# Integrating Monolithic and Free-parts Representations for Improved Face Verification in the Presence of Pose Mismatch

Simon Lucey Tsuhan Chen

Advanced Multimedia Processing Laboratory, Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA 15213, USA
slucey@ieee.org, tsuhan@cmu.edu

## Abstract

*Face images, varying under pose, are dramatically different in their "pixel" appearance even if they stem from the same subject. Our work concentrates specifically on the task of verifying faces when the gallery set stems from frontal face images, with the probe set stemming from a number of alternate poses (i.e. pose mismatch). An argument is put forward for attempting to recognize faces through integrating holistic/monolithic and free-parts representations. Canonical monolithic representations are investigated such as Eigenface and Fisherface techniques, as well as recent techniques that are able to deal with pose specifically, such as Eigen-light fields. Similarly, parts representations are investigated, with particular attention being paid to Free-Parts Gaussian Mixture Models (FP-GMMs) as a useful representation. A contribution is made via the analysis of what traits, in a face, are most useful for each representation. Finally, we are able to demonstrate that there is: a) benefit in combining free-parts and monolithic representations, and b) further benefit can be obtained by varying the weight placed on each representation as a function of view point.*

## 1. Introduction

Face verification with a change in viewpoint, between 2D gallery and 2D probe images, is inherently a difficult task (i.e. pose mismatch). Images taken of the face from one pose, for the same subject, are markedly different to images captured under another pose. One can tell from visual inspection that pixel variation due to pose change is far greater than the variation seen due to changes in identity. An example of this problem can be seen in Figure 1. In this paper we will be dealing specifically with the problem of trying to verify clients from non-frontal viewpoint probe images given that only a single frontal view image of that client exists in the gallery.

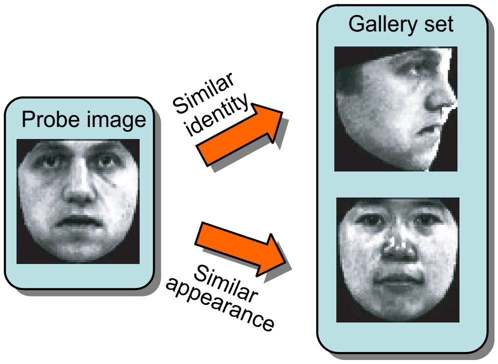

Figure 1: Example of the difficulty in recognizing subjects from different pose as images from the same pose, irrespective of identity, are more similar in terms of their pixel representation.

In cognitive science, theories abound over whether humans recognize faces based on component parts or holistic representations. In fact there is a large amount of literature [1, 2] indicating that both types of representations of the face are important in human face recognition in the presence of pose mismatch. We use the term *monolithic* in this paper to describe the holistic vectorized representation of the face based purely on pixel values within an image array, which can be associated with the holistic mechanism used in a human face recognition system. Similarly, we use the term *parts* to denote a representation of the face that can be considered as an ensemble of image patches of the image array. The employment of parts representations for object/face detection has recently gained much attention and success in machine vision literature [3, 4, 5]. For the task of face recognition we additionally categorize parts representations into two subsets namely *rigid*- and *free*- parts. Rigid-parts representations assume the weight/contribution of each patch in the image, towards the final recognition decision, is not homogeneous; but the position/structure of the patches within the image is preserved. Free-parts representations assume that the position/structure of patches within the image can be relaxed so they can "freely" move to varying extents. Both rigid- and free- parts representations assume there is minimal dependence between the appearance of other patches within the image.

Considerable work has already been performed with monolithic face representations, for automatic face recognition, in the presence of pose mismatch. Most notably techniques like Tensorfaces [6], Eigen-light fields [7] and Fisherfaces [8] have been employed with varying degrees of success. There has already been some preliminary work by Kanade and Yamada [9] demonstrating the benefit of a rigid-parts representation, where weightings for each patch in the face are learnt off-line, from a development set, as a function of pose. Monolithic and rigid-parts approaches are very similar in terms of the classification mechanism they employ as they both essentially compare gallery and probe "points" in a feature space; with rigid-parts approaches employing multiple points, with each point existing in separate and largely independent feature spaces. Hitherto, the benefit of employing a free-parts representation has not been fully investigated for the task of automated face verification in the presence of pose mismatch. Free-parts representations have an inherent advantage over monolithic and rigid-parts representations in that they compare "distributions" which are naturally able to cope with appearance variation. In this paper we will be focussing on comparing free-parts and monolithic representations as they are representative of "point" and "distribution" style classification mechanisms for verification.

