To be submitted for the 2007 International Joint Conference on Neural Networks

Independent Evaluation of Subspace Face Recognition Algorithms

Dhiresh R. Surajpal and Tshilidzi Marwala
School of Electrical & Information Engineering
University of Witwatersrand
South Africa
d.surajpal@ee.wits.ac.za

Abstract – **This investigation explores a comparative study of both the linear and kernel implementations of three of the most popular Appearance-based Face Recognition projection classes. These are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA). The experimental procedure provides a platform of equal working conditions and examines 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.**

Index Terms – CMS, Hybrid, McNemar, Rank, Subspace

I. Introduction

A human face is an extremely complex, dynamic and deformable 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 in response to the effects of emotion, temperature, reflectance properties and perspiration levels, thus creating a large variety and variability within the configurations of facial expression. Another avenue includes temporal 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 and hairstyles [2].

Much of the world's best commercial systems provide real time solutions for face detection, image registration, and image matching [5]. 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 [5], many of the computer vision systems reported in literature still employ the popular appearance-based paradigm for object recognition [6]. Appearance-based analysis, which is one of the oldest approaches, is still said to give the most promising results [4]. Among the most popular publicly available subspace approaches are the classes of Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA). 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 X is the matrix of input images).

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*. [12] and Liu *et al.* [13], for example, claim that ICA outperforms PCA, while Baek *et al*. [14] claim that PCA is better. Moghaddam [15] states that there is no significant difference. Beveridge et al. [16] claim that in their tests LDA performed uniformly worse than PCA, Martinez [6] states that LDA is better for some tasks, and Belhumeur *et al*. [17] and Navarrete *et al.* [18] 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-standardized 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 facial recognition (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 realize an independent comparative study that will greatly contribute to previous literary works. 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 most popular and 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 [27] and will be compared across all the techniques to conclusively yield the 'best' algorithm in each category.

While it may also 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 [19]. Each individual classifier has a different sensitivity to different changes in facial variation and as was reported by

Phillips *et al* [20], appearance-based methods show different levels of performance for different subsets of images. In their analysis of ICA and PCA, Bartlett *et al* [12] also reported that when incorrect classifications were made, it was very rare that the algorithms assigned the same incorrect identity class. The above findings strongly suggest that different classifiers contribute differently and hence offer complementary information about the classification task. A *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 this investigation will also propose a hybrid formulation that combines the 'best' algorithm from each category. Following Kittler's theoretical framework on combining classifiers [21], 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.

II. BACKGROUND

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* ($R^{N=m\ x\ n}$). This image space, however, constitutes a rather high-dimensional space and recognition therein would be deemed computationally infeasible [14,72]. 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* [7].

Theoretically it is common to model this subspace as a lower-dimensional *principle manifold*, embedded in a higher dimensional image space [14,72], wherein the *intrinsic* dimensionality is determined by the number of degrees of freedom within the face space. Gong *et al* [32] 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 *principle modes* of the underlying manifold, while retaining as much information (energy) from the original images as possible [4]. By doing this, subspace methodologies ensure that computational efficiency and hence the successful viability of face recognition algorithms can be achieved [31].

A. Principal Component Analysis

The Principal Component Analysis (PCA) procedure follows the description by Pentland and Turk as described in [33]. Given an s-dimensional vector representation of each face in a training set of M images, PCA tends to find a t-dimensional subspace whose basis vectors correspond to the maximum variance direction in the original image space. This new subspace is normally lower dimensional (t << s). All images of known faces are projected onto the face space to find sets of weights that describe the contribution of each vector. To identify an unknown image, that image is projected onto the face space as well to obtain its set of weights. By comparing a set of weights for the unknown face to sets of weights of known faces, the face can be identified. If the image elements are considered as random variables, the PCA basis vectors are defined as eigenvectors of the scatter matrix S defined as:

$$S = \sum_{i=1}^{M} (x_i - \mu).(x_i - \mu)^T \tag{1}$$

where μ is the mean of all images in the training set (the mean face) and x_i is the i^{th} image with its columns concatenated in a vector. The projection matrix W_{PCA} is composed of t eigenvectors corresponding to t largest eigenvalues, thus creating a t-dimensional face space.

