICA's performance, against PCA, very similar to that of LDA, albeit the spike at rank 7 shows that ICA would definitely still be the better choice.

Comparing the two leading classes in column 3, one finds no significant statistical difference between the classes of LDA and ICA. However, the early rank CMS chart, in conjunction with the low *p*-values

found in McNemar's $p(H0)$ graph, one gets the strong impression that it is the class of LDA that offers better subject separability and hence overall superior performance in temporal identification.

In summary, the intra-class assessments reveal that there is very little statistical distinction between the variants within each class. With the exception of PCA's polynomial algorithm, the 'best' algorithms in both the classes of LDA and ICA were selected based purely on intuitive analysis and while this may be the correct route in selecting a suitable algorithm of choice, it does not leave any room for making a clear cut decision into algorithm superiority. On the other hand, the inter-class testing have facilitated a more insightful decision, where LDA and ICA do clearly outperform the class of PCA.

## 4.7 Summary of Results

Using the described methodology this investigation evaluated the relative performance levels of the three most popular classes of appearance based subspace methodologies i.e. PCA, LDA and ICA. In conjunction with the assessment process one also sort to find the best metric combination that offered the best task specific advantage. The comparison was assessed in five categories of concern i.e. classification influences associated with changes in facial action, illumination, upper and lower facial obscurement and the effects of temporal delay. One first considered the rank-1 performances of each algorithm and confirmed their performance in accordance with the highest CMS metric curves that offered the highest accuracy levels across the rank spectrum; this either confirmed the rank-1 choices or revealed a metric that mostly offered a better overall performance or was not significantly different from the initial choice.

Having found the best projection-metric combination for each algorithm, one then wanted to establish the 'best' algorithm within each subspace class. Although the respective CMS curves offered a meaningful and sometimes very convincing indication of superiority order, it was felt that a deeper analysis would offer greater performance distinction. This was accomplished by employing McNemar's Hypothesis Protocol that aided one in making a decision that is not only intuitively correct but statistically sound as well. This combinatorial approach offered much more significant insight into the relative performance of each variant within the classes and also brought forth the best alternative that one could or should consider for inter-class comparisons.

The inter-class assessments were carried out in much the same fashion as the intra-class tests were, the CMS curves were used as the primary tool for obtaining an intuitive indication as to which class performed better and this was either confirmed, regionally clarified or nullified by the findings of McNemar's evaluation. The CMS charts, Decision Graphs and Confidence Probabilities were then cohesively used in revealing any significant, task specific, advantage that one class may offer over the next.

In the class of PCA, the following was categorically found:

*Expression*: There is no statistical difference between any of the variants, so surprisingly given the non-rigidity of the facial object, one would still expect a similar performance from the linear and non-linear implementations. The CMS curves do, however, indicate that it is the polynomial variant that offers the slight advantage in recognition rate. Based only on this intuitive deduction, the greater confidence is placed in the polynomial approach. Considering the metric combinations, one found that the best rank-1 results were obtained with the L1 measure and the best overall performance is offered by the L2 measure followed by cosine.

*Illumination*: The polynomial approach offered the best CMS accuracy levels, particularly in early ranks, where it was found to be statistically superior in ranks 1 and 2. The L1 measure was found to offer both the best rank-1 and overall performance that was closely followed by the L2 metric.

*Lower Occlusion*: These results were by far the lowest achieved and in some ways can even be considered as coincidental. Performance conclusions were reached simply on available evidence but one should leave room for further investigation. From current results, the linear and RBF implementations offered the best performance, with no statistical difference between the two. The CMS curves suggest that early rank advantage is given by the linear variant, while the RBF algorithm claims supremacy after rank-30. The Mahalanobis measure, surprisingly, offered the best metric results for this measure and was closely followed by the City-Block (L1) similarity measure.

*Upper Occlusion*: The polynomial variant again found intuitive and statistical superiority across most of the rank spectrum. The best results were found by employing the L1 measure at rank-1 and the Euclidean (L2), followed by the Cosine metrics for overall performance.

*Time Delay*: The polynomial algorithm, combined with the L1 metric, once again found early rank supremacy.

On average one could recommend that the best PCA performance levels are offered by the *Polynomial* algorithm in combination with the metrics of L1, followed by L2.

In the class of LDA, the results were as follows:

*Expression*: Both non-linear variants of RBF and Polynomial offered equal statistical advantage over the linear approach. The CMS curves were indicative of marginal polynomial superiority and it was

therefore selected for inter-class evaluation. The best metric results were reached using the Mahalanobis (Mah) measure, followed by cosine and Euclidean.

*Illumination*: The non-linear approaches again show early rank supremacy with the polynomial approach being statistically better than RBF at ranks 1 and 2. The Mahalanobis distance measure was also the best metric measure, followed by L2.

*Lower Occlusion*: The linear variant, without question, offered the best statistical and intuitive results across the spectrum. This conclusion however should be considered while also bearing in mind that the performance levels of appearance based methods are extremely sensitive to occlusion, specifically lower facial concealment. The best metric was again found to be the Mahalanobis measure, followed by cosine.

*Upper Occlusion*: The best performance was without a doubt offered by the polynomial approach in combination with the Mahalanobis metric.

*Time Delay*: There exists no statistical difference between any of the variants when it pertains to temporal face identification. CMS intuition does, however, suggest the marginally better performance being offered by the polynomial algorithm. Again the Mahalanobis measure offers the best results, followed by the cosine metric.

Overall one could suggest that better performance levels were obtained by the non-linear variants, specifically the *Polynomial* approach in combination with the Mahalanobis similarity measure.

In the class of ICA:

*Expression*: There was no statistical difference found between any of the ICA variants. *Architecture II*, in both the InfoMax and FastICA implementations, offered the best CMS advantage and *FA2* was selected solely on the fact that FastICA is computationally less costly than the InfoMax alternative. The best combinational metric was provided by the Cosine (*Cos*) measure.

*Illumination*: Again there was no statistical difference between any of the algorithms and *FA1* is selected purely on CMS intuition as the alternative that offers the most promising recognition rates. The Cosine metric again, also provides for the best combinatorial results.

*Lower Occlusion*: *Architecture I* was found to achieve the best statistical and intuitive early rank results. There was no clear distinction between the FastICA and InfoMax implementation, and again the FastICA variant was selected due to computational advantage. The Cosine metric proved clearly superior once again.

*Upper Occlusion*: The FastICA variants were found to perform much better than their InfoMax counterparts, with *Architecture I* reigning supremely and providing distinct rank 2 and 3 superiority over *Architecture II*. The cosine metric by far offers the best results.

*Time Delay*: Statistically, there was no significant difference between any of the approaches. *FA1* was however, again selected for inter-class evaluations based only on intuitive CMS performance. Cosine, once again offers the best metric results.

An overview of the ICA class shows that in the categories of expression, illumination and time delay, there is no significant statistical difference between any of the architectures and the choice of employing either the InfoMax or FastICA implementations does not affect the overall performance

rankings. In the case of Occlusion, however, *Architecture I* proved the most successful, reiterating the advantage that spatially localised vectors can offer over global (overlapping) feature vectors. In selecting the best metric combination for ICA, the Cosine measure was without a doubt the best distance measure in all categories.

In performing the Inter-class assessments the results were as follows:

*Expression*: LDA and ICA came out as the top classes, but only being superior to PCA at rank-1; other than that there was no statistical difference between any of the classes. Intuitively one could claim that PCA offers the best CMS result, followed by LDA and lastly ICA, however, this intuition would fall short of true performance conclusions, which in this case is similar for all the classes.