Recent work [10, 11, 12] has been conducted demonstrating that good performance can be attained by employing a free-parts representation in the task of frontal view face verification. Some generative models that have been previously employed to model these free-parts face distributions are: pseudo 2-D hidden Markov models (HMMs) [12] and Gaussian mixture models (GMMs) [10, 11]. GMMs can be thought of as a special subset of HMMs where no positional constraints are placed on the patch observations whatsoever. This is a highly desirable characteristic when trying to verify clients across pose as patch positions can vary wildly across viewpoints.

In this paper we will attempt to address the following two questions with respect to face verification via monolithic and free-parts representations:

**Q1:** Are areas of the face which are often associated with being the most salient and discriminative (i.e. eyes, nose and mouth) equally important for all representations of the face? Or can other traits such as skin texture play a larger role depending on the representation employed?

**Q2:** Is there any benefit in combining the match-scores resulting from a free-parts and monolithic representation? Can additional benefit be gained by combining these scores in an unequal manner?

As a result of answering the above questions we will also be presenting an algorithm which we refer to as the Free-parts and Holistic Integration (FHI) strategy. The FHI strategy is able to give substantial performance improvement in comparison to leading monolithic and free-parts approaches in the presence of pose mismatch.

# 2. Monolithic representations

It is outside the scope of this paper to perform a large scale evaluation of all possible monolithic approaches. Instead we will be taking a sample of techniques that are representative of current paradigms in pose robust face recognition. These paradigms differ largely by how they employ the development set in their off-line training. We define the development set as the set of observations used to obtain any data-dependent aspects of the verification algorithm (e.g. Eigenface, Fisherface, Eigen-light Field vectors, whitening transform, etc.), but does *not* provide any client specific information like those found in the gallery and probe sets.

Specifically, we will be considering the Eigenface algorithm [13] as a baseline due to its ubiquitous nature in face recognition literature. The Eigenface algorithm can be thought of as being representative of a paradigm that make matches based purely on pixel appearance. The Fisherface algorithm [14] is also considered as a baseline due to its simplicity and high performance in recent evaluations [15, 16, 17]. This algorithm can be thought of as being representative of a paradigm that attempts to learn the within-class and between-class differences between poses in the development set. Finally, the Eigenlight-fields technique will be used as a baseline due to its specificity to pose and its similar nature to other popular approaches such as Tensorfaces [6] as well as Lee and Kim's [8] pose transformation technique. These types of algorithms are representative of a paradigm that attempts to learn the relationships/transformations between each pose in the development set.

## 2.1. Eigen- and Fisher-faces

Eigen- and Fisher-face approaches have been around for quite some time and have enjoyed much success in full-frontal face recognition. In this paper we will be evaluating a specific type of Eigen- and Fisher-face strategy. The first, which will be referred to as MON-PCA, is the baseline Eigenface [13] technique which employs principal component analysis (PCA) to generate a subspace preserving the $K = 89$ most energy preserving modes. The whitened cosine distance (i.e. the cosine distance between two observations after performing the whitening transform [18] on each of them) is then employed to gain a measure of similarity between the gallery and probe observation vectors which result after mapping the original pixel images into the PCA generated subspace. The second technique which we shall refer to as MON-LDA, is a variant on the Fisherface [14] technique which employs linear discriminant analysis (LDA), after PCA, to generate a subspace preserving the $K = 89$ most discriminant modes. As suggested by [15, 16, 17] good performance can be attained if we employ the cosine distance to gain a measure of similarity.

## 2.2. Eigen-light Field Approach

Eigen-light fields were proposed by Gross *et al.* [7] as a technique for learning the dependencies that exist between monolithic representations of the face from different view points. In their paper Gross *et al.* argue that a face's light field is an ideal representation to perform face recognition under varying pose as the representation naturally encompasses all view points. A face was assumed to stem from only a finite set of poses $1, 2, \ldots, P$. In their work a light-field was represented as the concatenation of the vectorized view point images $\mathbf{x}_p$ such that $\boldsymbol{\ell} = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T\right]^T$ (i.e. the light-field was assumed to be represented accurately from $P$ sample viewpoints). From an ensemble of $K$ training light-fields $\{\boldsymbol{\ell}_k\}_{k=1}^K$ a set of eigenvectors $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^K$ (i.e. eigen-light fields) can be found through PCA that satisfy,

$$\boldsymbol{\ell} = \sum_{k=1}^{K} a_k \mathbf{v}_k + \overline{\boldsymbol{\ell}} = \mathbf{V}\mathbf{a} + \overline{\boldsymbol{\ell}} \tag{1}$$

where $\overline{\boldsymbol{\ell}}$ is the sample mean of the light fields. As long as $\boldsymbol{\ell}$ lies in the same approximate subspace as the eigen-light fields the vector $\mathbf{a}$ can be used as a compact pose-invariant representation of that subject's face. In practice however, one rarely has all possible view points to construct a complete light-field. In fact, it is quite common to only have a single gallery view point. In this common scenario a least squares approximation of $\mathbf{a}$ can be found by,