B. Independent Component Analysis

PCA considered image elements as random variables with Gaussian distribution and minimized second-order statistics. Clearly, for any non-Gaussian distribution, largest variances would not correspond to PCA basis vectors. Independent Component Analysis (ICA) [12] minimizes both second-order and higher-order dependencies in the input data and attempts to find the basis along which the data (when projected onto them) are – *statistically independent*. Bartlett et al. [12] provided two architectures of ICA for face recognition task: *Architecture I* (ICA1) - statistically independent basis images and *Architecture II* (ICA2) - factorial code representation. These architectures are used in combination with the two currently popular implementations of ICA in the form of Bell and Sejnowski's *InfoMax* algorithm [34] and Hyvarinen's *FastICA* approach [35].

C. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [17] finds the vectors in the underlying space that best discriminate among classes. For all samples of all classes the between-class scatter matrix S_B and the within-class scatter matrix S_W are defined by:

$$S_B = \sum_{i=1}^{c} N_i (x_i - \mu)(x_i - \mu)^T \tag{2}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_m \in X_i} (x_m - \mu_i)(x_m - \mu_i)^T \tag{3}$$

where N_i is the number of training samples in class i, c is the number of distinct classes, μ_i is the mean vector of samples belonging to class i and X_i represents the set of samples belonging to class i with x_m being the m^{th} image of that class. S_W represents the scatter of features around the mean of each face class and S_B represents the scatter of features around the overall mean for all face classes.

The goal is to maximize S_B while minimizing S_W, in other words, maximize the ratio $\det|S_B| / \det|S_W|$. This ratio is maximized when the column vectors of the projection matrix (W_{LDA}) are the eigenvectors of $S_W^{-1} \cdot S_B$. In order to prevent S_W from becoming singular, PCA is used as a preprocessing step and the final transformation is thus $W_{opt}^T = W_{LDA}^T W_{PCA}^T$.

D. Kernel Methods

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 [8]. From a classification viewpoint linear approaches, which only describe information based on second order statistics [1], 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 [24].

First introduced by Aizerman *et al* [25], 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]. It was not until recently that Schölkopf *et al* [9] extended the classical PCA algorithm to Kernel Principal Component Analysis (KPCA) that was shown to be able to extract nonlinear 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 [17]. Mika *et al* [11] and Baudat and Anouar [26], then proposed Kernel Linear Discriminant Analysis (KLDA), whose results are claimed to be superior to that of PCA, LDA, ICA and KPCA [7]. Kernel ICA was also introduced by Bach and Jordan [10] 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 [10] and therefore placing its evaluation beyond the scope of this investigation.

E. Hybrid Methods

In classifier fusion, the outputs of individual classifiers are combined by a second classifier according to a pre-defined combination rule. Classifier combination can essentially be implemented at three levels [23]: Fusion at the a) *Feature Extraction level* b) *Confidence or Matching score level* and c) *Decision level.*

The use of classifier fusion has produced many combination techniques over the years. One popular approach has been the idea of bagging [29], which manipulates the training-data with sub-sampling. Another common algorithm, boosting [30], 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* [21] 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*. The experimental results showed that the best performances were obtained when using the sum and median rules. Ross and Jain [23] also 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 [23]:

- Non-homogenous score types, whereby different methods are employed in obtaining these values such as distances and similarity measures (example 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* [23]. 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 [23].

III. Implementation

In refining the nature of the tasks being evaluated one established five main areas of focus [3]: *Viewing Angle, Illumination, Expression, Occlusion* and *Time Delay*. The AR Database [27] provided the most efficient publicly available source of subject samples, which overlapped most of the desired categories and effectively facilitated testing. The database contains over 3000 images of 135 individuals. The subjects were recorded over two sessions with a two week interval between shoots. During each session 13 images per subject under varying conditions, 1 neutral and 3 per variant of expression, illumination and occlusion (with lighting changes).