*Illumination*: LDA and ICA both claim statistical superiority over PCA for the first 7 ranks; ICA however, outperforms LDA for the first 3 ranks leading one to the conclusion that ICA is the best class to apply for the task of illumination changes

*Lower Occlusion*: While LDA may have outperformed PCA for the first 40 ranks, it was ICA (*Arch I*) that reigned statistically and intuitively supreme throughout the rank spectrum.

*Upper Occlusion*: ICA (*Arch I*) again performed the best amongst the three classes. PCA in this category, however, was found to statistically outclass LDA in both early and late rank evaluations.

*Time Delay*: Both LDA and ICA outperformed PCA for the first ten ranks. There was no statistical difference between these two classes, however, intuitive analysis would suggest that it is LDA that offers the best performance.

In summing up the class results, while it is true that the specific nature of the task may greatly influence the performance level of any algorithm, on average one could confidently recommend that the class of ICA is perhaps the most flexible and widely adaptable subspace methodology that could be applied, followed by the classes of LDA (non-linear) and PCA (non-linear), respectively.

In making this claim, there are also two important questions that need to be considered:

1) *Are these results significant with respect to the choice of training set?* Given that the objective of this investigation was to provide a real world training set i.e. as few neutral images of the subject possible, and then evaluate each algorithm due to pure variation in major performance influencing factors. One could confidently state that these results are significant and offer much needed insight into the practical demands of real world systems.

2) *Do these results depend on the number of subspace dimensions?* In validating the FERET heuristic during the implementation phase, rank-1 tests from dimension 1-220 for PCA and ICA and 1-110 for LDA were run. It was found that the metric ordering in each category was at some stages changed by the dimensionality selection i.e. the graphs crossed over. A particular example was the Mahalanobis measure whose accuracy level rose significantly as one neared 100% of dimensionality retention. As to how this exactly affects the assignment of the 'best' CMS metric, one cannot say at this stage as further investigation would be required, however, the overall performance and relative rankings of the individual algorithms are not believed to change. It is therefore vital that a standard of dimensional retention be adopted so that effective literary comparisons can be made. In this investigation the world standard FERET information retention of 96% was followed, thereby allowing for relatively comparable assessments.

## 4.8 Comparison to Previous Literary Works

At this point it is worth mentioning that many of the papers, against which these results will be compared to, do not employ the AR database and instead use other non-standard and the standard (FERET) datasets. Their evaluations were based on three main categories: expression, illumination and temporal changes. It should also be noted that most of these results are primarily based upon rank-1 accuracies and limited metric combinations. This should not, however, pose any problem as the comparison to their work will be made only on the *relative* performance of the results.

So just where do the results of this investigation fit in amongst prior literary conclusions?

In comparing PCA and ICA; firstly in the category of *temporal changes* Bartlett *et al* [20] showed that both ICA architectures (InfoMax) with the Cosine measure outperformed PCA (L2 and Cos). Draper *et al* [12], Liu *et al* [21], Phillips *et al* [63] and Delac *et al* [5] claim that ICA *Arch II* (Cos) finds the most superior result. The results in Section 4.5 are consistent with both these claims. Given that there was no statistical difference between any of the ICA variants, both architectures whether implemented using InfoMax or FastICA, are shown to outperform PCA. This completely contradicts the claim made in [12] however, where *Arch I* is claimed to be suboptimal for face recognition.

In the category of *expression*, [20] claims that both architectures are superior to PCA, [5] claims that the InfoMax algorithm with *Arch I* and [21] supports the performance of *Arch II*. The results in Section 4.1 are found to concur with all these claims, given that again there is no statistical difference between the architectures, either or both would outperform PCA for rank-1 accuracies. However, there is no performance distinction thereafter; in making this claim and these results disagree with those found by [5] who also claim PCA(Cos) outperforms ICA at higher ranks.

In the case of *illumination* [5] claims ICA *Arch II* outperforms PCA (L2) and that PCA(L1) outperforms ICA at higher ranks. Again the results of Section 4.2 concur with the findings that ICA outperforms PCA at early ranks, however any claim regarding later PCA statistical advantage is strongly rebuked.

In considering works examining all three categories (expression, illumination and time delay) Baek *et al* [22] claim that PCA (L1) outclasses ICA *Arch I* at rank-1 results. While the results of this investigation agree with the claim under the condition of expression changes, it also strongly contradicts this conclusion under the condition of illumination and temporal changes, where *Arch I* is found to offer better rank-1 results than PCA. Moghaddam [23], on the other hand, found that there existed no statistical rank-1 difference between the ICA *Arch II* - (L2) and PCA(L2). Findings in this paper disagree with this claim and show that in comparing the rank-1 results in the category of expression, PCA(L2) outperforms ICA(L2) and this performance is reversed in the categories of illumination and temporal change where ICA in fact yields significantly better results.

Turning to the class of LDA, Liu *et al* [21] claim that the ICA *Arch II* significantly outclasses LDA. Results in Section 4.2 may agree with this claim, however no major statistical difference was found between ICA and LDA in the categories of expression and time delay, in fact LDA can be intuitively supported as offering better performance levels. This claim is also evidently supported by [5] who shows that LDA either outperforms or is not significantly different to PCA and ICA in the category of expression.

In comparing LDA to PCA Belhumer [25] and Navarette *et al* [26] clearly demonstrate LDA's rank-1 superiority, which is consistent with findings in Section 4.1, 4.2, 4.3 and 4.5. The opposite claim is made by Beveridge *et al* [24] who states that LDA performs uniformly worse than PCA. The only evidence of this is in Section 4.4 (lower occlusion), however Beveridge's work made no effort in

exploring the effects of occlusion so in fact the results in this investigation can be said to completely contradict claims made by [24].

In addressing literary claims that were made with regard to the performance of non-linear kernel variations of PCA and LDA, the work in this paper partially supports the claim made by Yang [55], that KPCA and KDA achieve lower error rates than standard PCA, LDA and ICA. The non-linear kernel variants do in general achieve better results than the standard linear implementations of PCA and LDA, however, the most that can be said when being compared to ICA is that no-statistical difference can be found between the classes. Navarette [26] and Pavlou [1] demonstrate that KDA outperforms KPCA and this is consistent with finding in the categories of expression, illumination, lower occlusion and time delay.

While the results in this paper may not totally agree or completely disagree with many literary claims, one hopes that this investigation will give the reader greater insight as to why and under what conditions certain claims were made. In doing so this paper hopes to invoke a better sense of clarity when deciding which algorithm would be better suited for the task at hand.

# Chapter 5

# Hybrid Formulation

## 5.1 Introduction

While it may be true that a robust classifier could be designed to effectively handle any one of the performance influencing factors, it is extremely difficult for an appearance-based technique to robustly deal with all the influencing variations [27]. Each individual classifier has a different sensitivity to different changes in facial variation and as was reported by Phillips *et al* [28], appearance-based methods show different levels of performance for different subsets of images. In their analysis of ICA and PCA, Bartlett *et al* [20] also reported that when incorrect classifications were made, it was very rare ( < 3%) that the algorithms assigned the same incorrect identity class. The above findings strongly suggest that different classifiers contribute different and hence complementary information about the classification task. A combination or *fusion* scheme involving the different face classifiers, which integrates multiple sources of evidence is therefore more likely to yield an overall improvement in both the efficiency and accuracy of the identification system. And while this may not solve the problem regarding influencing factors, it will definitely alleviate the impact they have on performance levels. It is for this reason that Hybrid approaches have been the subject of intensive research over the last ten years and consequently have gained a very special status among face recognition systems.