$$\mathbf{a} \approx \mathbf{V}_p^{+}\mathbf{x}_p + \overline{\mathbf{x}}_p \tag{2}$$

where $\mathbf{V}_p$ is the subset, referring to pose $p$, of the complete set of eigen-light fields $\mathbf{V} = \left[\mathbf{V}_1, \ldots, \mathbf{V}_P\right]^T$. The Moore-Penrose inverse, denoted by the $^+$ superscript, of $\mathbf{V}_p$ needs to be found to gain the least squares solution, as the set of vectors contained in $\mathbf{V}_p$ are not assured of being orthonormal. Once the vector $\mathbf{a}$ is estimated the cosine distance is used to gain a match-score between gallery and probe images. Throughout the experimental portion of this paper we shall refer to this specific technique as LF-PCA.

# 3. Free-parts Representations

Learning the face as a distribution (i.e. many observations), as opposed to a single observation, has many appealing properties for face classification tasks. First, the many observations (representing the face) can exist in a low dimensional space circumventing problems associated with the "curse of dimensionality" [18] when training a classifier with high dimensional observations. Second, by representing a face with many observation points one naturally has more observations (of a lower dimensionality) to aid in the estimation of a classifier's parameters. Through the use of GMMs to model the face distribution, it has been shown [10, 11] that good verification performance can be attained by throwing away most position/structure information. We refer to this type of face model as a free-parts GMM (FP-GMM). In this subsection we briefly explain what features we use to estimate the FP-GMM, how it is estimated and how we evaluate the GMM during verification.

## 3.1. Free-parts GMMs

To estimate or evaluate a FP-GMM for a subject, the subject's geometrically and statistically normalized images are first decomposed into $16 \times 16$ pixel image patches with a $75\%$ overlap between horizontally and vertically adjacent patches. Each image patch has a 2D-DCT applied to it in order to compact the 256 elements into a feature vector $\mathbf{o}$ of dimensionality $D$. Based on preliminary experiments, we have chosen $D = 35$. Additional information about the generation of the feature representations can be obtained from [10, 11].

A GMM models the probability distribution of a $D$ dimensional random variable $\mathbf{o}$ as the sum of $M$ multivariate Gaussian functions,

$$f(\mathbf{o}|\boldsymbol{\lambda}) = \sum\nolimits_{m=1}^{M} w_m \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{3}$$

where $\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the evaluation of a normal distribution for observation $\mathbf{o}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The weighting of each mixture component is denoted by $w_m$ and must sum to unity across all components. In our work the covariance matrices in $\boldsymbol{\lambda}$ are assumed to be diagonal such that $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\sigma}\}$, as substantial benefit can be attained by reducing the number of parameters that need to be estimated.

Given a world model $\boldsymbol{\lambda}_{\mathtt{w}} = \{w_{\mathtt{w}_m}, \boldsymbol{\mu}_{\mathtt{w}_m}, \boldsymbol{\Sigma}_{\mathtt{w}_m}\}_{m=1}^{M}$ and training observations from a particular client, $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_R\}$, the GMM parameters for that client are estimated through relevance adaptation (RA) [10].

The world model is simply a single model trained from a large number of subject faces representative of the general population (i.e. the development set). In our work we have found best performance was attained when the world model was estimated using frontal observations *only*. This was done to ensure the final client model was discriminating against subject identity only, and not other poses present in the world model. The world model's parameters are estimated using the Expectation Maximization (EM) algorithm [19], configured to maximize the likelihood of training data. RA is an instance of the EM algorithm configured for maximum *a posteriori* (MAP) estimation, rather than simply maximum likelihood (ML). It has been noted that great benefit can be obtained in terms of estimating high performance robust FP-GMMs by employing RA when only small amounts of client specific observations exist (e.g. a single enrollment image). Using RA, parameters for client $\mathtt{c}$ are obtained using the following update equations:

$$w_{\mathtt{c}_m} = \beta\left[(1-\alpha_m^w)w_{\mathtt{w}_m} + \alpha_m^w \frac{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)}{\sum_{m=1}^M \sum_{r=1}^R \gamma_m(\mathbf{o}_r)}\right] \quad (4)$$

$$\boldsymbol{\mu}_{\mathtt{c}_m} = (1-\alpha_m^\mu)\boldsymbol{\mu}_{\mathtt{w}_m} + \alpha_m^\mu \frac{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)\mathbf{o}_r}{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)} \quad (5)$$

$$\boldsymbol{\sigma}_{\mathtt{c}_m} = (1-\alpha_m^\sigma)\left(\boldsymbol{\sigma}_{\mathtt{w}_m} + \boldsymbol{\mu}_{\mathtt{w}_m}^2\right) + \alpha_m^\sigma \frac{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)\mathbf{o}_r^2}{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)} - \boldsymbol{\mu}_{\mathtt{c}_m}^2 \quad (6)$$