Each image was resized to dimensions of (125 x 165 x 3) and pre-processed using the traditional procedure depicted in figure 1. The pre-processed database formulated a baseline system [5], which was then divided into gallery and probe sets. 110 subjects (55 males and 55 females) were randomly selected and since it is desirable to have no overlap between the training and testing images, subject images were divided 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, illumination and occlusion over time given that few training samples may be available.

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). 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. 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.

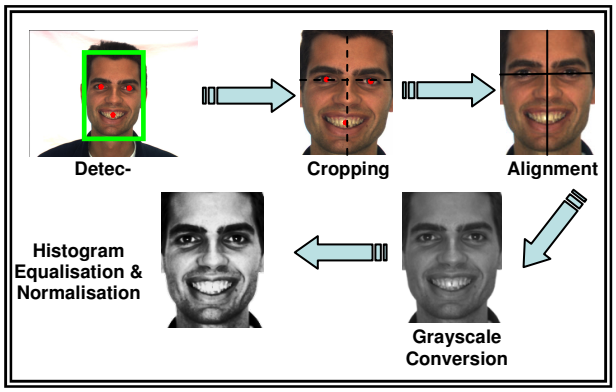

Figure 1 Depiction of Pre-processing Procedure

Training each algorithm followed the procedure in figure 2, whereby all the training images are arranged into a column-wise input matrix and sequentially projected into the subspace to yield the projection database. 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*'. Having no closed form expression to explicitly determine this region, one adopted the FERET heuristic that suggests selecting 40% of the dimensionality would retain approximately 96% of the energy spectrum (signal information) [5,12,32]. This was applied to the AR Database where it was found that optimal dimensional retention for PCA and ICA lied between 116 – 199 and between 72 – 74 for LDA.

The testing procedure, using the nearest neighbor approach, facilitates each probe image to be classified with a class label, the results of which are then put forth to the evaluation phase, where the comparative assessment will be 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 [22], offering intuitive insight into which algorithm performance throughout the rank spectrum. Finally statistical measures are also applied in the form of McNemar's Hypothesis Protocol [28] that offers the practical insight pertaining to *what point does the difference in performance results actually become significant.*

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 [26] and Franc [68], which proved that Kernel implementations employing linear functions (1st 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. Looking at the class of ICA, the Kernel implementation is not yet feasible for large input spaces; the results will therefore be evaluated demonstrating the algorithms of InfoMax and FastICA in combination with both Architectures I and II.

Upon establishing the 'best' projection-metric combinations for each evaluation category, these approaches will then be combined to realize a fusion strategy that will effectively demonstrate the advantageous results of hybrid formulations. Ross and Jain [23] reported their conclusive findings which indicate that the approach of classification before fusion actually performs poorer than the confidence score level.

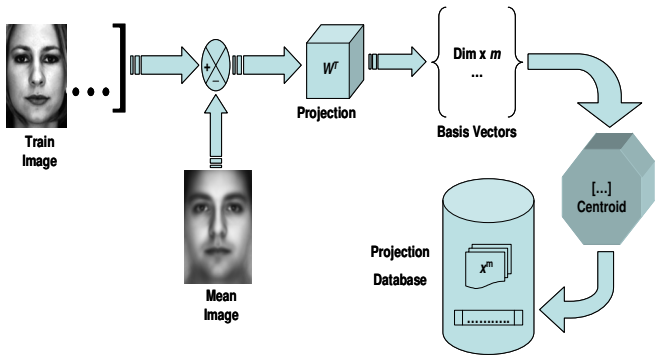

Figure 2 Generic Training Model

The matching score level was found to offer the best tradeoffs in terms of information content and ease of fusion [23]. 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. Normalisation of the differing metric measures are performed employing the MinMax scheme, shown below, resulting in a common range of [0 100].