In classifier fusion, the outputs of individual classifiers are combined by a second (*combination*) classifier according to a pre-defined combination rule. Classifier combination can essentially be implemented at three levels:

a) *Fusion at the feature extraction level*: Data obtained at sensor level usually enter a system as feature vectors. Since sensor information are generally independent of each other, it is acceptable to concatenate two or more feature vectors into a single more discriminating higher dimensional vector associated with every class.
b) *Fusion at a confidence or matching score level*: Each system provides a similarity or matching score indicating the proximity of the input (probe) vector with regard to the template (training) vector. These scores are then combined either directly or by functional relation to reveal a confidence score that is employed to assert the veracity of the claimed identity.
c) *Fusion at a decision level*: The feature vectors are independently classified by each classifier and the final accept/reject decisions are then consolidated to yield the final classification. An example of this type of fusion would be a majority rules voting system.

The use of classifier fusion has produced many combination techniques over the years. One popular approach has been the idea of bagging [70], which manipulates the training-data with sub-sampling. Another common algorithm, boosting [71], also manipulates the training data, but with emphasis on the training of samples that are difficult to classify. Recently, probability-based strategies have become popular in pattern recognition; Kittler *et al* [29] provided a theoretical framework for combining various base classifiers. They reviewed several common strategies, which included the *product rule, sum rule, max rule, min rule* and *median rule*. These rules make final decision based on posteriori probabilities generated by base classifiers. The experimental results showed that the best performances were obtained when using the sum and median rules. Ross and Jain [49] showed that when the error

introduced by each classier, due to problems in acquisition and feature extraction processes, is unknown, the errors in estimating the posteriori probabilities become very large and hence it is better to directly combine individual scores.

In combining these scores however, it is common that one may experience one or more of the following problems [49]:

- Non-homogenous score types, whereby different methods are employed in obtaining these values such as distances and similarity measures (L1 and Cosine).
- The score ranges may be entirely different e.g. [0,100] or [0,1000].
- The score distributions may be entirely different.

An important part of the fusion process is therefore score normalisation, whereby the scores obtained from multiple frameworks are modified and transformed into a domain of common scale and range before combining them [49,69]. Popular normalisation schemes includes the approaches of *MinMax, Decimal Scaling, Z-score, Median, Tanh Estimators* and *Double Sigmoid Functions* [49]. Each approach may of course perform better given a certain type of data, but in general the schemes of MinMax and Z-score approaches were found to be the most robust and efficient [49].

Over the years many practical applications of combining multiple classifiers have been developed. Brunelli and Falavigna [41] presented a person identification system by combining outputs from classifiers based on audio and visual cues. Jain *et al* [67] integrated multiple fingerprint matchers to develop a robust finger print verification system. Hong and Jain [42] implemented a decision fusion scheme to combine faces and fingerprints for personal identification and Marcialis and Roli [16] exploited the fusion of PCA and LDA for face verification.

This investigation will propose a simple weighted sum fusion strategy that will combine the best results of the previous section in an effort to provide a publicly available subspace methodology that yields good overall performance across all influencing factors concerned.

## 5.2 Fusion Strategy

Looking at the results from section 4, two classes of subspace methods stood out, those being the classes of ICA and LDA. The hybrid formulation will seek to combine and exploit the powerful data representation of ICA and the unique class discrimination of LDA. Of course the best projection-metric combination from each class is selected so as to develop the most advantageous hybrid performance available. In the LDA class, the polynomial approach in combination with the Mahalanobis similarity measure was selected. In the class of ICA, in the categories of Expression, Illumination and Time delay, it was observed that there were no statistical differences between any of the variants, however FA2 (*Cosine*) did seem intuitively better in the category of Expression and FA1 (*Cosine*) did come out very strong in the categories of Illumination and Time delay; also in the occlusion categories FA1 was clearly the superior algorithm. It was therefore decided to combine both ICA architectures, *Arch I* and *II*, using the computationally more efficient FastICA implementation. By incorporating both architectures, one hoped to harness the advantages offered by both spatially localised and global independent components.

In considering at which level the classifiers should be combined, the first level of feature combination had to be ruled out in this application, given that all the images (feature vectors) were obtained using the same sensor (camera equipment) and therefore did not offer any independent advantage. In the case of decision-level fusion, also known as *hard fusion*, Ross and Jain [49] reported their conclusive

findings which indicate that the approach of classification before fusion, actually performs poorer than the confidence score level. The matching score level or *soft fusion* was found to offer the best tradeoffs in terms of information content and ease of fusion [49]. It was therefore selected as the most appropriate level of fusion for this particular application. The similarity measures from the relevant metric of each algorithm will be taken as inputs to the combinational classifier. The similarity scores inherently contain the top rank-*N* information and therefore serve as the best information medium. Normalisation of both the Cosine and Mahalanobis measures are performed employing the MinMax scheme, shown below, resulting in a common range of [0 100].

$$S' = \frac{S - \min\{S_k\}}{\max\{S_k\} - \min\{S_k\}} \quad \text{for k = 1,2,…,} P \tag{39}$$

In combining the different metric measures, the *weighted sum rule* is selected as the fusion rule. Despite its simplicity, the sum rule often outperforms other combination schemes [29, 49] and because of its linear model it is proven to be more tolerant to noise signals, unlike the *product* rule that severely magnifies any noise contributions. The combined matching score will be calculated as follows:

$$S_{comb} = w_1 S'_{FA1} + w_2 S'_{FA2} + w_3 S'_{LDA} \tag{40}$$

The weights associated with each classifier's matching score are defined by a confidence function to represent the contributional degree of each classifier. This paper proposes two methods that establish these weight values:

*Method 1*: This approach is an intuitive generalisation, whereby the weightings are defined as per class performance level in each of the five relevant evaluation categories presented in Section 4. LDA performed well in the categories of Expression and Time Delay, hence obtained two fifths or 40% of the weighting; similarly FA1 outperformed the other algorithms in both occlusion categories and will also receive 40% of the weighting. ICA as a class performed well in the category of illumination, and

given there was no statistical difference between the variants and that *Arch II* must also be incorporated, FA2 is assigned the final 20% of the weighting.

*Method 2*: Because recognition accuracy of each component classifier is directly related the confidence one has in its abilities, one can generate a confidence function as a weighting function. Letting $r_i$ be the recognition accuracy of each classifier, the sum of recognition accuracies is given by:

$$r_{sum} = \sum_{i=1}^{q} r_i \tag{41}$$

where $q$ is the number of classifiers being combined. The associated weightings are then given by:

$$\sum_{i=1}^{q} w_i = \sum_{i=1}^{q} \frac{r_i}{r_{sum}} = 1 \tag{42}$$

Once the hybrid matching score has been calculated, final classification and evaluation is performed. The findings of the parallel fusion formulation are presented in the proceeding sub-section. (Also the journal paper to be submitted for the *2007 International Joint Conference on Neural Networks* can be found in Appendix C)

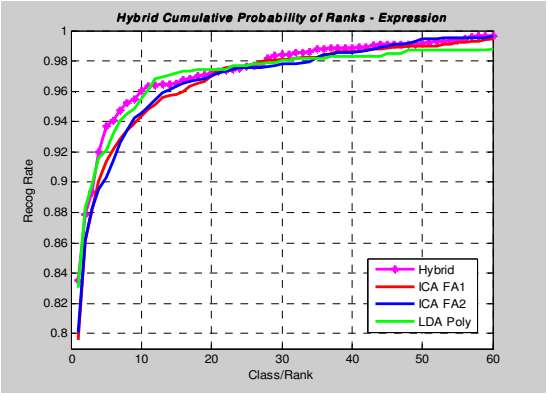
## 5.3 Hybrid Results and Comparisons

The results in this section are presented in much the same way as in Section 4; table 6 below captures the rank-1 results for the hybrid implementations of methods 1 and 2 in each evaluation category. For comparative purposes the rank-1 results of the component algorithms have also been inserted. Following the table one will find the best categorical CMS curves (between methods 1 and 2) and commentary regarding McNemar's protocol.