where $\gamma_m(\mathbf{o})$ is the occupation probability for component $m$, $\boldsymbol{\mu}^2$ indicates that each element in $\boldsymbol{\mu}$ is squared, and $\alpha_m^\rho$ is a weight used to tune the relative importance of the prior; it is defined as:

$$\alpha_m^\rho = \frac{\sum_{r=1}^R \gamma_m(\mathbf{o}_r)}{\tau^\rho + \sum_{r=1}^R \gamma_m(\mathbf{o}_r)} \quad (7)$$

where $\tau^\rho$ is a *relevance* factor. The above definition of $\alpha_m^\rho$ can limit the adaptation to only the Gaussians for which there is sufficient data. We have found effective performance can be attained by using a single relevance factor ($\tau = \tau^w = \tau^\mu = \tau^\sigma$). Based on empirical evaluation on many data sets, we have chosen $\tau = 10$. The scale factor, $\beta$, in Equation 4 is computed to ensure that all the adapted component weights sum to unity. The adaptation procedure is iterative, thus an initial client model is required. This is accomplished by copying the world model.

In RA, the distributions are estimated by finding and using observations that aid in discriminating client models from the world model. As such, the distributions should not be considered as generative distributions (i.e. distributions that can be used for producing synthetic observations representative of a particular client). In this sense the GMM based classifier, trained via RA, is inherently discriminative and is able to obtain good classification performance with sparse amounts of training data. Additional information on RA can be found in [10].

### 3.2. Evaluating a FP-GMM

To evaluate a sequence of observations, generated from a claimant's probe image, we obtain the average log-likelihood,

$$\mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_c) = \frac{1}{R}\sum_{r=1}^R \log f(\mathbf{o}_r|\boldsymbol{\lambda}_c) \quad (8)$$

Given the average log-likelihood, for the client and world models, one can then calculate the log-likelihood ratio,

$$\Lambda(\mathbf{O}) = \mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_c) - \mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_w) \quad (9)$$

For our work we found good performance across pose could be attained if we employed GMMs with $512$ components.

## 4. Face Database and Verification

Experiments were performed on a subset of the FERET database [20], specifically images stemming from the *ba*, *bb*, *bc*, *bd*, *be*, *bf*, *bg*, *bh*, and *bi* subsets; which approximately refer to rotation's about the vertical axis of $0^o$, $+60^o$, $+40^o$, $+25^o$, $+15^o$, $-15^o$, $-25^o$, $-40^o$, $-60^o$ respectively. The database contains $200$ subjects which were randomly divided into an evaluation and development set both containing $90$ subjects. The remaining $20$ subjects were used as an imposter set for our verification experiments. As mentioned previously, the development set is used to obtain any data-dependent aspects of the verification system (e.g. subspace, world models etc.). The evaluation and imposter sets are where the performance rates for the verification system are obtained.

Traditionally, before performing the act of face recognition, some sort of geometric pre-processing has to go on to remove variations in the face due to rotation and scale. The distance and angle between the eyes has long been regarded as an accurate measure of scale and rotation in a face. However, this type of geometric normalization, based purely on the eye position, becomes problematic when faced with depth pose rotation due to a stretching of the image in the y-axis. In our work we chose to employ the distance from the eye line to the nose tip vertically to remedy the stretching problem. The final geometrically normalized cropped faces formed an $98 \times 115$ array of pixels.

The face verification task is the binary process of accepting or rejecting the identity claim (i.e. the log-likelihood ratio or cosine distance match-score from the free-parts and monolithic recognizers respectively) made by a subject under test. A threshold $Th$ needs to be found so as to make the decision. Face verification performance is evaluated in terms of two types of error: a) being false rejection (FR) error, where a true client is rejected against their own claim, and b) false acceptance (FA) errors, where an impostor is accepted as the falsely claimed subject. The FA and FR errors increase or decrease in contrast to each other based on the decision threshold $Th$ set within the system. A simple measure for overall performance of a verification system is found by determining the equal error rate (EER) for the system, where FA = FR.