$$S' = \frac{S - \min\{S_k\}}{\max\{S_k\} - \min\{S_k\}} \quad \text{for } k = 1,2,\ldots,P \qquad (4)$$

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. The combined matching score will be calculated as follows:

$$S_{comb} = w_1 S'_{FA1} + w_2 S'_{FA2} + w_3 S'_{LDA} \qquad (5)$$

The weights associated with each classifier's matching score are defined by a confidence function to represent the relative contribution 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. If LDA, for example, performs well in the categories of Expression and Time Delay, it will obtain two fifths or 40% of the weighting; similarly if FA1 outperformed the other algorithms in both occlusion categories it will also receive 40% of the weighting; with the remaining 20% going to the algorithm performing best in the category of illumination.

Method 2: 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:

$$r_{sum} = \sum_{i=1}^{q} r_i \tag{6}$$

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

$$\sum_{i=1}^{q} w_i = \sum_{i=1}^{q} \frac{r_i}{r_{sum}} = 1 \tag{7}$$

Once the hybrid matching score has been calculated, final classification and evaluation is performed.

IV. Results and Analysis

Using the described methodology the relative performance levels of the three most popular classes of appearance based subspace methodologies i.e. PCA, LDA and ICA were investigated. In conjunction with the assessment process one also sort to find the best metric combination that offered the best task specific advantage. 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 offered a better overall performance.

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, rank-1 results shown in table 1, were carried out in much the same fashion as the intra-class tests. 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 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 class.

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

Expression: There is no statistical difference between any of the variants, so 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.

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.

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*.

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.

Table 1 Inter-class Rank-1 results

Inter-class Rank 1 Results					CMS Results	
	L1	***L2***	***Cos***	***Mah***	***Highest Curve***	***Same rank1***
EXPRESSION						
PCA – Poly	***79.73%***	77.32%	77.55%	77.46%	L2	No
LDA – Poly	82.18%	81.77%	81.96%	***83.25%***	Mah	Yes
ICA – FA2	70.14%	72.46%	***80.05%***	72.46%	Cos	Yes
ILLUMINATION						
PCA – Poly	***57.09%***	52.86%	52.46%	55.52%	L1	Yes
LDA – Poly	71.18%	72.50%	70.86%	***75.18%***	Mah	Yes
ICA – FA1	71.46%	74.59%	***84.14%***	74.59%	Cos	Yes
LOWER OCC.						
PCA – Linear	2.91%	3.96%	3.77%	***7.23%***	Mah	Yes
LDA – Linear	14.96%	14.96%	***16.46%***	9.77%	Cos	Yes
ICA – FA1	9.50%	10.18%	***27.86%***	10.18%	Cos	Yes
UPPER OCC.						
PCA – Poly	31.46%	***32.59%***	28.96%	29.82%	L2	Yes
LDA – Poly	22.91%	21.36%	21.32%	***27.91%***	Mah	Yes
ICA – FA1	34.36%	31.36%	***51.59%***	31.36%	Cos	Yes
TIME DELAY						
PCA – Poly	78.18%	74.55%	73.64%	***80.00%***	L1	No
LDA – Poly	90.00%	88.18%	86.36%	***90.91%***	Mah	Yes
ICA – FA1	87.27%	85.46%	***90.00%***	85.46%	Cos	Yes

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.