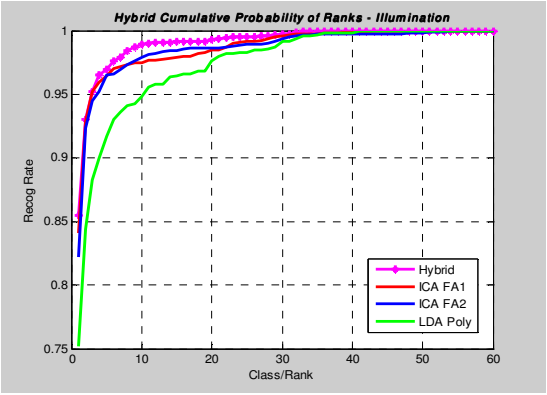
Looking at Table 6 and Figure 41, one observes that the Hybrid algorithm performs exceptionally well, having the best overall recognition rate performance in four out of the five categories; only in the category of Time Delay did LDA perform better, but only by a tiny magnitude.

| Rank 1 Hybrid Results | | | | | | CMS Results | |
|---|---|---|---|---|---|---|---|
| | *FA1* | *FA2* | *LDA* | *Hybrid* | | *Highest Curve* | *Same as rank-1* |
| **CATEGORY** | | | | *Method 1* | *Method 2* | | |
| *Expression* | 79.59% | 80.05% | 83.05% | ***83.5%*** | 82.36% | Hybrid | Yes |
| *Illumination* | 84.14% | 82.23% | 75.18% | ***85.5%*** | 84.95% | Hybrid | Yes |
| *Lower Occ* | 27.86% | ***28.27%*** | 5.54% | 17.5% | 27.77% | Hybrid | No |
| *Upper Occ* | 51.59% | 49.63% | 27.91% | 51,59% | ***52.81%*** | Hybrid | Yes |
| *Time Delay* | 90.00% | 90.00% | ***90.91%*** | 90.00% | 90.00% | LDA | Yes |

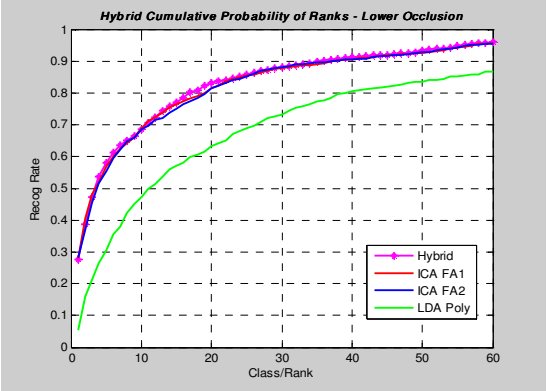
**Table 6 Rank-1 Comparative Hybrid Results**



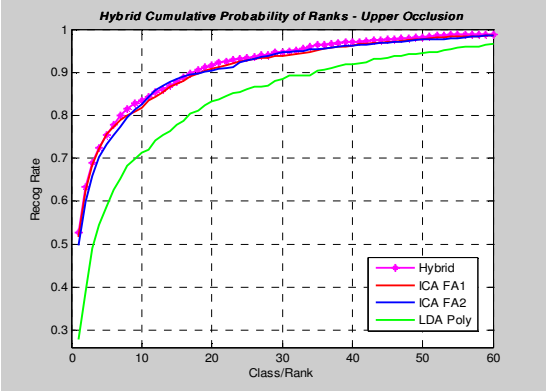

(a)



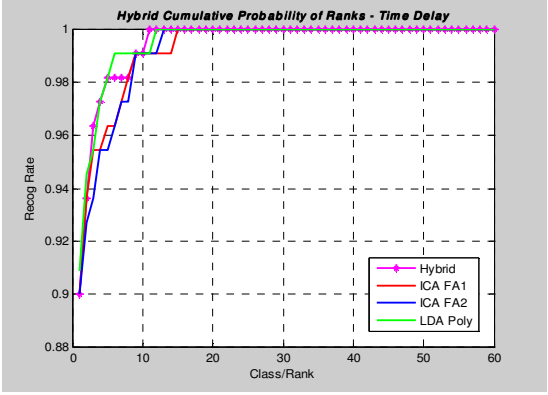

(b)




(c)




(d)




(e)

**Figure 41 Hybrid CMS Charts**

**(a) Expression (b) Illumination (c) Lower Occlusion (d) Upper Occlusion (e) Time Delay**

Comparing the Hybrid weighting approaches, although very different, both methods performed very well, with method 1 finding superior claim in the categories of Expression and Illumination and method 2 being the better performer in the Occlusion categories. Both performed equally well in the category of Time Delay. Statistically there is no significant difference between the results of either approach.

Turning to McNemar's analyses, the categorical results were as follows:

*Expression*: Statistically there is absolutely no significant difference between the Hybrid results and any of the constituent algorithms. The only advantage that is offered by the Hybrid formulation, over the ICA algorithms, is the marginal superiority in CMS accuracy levels.

*Illumination*: Comparing the Hybrid and ICA algorithms, one again finds no statistical difference between the algorithms. When comparing the results to LDA however, one finds that the Hybrid offers a CMS advantage for the first 15 ranks and is statistically superior for the first 4 ranks.

*Lower Occlusion*: Again there is no significant difference between the Hybrid and ICA results, but the higher CMS curve and low $p$-values between ranks 1-9 and 20-40 do indicate a higher confidence in the Hybrid performance. When comparing the results to LDA, as one would expect, the Hybrid formulation displays 100% confidence in statistical superiority throughout the rank spectrum.

*Upper Occlusion*: The results mimic those found in lower occlusion, whereby no significant difference exists between the Hybrid and ICA algorithms and the Hybrid approach offers complete statistical and intuitive superiority over LDA.

*Time Delay*: Statistically there is no difference between any of the algorithms. The Hybrid approach only offers a small early rank accuracy level advantage over the ICA constituents.

In summary, this investigation has proposed an integration scheme, which combines the output matching scores of the best categorical subspace methodologies. This was done in an effort to establish a combinatorial algorithm that yields improved recognition accuracies in the realm of face identification. The experimental results, although not vastly superior are nonetheless very encouraging and highlight the fact that combinational strategies can in general lead to more accurate face recognition levels than those achieved by individual classifiers. Although the proposed approach only explores one aspect of hybrid synthesis and the results are not statistically superior to the best categorical constituent algorithms, the framework has been made scalable so that future investigations can easily incorporate and improve other face recognition modules in the quest to realise a truly superiorly performing hybridised subspace methodology.

# Chapter 6

# Conclusion

## 6.1 Conclusion

This research investigation presented a rather rare comparative study of three of the most popular appearance-based face recognition projection classes, PCA, LDA and ICA along with the four most widely accepted similarity measures of City Block (L1), Euclidean (L2), Cosine and the Mahalanobis metrics. Although comparisons between these classes can become fairly complex given the different task natures, the algorithm architectures and the distance metrics that must be taken into account, an important aspect of this study was the completely equal working conditions that were provided in order to facilitate fair and proper comparative levels of evaluation. In doing so, one was able to realise an independent study that significantly contributes to prior literary findings, either by verifying previous results, offering further insight into why certain conclusions were made or by providing a better understanding as to why certain claims should be disputed and under which conditions they may hold true. In the process, by firstly exploring previous literature with respect to each other and secondly by relating the important findings of this paper to previous works one was also able to meet the primary objective in providing an amateur, in the field of face recognition, with a good understanding of publicly available subspace techniques.