## 5. Leading monolithic techniques

Before embarking on our analysis of the differences between leading monolithic representations and our proposed free-parts representation algorithm (i.e. FP-GMMs) it is first important to establish which monolithic technique performs best in the presence of pose mismatch for our experimental framework. Interesting work has already been conducted by Lee and Kim [8] concerning whether it is better to: a) learn the within-class and between-class differences

| Pose | MON-PCA | LF-PCA | MON-LDA |
|---:|---:|---:|---:|
| -60 | 26.67 | 14.58 | 13.33 |
| -40 | 17.78 | 12.20 | 9.93 |
| -25 | 10.22 | 11.27 | 6.64 |
| -15 | 6.67 | 10.00 | 4.56 |
| 15 | 6.67 | 11.11 | 6.49 |
| 25 | 8.89 | 11.19 | 5.58 |
| 40 | 15.55 | 14.24 | 9.05 |
| 60 | 24.44 | 13.56 | 11.11 |
| ***Average*** | *14.61* | *12.27* | *8.34* |

Table 1: Comparison of monolithic paradigms for good performance in the presence of a pose mismatch. In these results one can see for a modest development set size of 90 subjects across 9 poses a strategy of learning the within-class differences, through LDA, performs best overall.

between poses (i.e. discriminant analysis), or b) learn the relationships (i.e. transformations) between each possible within-class variation/view-point.

The technique Lee and Kim employed to learn transformations between poses, although not explicitly the same, is very similar to previous techniques like Eigen-light Fields [7] and Tensorfaces [6]. In all three techniques a least squares linear mapping is learnt to transform from a previously unseen pose of the claimant to one or many viewpoints (in the case of light-fields) that have been seen in enrollment. In Lee and Kim's work a combination between the two paradigms seemed to work best, where one first transformed the probe image to a frontal view and then applied discriminant analysis to the result. This approach was however, dependent on having ample development observations (they used over 245 subjects in their development set with only 5 poses) to learn both the transformation and discriminant analysis subspace.

In our work we opted to only compare the paradigms of discriminant analysis and transformation through the MON-LDA and LF-PCA approaches respectively, as the development set we were employing (only 90 subjects with 9 poses) gave poor results when trying to combine both paradigms. In Table 1 one can see the results for all the monolithic approaches outlined in this paper.

Although by no means comprehensive, this analysis is informative as it demonstrates that a monolithic paradigm that attempts to learn the within-class and between-class differences (e.g. MON-LDA), as opposed to learning the within-class relationships/transformations (e.g. LF-PCA) tends to perform better with our pre-defined development set. As expected, both techniques on average performed better than simple appearance techniques like Eigenfaces (MON-PCA). An open issue for further investigation is how the size and variation of the development set can affect which monolithic paradigm to employ.

# 6. Q1: Face traits and representation?

In the plethora of work that has been done with monolithic and rigid-parts approaches, for frontal view face recognition, it has been demonstrated that the eye, nose and mouth regions are considered most salient for the purposes of face recognition. Most notably the work by Moghaddam and Pentland [21] concerning modular eigenspaces depicted the superior performance attained by individually modelling components of the face (eyes, nose, mouth) and discarding the residual part of the face. A problem with face recognition across pose however, using monolithic and rigid-parts techniques, is that these salient areas are most often the most warped and distorted during pose variation due to their 3D nature (e.g. the nose). An important question was raised during the development of our work; do free-parts representations of the face rely on these same salient areas prone to large non-linear variation from pose change?

In Figure 2 one can see a number of evaluation images, in the first column, along with their associated log-likelihood ratio (LLR) score maps generated from the FP-GMM for each image patch, in the second column. If one was to take the sum of the LLR values in the map, they would result in the final LLR values for that claimant image which is consistent with Equations 8 and 9. Inspecting the LLR-maps in Figure 2 one can see that regions of the face that are often associated with being most salient for recognition in monolithic and rigid-parts representations (i.e. the eye and nose regions) are extremely dark indicating their minimal contribution to the free-parts verification process. Other areas of the face which have often thought to be of minimal benefit in monolithic and rigid-parts representations, such as the brow and cheeks and most notably in this example the nose bridge, demonstrate a very high contribution to the free-parts verification process. This leads us to propose a hypothesis. Do free-parts techniques like the one employed by the FP-GMM actually learn the client's skin texture and not other traits (i.e. eyes and nose) of the face long thought to be essential for good face recognition?

There is strong evidence to support this hypothesis. In previous work [22], for the task of frontal face verification, a complimentary relationship between monolithic and free-parts based representations was first established. In those experiments the authors were able to demonstrate that monolithic type approaches like Fisherfaces operate predominantly on the lower-frequency information contained in the face whereas free-parts based techniques like our own FP-GMM technique are quite dependent on higher-frequency traits, like skin texture, contained in the face while largely ignoring the global structure of the face image.