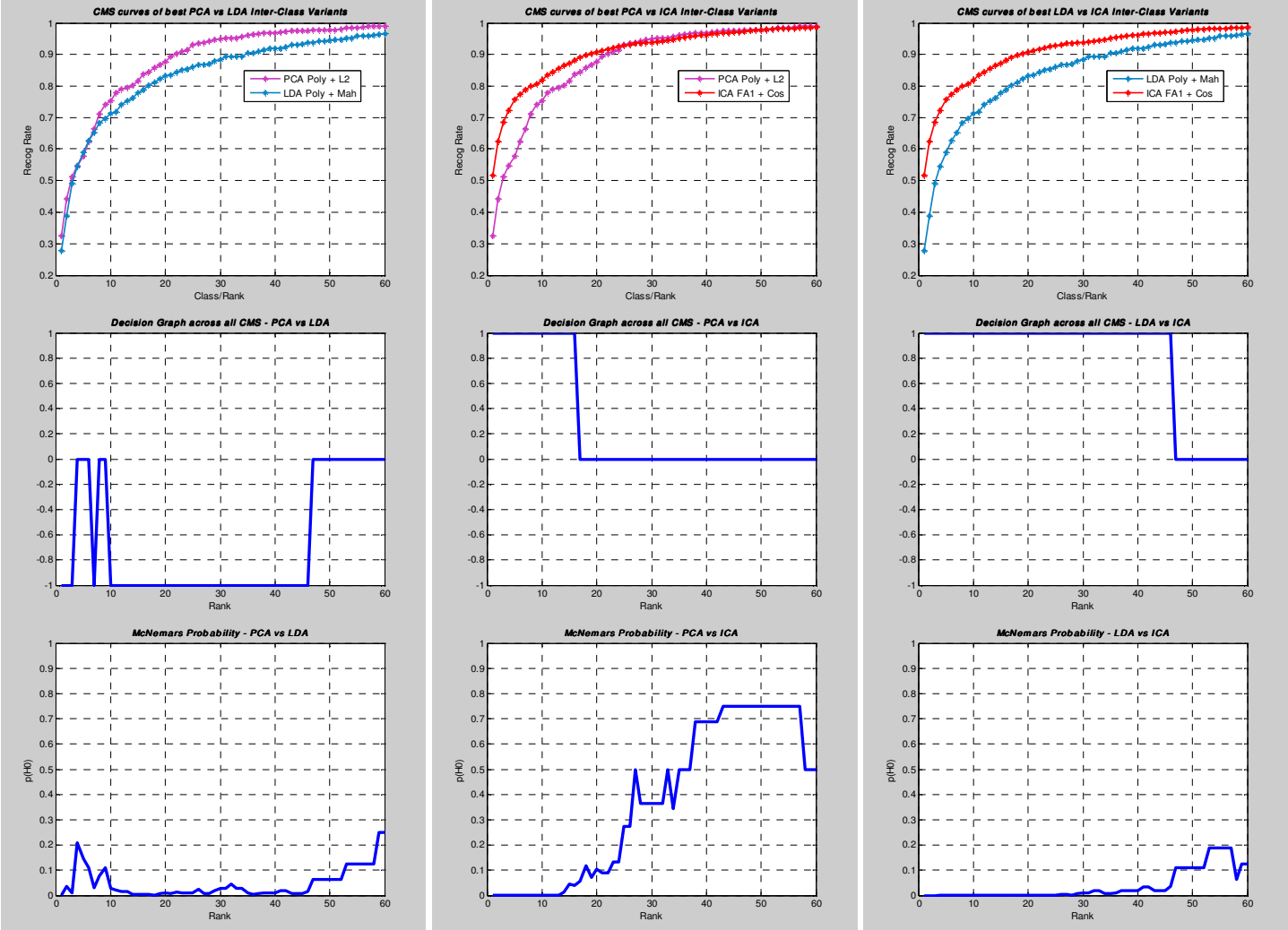

Figure 3 Example of McNemar's Test for the top 3 inter-class algorithms – Upper Occlusion
Column 1: PCA Poly(L2) vs LDA Poly(Mah), Column 2: PCA Poly(L2) vs ICA FA1(Cos), Column 3: LDA Poly(Mah) vs ICA FA1(Cos)

In 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.

Two classes of subspace methods stood out, those being the classes of ICA and LDA. The hybrid formulation will thus seek to combine and exploit the powerful data representation of ICA and the unique class discriminality 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.