The experimental setup yielded 10 different projection-metric combinations to be evaluated, comparing both the linear and non-linear variants of the PCA and LDA classes as well as both Architecture I and II in the class of ICA, along with both the InfoMax and FastICA implementations. In presenting any findings the evaluation procedure incorporated both the intuitive recommendations provided by the FERET standard of CMS charts and statistical analyses in the form of McNemar's protocol. In general it was found that one of the most important influencing factors was in fact the nature of the specific task, some applications such as facial identity recognition, are rather holistic and may require spatially overlapping feature vectors for higher performance levels, while other tasks such as occlusion, are more localised and perform better with spatially disjoint feature vectors.

In the categories of *Expression*, *Illumination* and *Temporal* variations it was found that the classes of LDA and ICA outperformed PCA at early ranks, with LDA offering the intuitively higher performance levels in the categories of expression and time delay and ICA claiming intuitive and statistical superiority in the task of illumination. In the categories of *Occlusion*, both upper and lower, the locally spatialised feature vectors introduced by ICA *Arch I*, claimed unprecedented supremacy throughout most of the rank spectrum. In evaluating the best metric combinations, one found that the City Block measure followed by L2 was most successful when used with PCA. The Mahalanobis metric followed by Cosine found good results in the class of LDA and the mutually orthogonal basis vectors offered by the Cosine measure was by far the most successful when combined with any of the ICA variants. In summation of the class evaluations, while it is true that the specific nature of the task may greatly influence the performance level of any algorithm, on average one could confidently recommend that the class of ICA is perhaps the most flexible and widely adaptable subspace methodology that could be applied, followed by the *non-linear* variants of LDA and PCA respectively.

Having established the 'best' algorithms for each category, this investigation wanted to explore the complementary classification nature offered by these algorithms and consequently proposed a fusion scheme to explore the impact of a hybridised face recognition module that was more likely to boost the accuracy levels of the identification system. Following Kittler's [29] theoretical framework for combining classifiers, a simple weighted sum fusion strategy was adopted for evaluation. Although this approach only explored one technique of hybrid synthesis and the results are not statistically superior to the best categorical constituent algorithms, the CMS charts are very encouraging and highlight the fact that combinational strategies can in general lead to more accurate face recognition levels than those achieved by individual classifiers. The framework has been made scalable so that future investigations can easily incorporate and improve other face recognition modules in the quest to realise a truly superior performing hybridised subspace methodology.

As with any comparative study, there will be limitations. This study followed a specific implementation methodology and consequently did not explore certain aspects. For one, the effects of registration errors or different pre-processing schemes were not incorporated. Only three classes of subspace techniques were evaluated, many other methodologies also exist and deserve equal investigative attention. Also this paper only employed two evaluation criteria, the FERET standard of Cumulative Matching Score charts and the statistically defining McNemar's protocol, one also did not explicitly evaluate the impact of computational costs and more importantly the influence on performance accuracies due to the underlying data distributions. As a result, it is difficult to recommend, with 100% confidence, the best technique for any novel domain. This investigation therefore considers completely exploring and understanding this relationship, by means of a larger scale database creation and other statistical evaluations, to be future work both for the author and for others in the field of appearance-based object recognition.

# References


[1] M. Pavlou, '*Face Kernel Extraction from Local Features*', Doctoral Thesis, Faculty of Engineering and Physical Sciences, University of Manchester, 2005.

[2] A. Nes, '*Hybrid Systems for Face Recognition*', Master of Science Graduate Thesis, Faculty of Information Technology, Norwegian University of Science and Technology, 2003.

[3] S. Tseng, '*Comparison of Holistic and Feature Based Approaches to Face Recognition*', Master of Science Graduate Thesis, Faculty of Applied Science, Royal Melbourne Institute of Technology University, 2003.

[4] R. Gross, J. Shi and J. Cohn, '*Quo vadis Face Recognition?*', CMU-RI-TR-01-17, Robotics Institute, Carnegie Mellon University, 2001.

[5] K. Delac, M. Grgic, S. Grgic, '*Independent Comparative study of PCA, ICA and LDA on the FERET Data set*', Proc. of the 4th International Symposium on Image and Signal Processing and Analysis, pp 289-294, 2005.

[6] W. Zhao and R. Chellappa, '*Image-based Face Recognition: Issues and Methods*', in Image Recognition and Classification by B. Javidi, pages 375-402. Mercel Dekker, 2002.

[7] Proceedings of the International Conferences on Automatic Face and Gesture Recognition, 1995-1998.

[8] Proceedings of the International Conferences on Audio and Video-Based Person Authentication, 1997, 1999.

[9] P.J. Phillips, H. Moon, P. Rauss and S.A. Rizvi, '*The FERET Evaluation Methodology for Face Recognition Algorithms*', in Proc. Conference on Computer Vision and Pattern Recognition, pp 137-143, 1997.

[10] S.A. Rizvi, P.J. Phillips and H. Moon, '*The FERET Verification Testing Protocol for Face Recognition Algorithms*', in Proc. International Conference on Automatic Face and Gesture Recognition, pp 48-53, 1997.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, '*XM2VTSDB: The Extended M2VTS Database*', in Proc. International Conference on Audio and Video-Based Person Authentification, pp 72-77, 1999.

[12] B.A. Draper, K.Baek, M.S. Bartlett and J.R Beveridge, '*Recognizing Faces with PCA and ICA*', Computer Vision and Image Understanding, Vol. 91, pp 115-137, 2003.

[13] A.M. Martinez and A.C Kak, '*PCA versus LDA*', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, pp 228-233, 2001.

[14] G.Shakhnarovich and B.Moghaddam, '*Face recognition in Subspaces*', Handbook of Face Recognition, Springer-Verlag, 2004.

[15] Y. Adini, Y. Moses, and S. Ullman, '*Face recognition: The problem of compensating for changes in illumination direction*', IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 721-732, 1997.

[16] G. L. Marcialis and F. Roli, "*Fusion of LDA and PCA for face verification*," in Biometric Authentication, Vol. 2359, pp. 30-37, 2002

[17] B. Schölkopf and K. R. Müller, '*Nonlinear component analysis as a kernel eigenvalue problem*', Neural Computation, pp 1299-1319, 1998.

[18] F.R. Bach and M.I.Jordan, '*Kernel Independent Component Analysis*', Journal of Machine Learning Research, Vol. 3, pp 1-48, 2002.

[19] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K. R. Müller, '*Fisher discriminant analysis with kernels*', Neural Networks for Signal Processing IX, Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp 41–48, 1999.

[20] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, '*Face Recognition by Independent Component Analysis*', IEEE Trans. on Neural Networks, Vol. 13, pp. 1450-1464, 2002.

[21] C. Liu and H. Wechsler, '*Comparative Assessment of Independent Component Analysis (ICA) for Face Recognition*', Second International Conference on Audio and Video-based Biometric Person Authentication, AVBPA'99,Washington D. C., USA, March 22-24, 1999.

[22] K. Baek, B. Draper, J.R. Beveridge, and K. She, '*PCA vs. ICA: A Comparison on the FERET Data Set*', Proc. of the Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing, pp 824-827, 2002.

[23] B. Moghaddam, '*Principal Manifolds and Probabilistic Subspaces for Visual Recognition*', IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, pp. 780-788, 2002.

[24] J.R. Beveridge, K. She, B. Draper, and G.H. Givens, '*A Nonparametric Statistical Comparison of Principal Component and Linear Discriminant Subspaces for Face Recognition*', Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 535- 542, 2001

[25] V. Belhumeur, J. Hespanha, and D. Kriegman. '*Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection*', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19 pp 711-720, 1997.

[26] P. Navarrete and J. Ruiz-del-Solar, '*Analysis and Comparison of Eigenspace-Based Face Recognition Approaches*', International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, pp. 817-830, 2002.

[27] A. K. Jain, X. Lu, and Y. Wang, 'Combining Classifiers for Face Recognition', In Proc. of IEEE International Conference on Multimedia and Expo, pp. 13-16, Baltimore, MD, 2003.