In this correspondence we have devised an experiment where we have attempted to remove those areas of the face thought to contribute most highly to the verification process

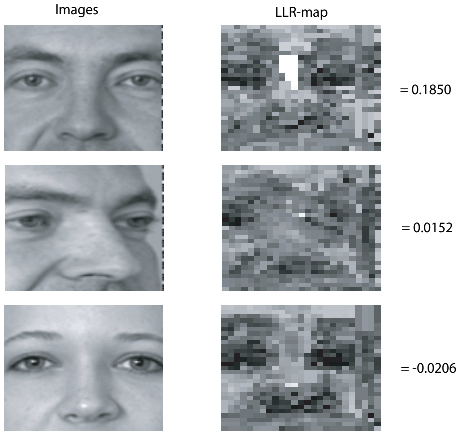

Figure 2: Depiction of grayscale images in column 1 with their respective FP-GMM log-likelihood ratio (LLR) maps in column 2. Row 1 depicts the client's train image used to estimate the FP-GMM. Row 2 depicts the client in a non-frontal pose, which was not employed in training. Row 3 depicts an imposter in frontal view. To the right of column 2 one can see the total LLR values for each image demonstrating the pose invariant properties of the FP-GMM algorithm (i.e. row 2 has a higher total LLR than row 3).

(i.e. the eyes and nose regions). We decided to form an experiment where we compare the performance of the FP-GMM and MON-LDA techniques, which for the purposes of this paper are representative of free-parts and monolithic techniques respectively. A depiction of the masks used to ignore these regions for each pose can be seen in Figure 3. One can see these masks are pose dependent as the size of the eyes as well as the position and size of the nose vary as a function of viewpoint.

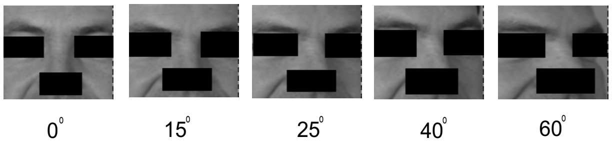

Figure 3: Depiction of example FERET pose images having the eye & nose regions ignored.

For the MON-LDA approach a similar technique to the one employed in Section 2.2 for Eigen-light Fields was used to cope with the problem of enrolling and evaluating face images with missing data[1]. One can see the results in Table 2, in terms of the difference with the normal non-masked MON-LDA approach, for the situation where only the eye & nose regions as well as the residual region (i.e. skin areas) were used to verify clients. To assure the accuracy of the technique used with the MON-LDA approach for enrolling and evaluating faces with missing data we also conducted tests where random face masks were generated for each pose so that 80% and 50% of the pixels remained.

| POSE | Eyes & Nose | Residual | Rnd (80%) | Rnd (50%) |
|---:|---:|---:|---:|---:|
| **-60** | 30.08 | 26.58 | -0.23 | 1.11 |
| **-40** | 22.29 | 22.29 | 0.00 | 0.06 |
| **-25** | 17.60 | 16.69 | -0.14 | -0.97 |
| **-15** | 13.04 | 5.45 | 0.15 | -0.10 |
| **15** | 6.87 | 12.37 | -0.69 | -0.88 |
| **25** | 23.41 | 25.53 | 0.10 | -0.04 |
| **40** | 37.62 | 44.04 | -0.03 | 2.03 |
| **60** | 37.78 | 53.53 | 0.00 | 0.00 |
| ***Average*** | 23.59 | 25.81 | -0.11 | 0.15 |

Table 2: Results demonstrating the subtracted difference between the original MON-LDA EERs (%) and those for representations where some areas of the face is masked. The *Eye & Nose* masks were for experiments where only those areas were available. The *Residual* masks were for the opposite situation where the eye and nose regions were not available. To validate our results we also employed random masks (*Rnd*) with a percentage (50 and 80 %) of pixels being employed. One can see the dramatic deteriorating affect in performance of removing both eye & nose regions as well as the residual skin regions. Employing the random skin masks however, had a negligible effect.

One can see that the MON-LDA technique is very much a technique reliant on a holistic representation of the face with neither the eye & nose or residual skin masks giving dominant results. Employing random masks resulted in no performance degradation whatsoever for the 80% scenario (with results actually being slightly better in most cases) and only slightly poorer for the 50% scenario. This result demonstrates that the missing data technique being employed for the MON-LDA algorithm is valid and also gives some additional evidence that the current MON-LDA representation may be over sampled and could perform well using only lower frequency detail. Results for both the eye & nose and residual masks were considerably poorer in relation to their original values with results becoming catastrophic the further in viewpoint from the frontal pose the evaluation faces became.