Looking at table 2 and figure 4, 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.

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. 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.

	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 2 Rank-1 Comparative Hybrid Results

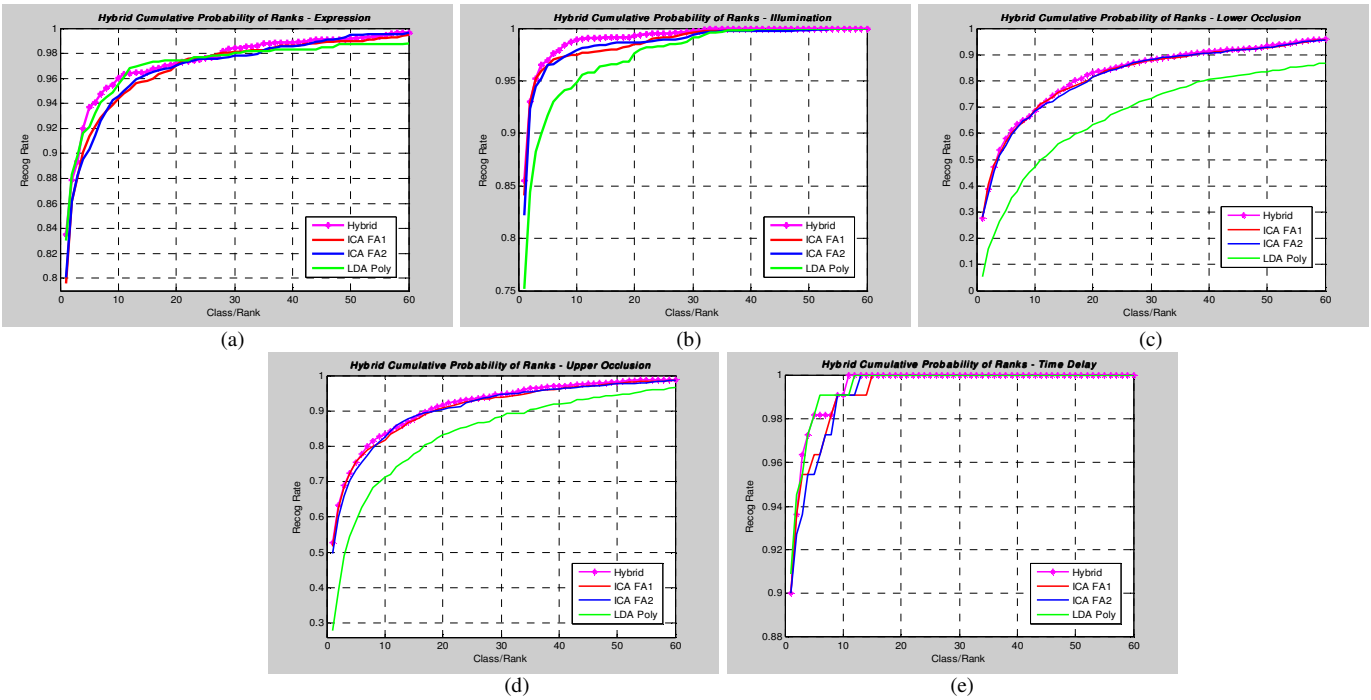

Figure 4 Hybrid CMS Charts
(a) Expression (b) Illumination (c) Lower Occlusion (d) Upper Occlusion (e) Time Delay

V. Conclusion

This research investigation presented a 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 evaluated the linear and kernel variants of the respective classes and provided both intuitive and statistical evidence into their comparative standings. This work 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. 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 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.

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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

[17] 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.

[18] 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.

[19] 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.

[20] 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

[21] 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.

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

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

[24] 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.

[25] 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.

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

[27] 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

[28] 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.

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

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

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

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

[33] M. Turk and A. Pentland, '*Eigenfaces for Recognition*', Vision and Modeling Group, Massachusetts Institute of Technology, 2004.

[34] 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.

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