[28] P.J. Phillips, H. Moon, S.A. Rizvi and P.J. Rauss, '*The FERET evaluation methodology for Face Recognition algorithms*', IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, pp. 1090-1104, 2000

[29] J. Kittler, M. Hatef, R. Duin, and J. Matas, '*On combining classifiers*', IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, pp. 226-239, 1998.

[30] K. Delac, M. Grgic and P. Liatis, '*Appearance-based Statistical Methods for Face Recognition*', 47th International Symposium ELMAR, Zadar, Croatia, 8-10 June 2005.

[31] S. Gong, S.J. McKenna and A. Psarrou, '*Dynamic vision: From images to face recognition*', Imperial College Press, London, 2000.

[32] H.K. Ekenel and B. Sankur, '*Feature selection in the independent component subspace for face recognition*', Pattern Recognition Letters, Vol. 25, pp 1377-1388, 2004.

[33] D.I. Domboulas, '*Infrared Imaging Face Recognition using Nonlinear Kernel-based Classifiers*', Master of Science Graduate Thesis, Naval Postgraduate School, Monterey, California, 2004.

[34] M. Kirby and L. Sirovich, '*Application of the Karhunen -Loeve Procedure for the Characterization of Human Faces*', IEEE Trans. PAMI, Vol. 12, pp 103-108, 1990.

[35] M. Turk and A. Pentland, '*Eigenfaces for Recognition*', Journal of Cognitive Neuroscience, Vol. 3, pp. 71-86, 1991.

[36] A. Pentland, B. Moghaddam and T. Starner, '*View-Based and Modular Eigenspaces for Face Recognition*', In *IEEE* Conference on Computer Vision and Pattern Recognition, Seattle, pp 84-91, 1994.

[37] G.D. Finlayson, J. Dueck, B. V. Funt, M.S. Drew, '*Colour Eigenfaces*', Proc. Third International Workshop on Image and Signal Processing Advances in Computational Intelligence, Manchester, UK, 4-7 November 1996.

[38] B. Moghaddam and A. Pentland, '*Beyond Euclidean Eigenspaces: Bayesian Matching for Visual Recognition*', Face Recognition: From Theories to Applications, Springer Verlag, 1998, ISBN 3-540-64410-5.

[39] A. Selinger and D.A. Socolinsky, '*Appearance-Based Facial Recognition Using Visible and Thermal Imagery: A comparative Study*', Equinox Corporation, 2002.

[40] R. A. Fisher, '*The Use of Multiple Measures in Taxonomic Problems*', Ann. Eugenics, Vol. 7, pp. 179-188, 1936.

[41] R. Brunelli and D. Falavigna, '*Person identification using multiple cues*', IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17, pp 955-966, 1995.

[42] L. Hong and A.K Jain, '*Integrating faces and fingerprint for personal identification*', IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, pp 1295-1307, 1998.

[43] C. Jutten and J. Herault, '*Blind separation of sources in an adaptive algorithm based on neuromimetic architecture*', Signal Processing, Vol. 24, pp. 1-10, 1991.

[44] P. Comon, '*Independent component analysis - A new concept?*', Signal Processing, Vol. 36, pp. 287-314, 1994.

[45] A. J. Bell and T. J. Sejnowski, '*An information-maximization approach to blind separation and blind deconvolution*', Neural Computing, Vol. 7, pp. 1129-1159, 1995.

[46] M. Zibulevsky and B. A. Pearlmutter, '*Blind Separation of Sources with Sparse Representations in a Given Signal Dictionary*', International Workshop on Independent Component Analysis and Blind Signal Separation, Neural Computation, Vol. 13, pp 863-882, 2000.

[47] A. Hyvarinen, '*The Fixed-point Algorithm and Maximum Likelihood Estimation for Independent Component Analysis*', Neural Processing Letters, Vol. 10, pp. 1-5, 1999.

[48] S. Z. Li, X. Lv, and H. Zhang, '*View-subspace analysis of multi-view face patterns*', In Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, Canada, July 13 2001.

[49] A. Ross and A. Jain, '*Information fusion in biometrics*', Pattern Recognition Letters, Vol. 24, pp 2115-2125, 2003.

[50] M. Yang, '*Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods*', In Proc. 5th IEEE International Conference on Automatic and Gesture Recognition, pp 215-220, 2002.

[51] M.A. Aizerman, E.M. Braverman and L.I. Rosonoer, '*Theoretical foundations of the potential function method in pattern recognition learning*', Automation and Remote Control, Vol. 25, pp. 821–837, 1964.

[52] V. Vapnik, '*The Nature Of Statistical Learning Theory*', Springer-Verlag, 1995. ISBN 0387987800.

[53] G. Baudat and F. Anouar, '*Generalized discriminant analysis using a kernel approach*', Neural Computation, Vol. 12, pp 2385-2404, 2000.

[54] M. Welling, '*Fisher Linear Discriminant Analysis*', Department of Computer Science, University of Toronto, 2005. http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf

[55] M. Yang, '*Face Recognition using Kernel Methods*', Fifth IEEE International Conference on Automatic Face and Gesture Recognition, p. 0215, 2002.

[56] A. R. Martinez and R. Benavente, 'The AR face database. Technical Report', Computer Vision Centre Technical Report, Barcelona, Spain, 1998. http://rvl1.ecn.purdue.edu/ARdatabase

[57] V. Franc and V. Hlaváč, '*Statistical Pattern Recognition Toolbox for Matlab*', 2004. http://cmp.felk.cvut.cz/cmp/cmp software.html

[58] H. Gävert, J. Hurri, J. Särelä, and A. Hyvärinen, '*FastICA Toolbox for Matlab7.x and 6.x*', 2005. http://www.cis.hut.fi/projects/ica/fastica/code/dlcode.html

[59] P. Eisert and B. Girod, '*Illumination Compensated Motion Estimation for Analysis Synthesis Coding*', 3D Image Analysis and Synthesis, pp 61-66, 1996.

[60] S. Malassiotis and M.G. Strintzis, '*Robust Face Recognition using 2D and 3D Data: Pose and Illumination Compensation*', Informatics and Telematics Institute, Greece, 2004. www.iti.gr/files/pr2005face.pdf

[61] H. Moon and J. Phillips, '*Analysis of PCA-based Face Recognition Algorithms*', in Empirical Evaluation Techniques in Computer Vision, IEEE Computer Society Press, pp 57-71, 1998.

[62] R.O. Duda, P.E. Hart and D.G. Stork, '*Pattern Classification*', 2nd Edition, Wiley Interscience, New York, 2001.

[63] P.J. Philips, P.J. Rauss and S.Z. Der, '*FERET Recognition Algorithm Development and Test Results*', Technical Report ARL–TR–995, Adelphi, Maryland, 1996.

[64] W. Yambor, B. Draper and R. Beveridge, '*Analyzing PCA-Based Face Recognition Algorithms: Eigenvector Selection and Distance Measures*', Empirical Evaluation Methods in Computer Vision, Singapore: World Scientific Press, 2002.

[65] G. Givens, R. Beveridge, B. Draper and D. Bolme, '*Using a Generalized Linear Mixed Model to Study the Configuration Space of a PCA+LDA Human Face Recognition Algorithm*', Technical Report, Computer Science, Colorado State University, 2003, ISBN 978-3-540-22958-2.

[66] M. Grgic, K. Delac, S. Grgic, '*Face Recognition: Hypothesis Testing across all Ranks*', Technical Report: FER-VCL-TR-2005-02, University of Zagreb, Croatia, 2005.

[67] A. K. Jain, S. Prabhakar, and S. Chen, '*Combining multiple matchers for a high security fingerprint verification system*', Pattern Recognition Letters, vol. 20, pp. 1371- 1379, 1999.