A noticeably different result occurs in our analysis of this same experiment with the FP-GMM approach. One can see in Table 3 that there is minimal difference between representations where the face contains and does not contain eye & nose information. Further, for the FP-GMM algorithm where only the eyes & nose regions were employed, performance is quite poor in comparison to results attained from FP-GMM algorithm when employing the entire face or residual face area. One can assume from this result that the residual skin area is the dominant trait being used for verification with the FP-GMM approach. Interestingly however, the eye & nose only performance is still comparable with the leading monolithic technique (i.e. MON-LDA) for slightly off frontal view-points. One hy-

[1] The subspace generated from LDA is usually not orthonormal so an additional MP inverse had to be applied before applying the technique used in Section 2.2.

| POSE | Eyes & Nose | Residual |
|---:|---:|---:|
| **-60** | 7.7778 | -3.33 |
| **-40** | 11.09872 | 0.25 |
| **-25** | 4.44445 | 0.00 |
| **-15** | 2.19136 | -0.77 |
| **15** | 3.40741 | -0.10 |
| **25** | 8.84255 | 1.06 |
| **40** | 13.5463 | -0.06 |
| **60** | 22.7161 | 5.70 |
| ***Average*** | 9.25 | 0.34 |

Table 3: Results demonstrating the subtracted difference between the original FP-GMM EERs (%) and those for representations where some areas of the face is masked. The *Eye & Nose* masks were for experiments where only those areas were available. The *Residual* masks were for the opposite situation where the eye and nose regions were not available. Note: there is minimal affect in performance when the eyes and nose are removed, but there is a substantial deteriorating affect in performance when only the eyes and nose remain (i.e. no skin to texture to process).

pothesis for this result could be that there is some benefit in obtaining different FP-GMM representations for different salient regions of the face; as their may be a tendency for the FP-GMM algorithm to learn the most dominating trait (i.e. the skin texture) and not other traits when learning is done in an unsupervised manner.

## 7. Q2: A rationale for integration?

One can see from the previous section that there is strong evidence that the monolithic and free-parts representations employ different or at least place unequal weights on different traits of the face. Heuristically we hypothesize that there should be some benefit in combining these two representations, which the concluding experiments of the paper attempt to explore. We refer to the combination of these two representations as a Free-parts and Holistic Integration (FHI) strategy.

We employ the sum rule for combining the match scores from the classifiers of the two representations. Kittler *et al.* [23] demonstrated that the sum rule can obtain good performance in classifier combination, when the two classifiers are diverse and produce match scores approximately representative of their true a posteriori probabilities. The final combined match-score is generated by,

$$ms = \alpha \underbrace{\text{logsig}(\Lambda)}_{\text{free parts}} + (1-\alpha) \underbrace{\text{logsig}(d_{COS})}_{\text{monolithic}} \qquad (10)$$

where $\text{logsig}(a) = 1/(1+\exp(-a))$ is used to try and make the match scores obtained from the MON-LDA and FP-GMM algorithms more representative of their true a posteriori probabilities. The employment of the logsig operation results in the synergetic combination of match scores from the MON-LDA and FP-GMM algorithms. A weighting factor $\alpha$, which was allowed to vary between zero and one, was employed with the sum rule so as to place more emphasis on one representation over another as a function of pose. One can see in Figure 4 an example of how the variation of $\alpha$ can affect performance of the FHI algorithm. Through a cross-validation process the performance seen in Figure 4 tended to vary however, one could see two trends emerge. First, that the weighting factor should be greater than $0.5$ for all poses, indicating more emphasis should nearly always be placed on the free-parts representation than the monolithic representation in the presence of pose mismatch. Second, the further in viewpoint from the frontal pose the probe image becomes the more sensitive the FHI algorithm becomes to the correct selection of $\alpha$, as depicted in Figure 4. From cross-validation we found an $\alpha = 0.75$ performed best at the larger viewpoints of +/- $45^o$ and $60^o$ with the smaller viewpoints being largely insensitive to the selection.

Results for FHI strategies where equal $\alpha = 0.5$ and unequal weightings $\alpha = 0.75$ were employed, can be seen in Figure 5, compared to leading monolithic (MON-LDA) and free-parts (FP-GMM) algorithms. One can see the FHI strategy outperforms both monolithic and free-parts algorithms across all poses and in most cases by a substantial margin. One can also see that the accurate selection of an appropriate weight $\alpha$ makes a difference in verification performance for larger non-frontal viewpoints.

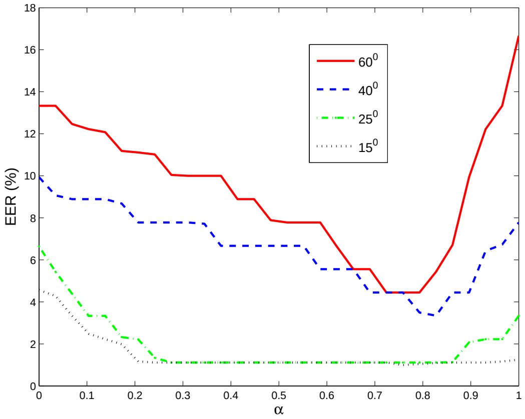

Figure 4: Effect of varying $\alpha$ in the FHI strategy for various poses. Note: minimum EERs are achieved at different values of $\alpha$ depending on pose. Larger non-frontal pose angles are far more sensitive to the correct selection of $\alpha$ than smaller non-frontal pose angles.

## 8. Summary and Conclusions

The FHI results presented in this paper give convincing evidence that there is benefit in combining monolithic and free-parts representations for the purposes of automatic face verification in the presence of pose mismatch. We have additionally made the novel contribution in offering evidence

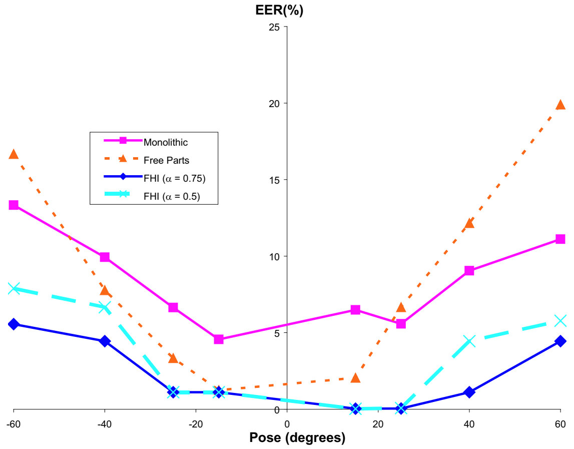

Figure 5: Final results demonstrating the benefit of a FHI strategy across all poses compared with monolithic (MON-LDA) and free-parts (FP-GMM) representations. Results also demonstrate that an unequal weighting of $\alpha = 0.75$ between monolithic and free-parts match-scores produces improved results at the larger non-frontal viewpoints.

that free-parts representations of the face may be placing greater emphasis on traits of the face, such as skin texture, that canonical monolithic representations at the moment do not employ. This insight gives further explanation into why these two representations are able to be integrated in such a synergetic manner, as they are attempting to verify subjects based on two different and diverse traits of the face.

Currently our FHI framework uses an adhoc technique to calculate an appropriate weighting factor for use across all poses. In future work we would like to explore a more empirical and pose dependent weighting strategy for larger viewpoints. In additional future work we would like to incorporate a rigid-parts based algorithm into our integration strategy to see if further synergetic performance can be attained.

# References

[1] J. W. Tanaka and M. J. Farah, "The holistic representation of faces," in *Perception of Faces, Objects, and Scenes* (M. A. Peterson and G. Rhodes, eds.), ch. 2, pp. 53–74, Oxford University Press, Inc., 2003.

[2] J. E. Murray, G. Rhodes, and M. Schuchinsky, "When is a face not a face?," in *Perception of Faces, Objects, and Scenes* (M. A. Peterson and G. Rhodes, eds.), ch. 3, pp. 75–91, Oxford University Press, Inc., 2003.

[3] H. Schneiderman and T. Kanade, "A histogram-based method for detection of faces and cars," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 504–507, September 2000.

[4] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *European Conference on Computer Vision (ECCV)*, pp. 18–32, 2000.

[5] M. Weber, M. Welling, and P. Perona, "Towards automatic discovery of object categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 101–108, June 2000.

[6] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: TensorFaces," in *European Conference on Computer Vision (ECCV)*, vol. 2350 of *Lecture Notes in Computer Science*, (Berlin), pp. 447–460, Springer-Verlag, 2002.

[7] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Trans. PAMI*, vol. 26, pp. 449–465, April 2004.

[8] H. Lee and D. Kim, "Pose invariant face recognition using linear pose transformation in feature space," in *European Conference on Computer Vision (ECCV)*, 2004.

[9] T. Kanade and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, (Kobe, Japan), pp. 954–958, July 2003.

[10] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. II, (Washington D.C.), pp. 855–861, June 2004.

[11] C. Sanderson and K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.

[12] S. Eickeler, S. Muller, and S. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image and Vision Computing*, vol. 18, no. 4, pp. 279–287, 2000.

[13] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[15] P. Navarrete and J. Ruiz-del-Solar, "Analysis and comparison of eigenspace-based face recognition approaches," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 7, pp. 817–830, 2002.

[16] J. Ruiz-del-Solar and P. Navarrete, "Towards a generalized eigenspace-based face recognition framework," in *4th Int. Workshop on Statistical Techniques in Pattern Recognition*, (Windsor, Canada), August 2002.

[17] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer, "A comparative study of automatic face verification algorithms on the BANCA database," in *AVBPA*, pp. 35–43, 2003.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: John Wiley and Sons, Inc., 2nd ed., 2001.

[19] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[20] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 10, no. 22, pp. 1090–1104, 2000.

[21] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object recognition," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 696–710, 1997.

[22] S. Lucey, "The symbiotic relationship of parts and monolithic face representations in verification," in *International Workshop on Face Processing in Video (FPIV)*, (Washington D.C.), June 2004.

[23] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. PAMI*, vol. 20, pp. 226–239, March 1998.