[68] V. Franc and V. Hlaváč, '*Statistical Pattern Recognition Toolbox for Matlab – User Guide*', 2004. ftp://cmp.felk.cvut.cz/pub/cmp/articles/Franc-TR-2004-08.pdf

[69] A. Godil, S. Ressler, P. Grother, '*Face Recognition using 3D Facial Shape and Colour Map Information: Comparison and Combination*', National Institute of Standards and Technology, Gaithersburg, 2004.

[70] L. Breiman, '*Bagging Predictors*', Machine Learning, Vol. 26, pp. 123-140, 1996.

[71] Y. Freund, R.E. Shaphire, '*A Decision-theoretic generalization of on-line learning and an application to boosting*', Journal of Computer and System Sciences, Vol. 55, pp. 119-139, 1995.

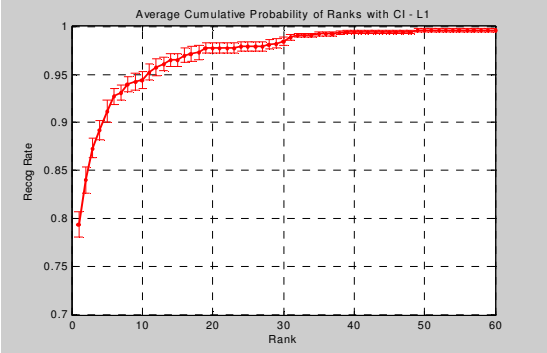
[72] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, '*The FERET Evaluation Methodology for Face-Recognition Algorithms*,' IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp.1090-1104, 2000.

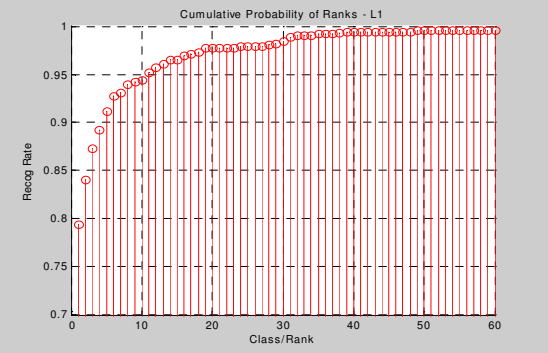
# APPENDIX A

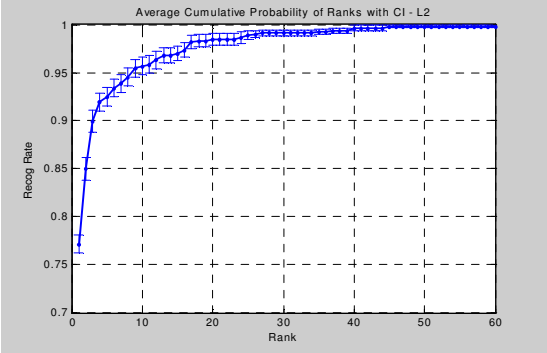
## Sample of Graphs found in Appendix A of attached CD

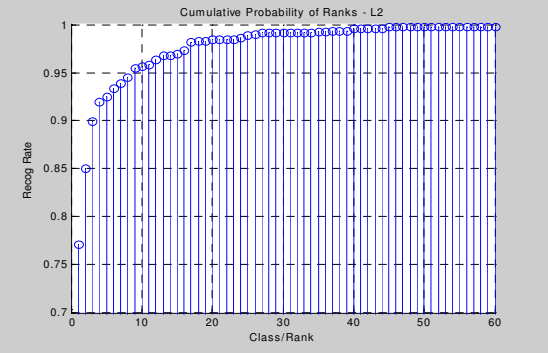
## *Expression:* PCA - Linear

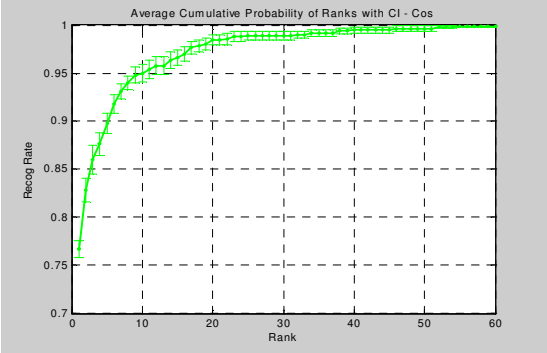
Confidence Interval and Stem plots

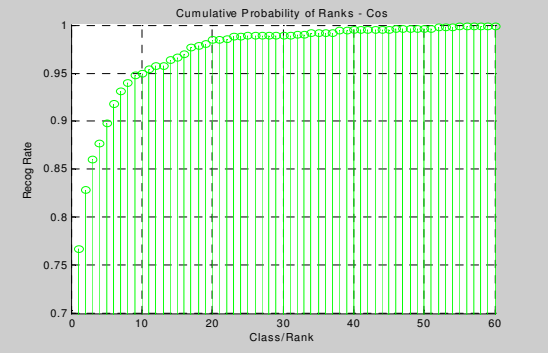


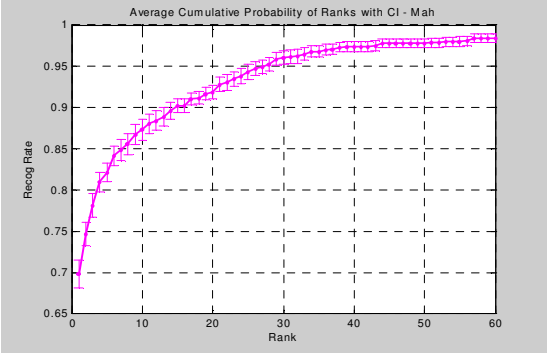



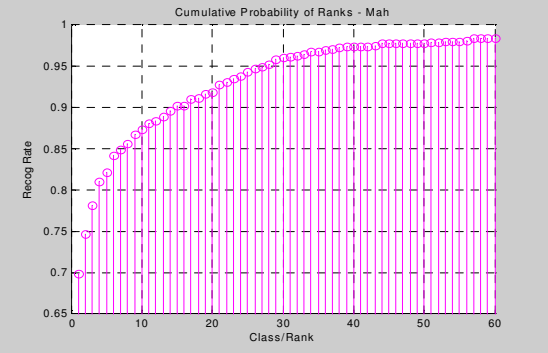



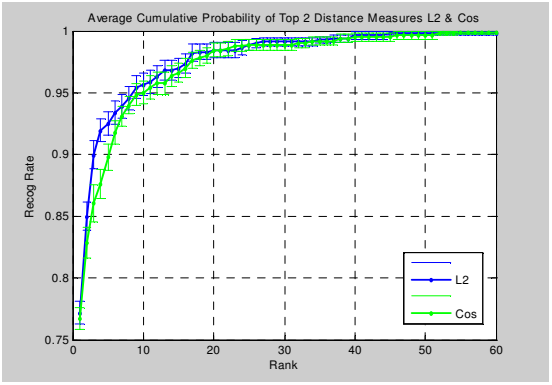



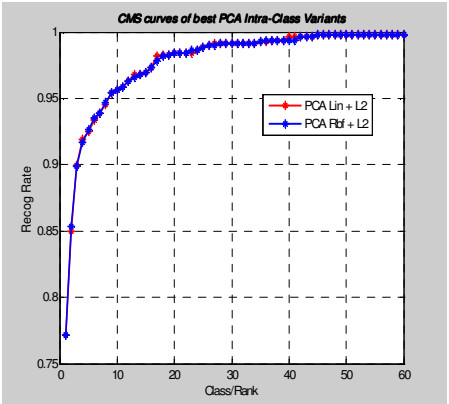



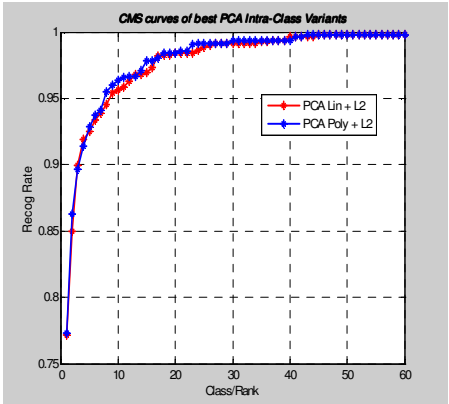



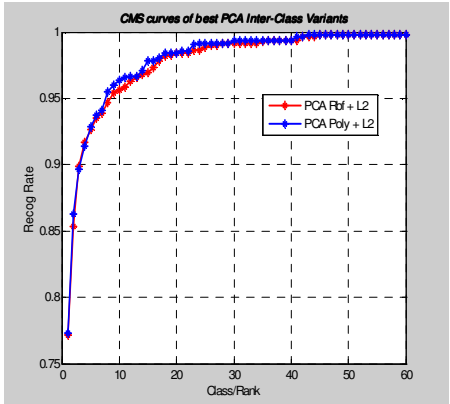



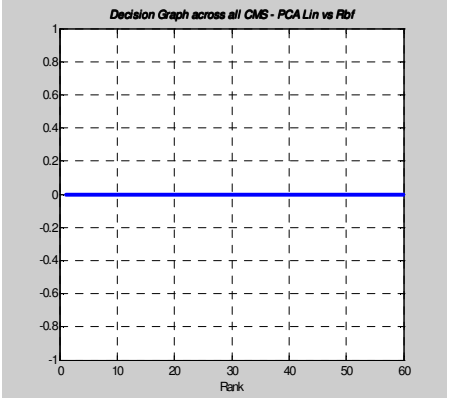


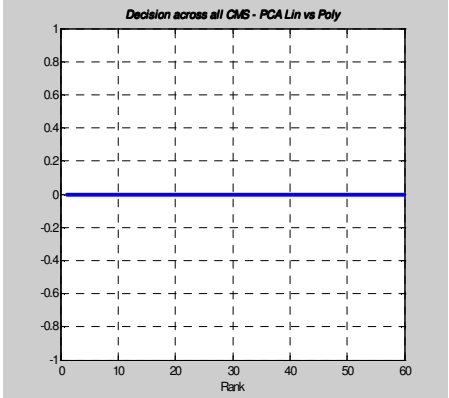

McNemar's Testing for the PCA class in the category of Expression

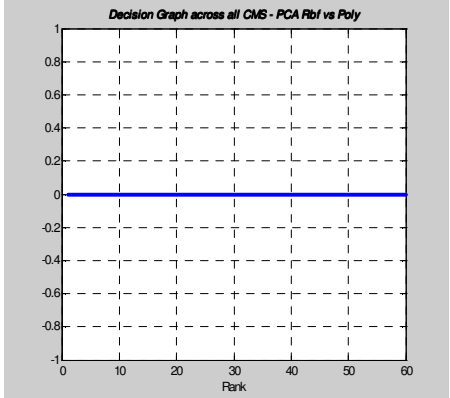


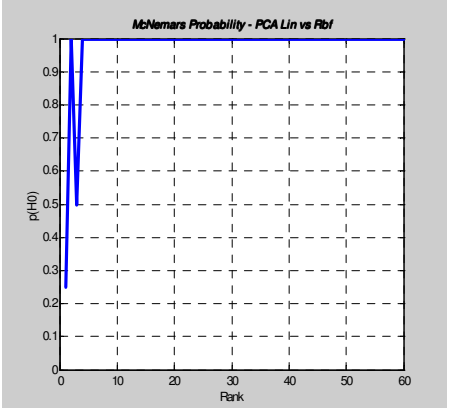



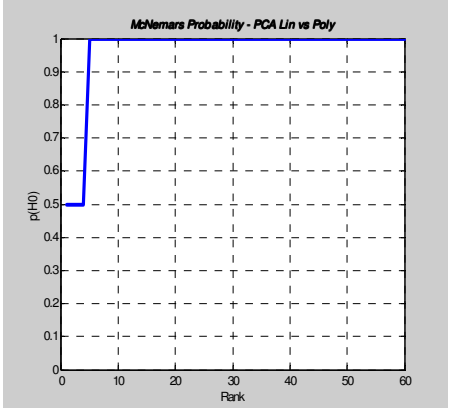



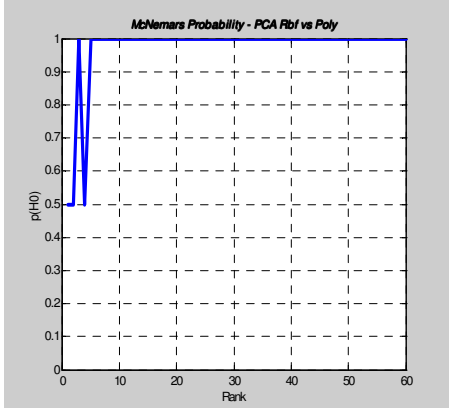



# APPENDIX B

# *Summary of Matlab Modules (m files) found in Appendix B of CD*

## Pre-processing

*Acc_vs_Dim* - Script file employed to obtain rank-1 Accuracy and Dimensionality information, used in conjunction with *Eigen_plot* in verifying the FERET Heuristic for 96% of energy retention.

*Add0* - Automated image ordering and renaming module.

*Alignment* - Publicly available GUI interface foe image resizing and alignment. See www.mathworks.com

*Div_Select_Auto* - Script file for separating images into various training and probe sets prior to any processing.

*Eigen_cont* - Eigen/Energy contribution versus subspace dimensionality.

*Eigen_plot* - Script file for energy contribution.

*Errorbar_mine* - Edited version of Matlab's Errorbar function. Used in creating Confidence Intervals.

*Input_Create* - Input Matrix creation for concatenating training and testing images for use in Matlab environment.

*Preprop_2_hist* - Greyscale conversion, Normalisation and Histogram Equalisation.

*Proj_norm* - Normalisation module.

*Rand_Select* - Random selection and creation of training and testing sets, used for verifying and establishing testing methodology.

*Test_Select_2* - Script file for probe set creation.

## Implementation and Testing

*Ave_Dist* - Function for obtaining average distance matrix over 10 probe tests.

*Class_Rank* - Class Classification and Cumulative Rank performance levels.

*Confid_graph* - Function for plotting CMS curves with Confidence Intervals.

*Confid_graph_t2* - Top 2 CMS metric plots with Confidence Intervals.

*Hyp_Test* - McNemar's Hypothesis Testing.

*Rank_All_graph* - CMS charts for all the distance metrics.

*Stem_graph* - Stem plot of the CMS charts.

*Subfigure* - Useful, publicly available function for plotting multiple images in a single figure. See www.mathworks.com

*TestAll* - Main script file for testing each algorithm with every probe set. Yields the similarity measures, averaged accuracies, CMS charts and confidence levels.

### Subspace Algorithms

*ICA_fast_A1*
*ICA_fast_A2*
*ICA_info_A1*
*ICA_info_A2*
*RunLDA*
*RunPCA*

### Toolbox Variations

*fastica_mine* - Adaptation from ref [58]
*kpca_mine* - Kernel PCA algorithm ref [57]
*runica_mine_fast* - Adaptation from ref [20]
*runica_mine_info* - Adaptation from ref [20]

## Hybrid formulation

*Ave_Dist_hybrid* - Hybrid implementation for average distance matrix over 10 probe tests.

*Hybrid_2* - Main function for hybrid formulation, takes similarity measures of component algorithms as inputs and outputs Hybrid comparative results to these components.

# APPENDIX C