# MUTUAL FEATURES FOR ROBUST IDENTIFICATION AND VERIFICATION

*Heiko Claussen*[*]*, Justinian Rosca*

Siemens Corporate Research Inc.
755 College Road East
Princeton, NJ 08540

*Robert Damper*

School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK

## ABSTRACT

Noisy or distorted video/audio training sets represent constant challenges in automated identification and verification tasks. We propose the method of Mutual Interdependence Analysis (MIA) to extract "mutual features" from a high dimensional training set. Mutual features represent a class of objects through a unique direction in the span of the inputs that minimizes the scatter of the projected samples of the class. They capture invariant properties of the object class and can therefore be used for classification. The effectiveness of our approach is tested on real data from face and speaker recognition problems. We show that "mutual faces" extracted from the Yale database are illumination invariant, and obtain identification error rates of 2.2% in leave-one-out tests for differently illuminated images. Also, "mutual speaker signatures" for text independent speaker verification achieve state-of-the-art equal error rates of 6.8% on the NTIMIT database.

***Index Terms***— Algorithms, Signal Processing, Pattern Classification, Signal Analysis, Speaker/Face Recognition.

## 1. INTRODUCTION

Principal Component Analysis (PCA) is a ubiquitous feature extraction and dimensionality reduction method [1], [2]. Principal components/functions are given by the directions of maximum variance in the data. The directions of minimum variance, or minor components, have received much less attention in the literature. However, Minor Component Analysis (MCA) is important in certain signal processing applications e.g. spectral estimation, curve and hyper-surface fitting, cognitive perception and computer vision [3].

Both, PCA and MCA principles are successfully utilized in classification problems. On the one hand, PCA can find the directions of maximum scatter between classes representing effective contrasts for classification. On the other hand, MCA can extract invariant representations of each class by finding a direction that minimizes the intra-class scatter. Utilizing Bayesian statistics with Gaussian priors of equal covariance, one can derive the Fisher linear discriminant analysis (LDA) [4]. This classical technique finds a trade-off between minimizing intra-class scatter and maximizing between-class scatter. Thus, the cost function of LDA can be defined by a combination of the PCA and MCA principles.

Data-dependent transformations (like PCA, independent component analysis or ICA, MCA), in contrast to general-purpose transformations (Fourier, wavelet analysis), can extract powerful representations to reason about new inputs with similar underlying structure to the training data [5]. However, when attempting to extract interdependencies in a dataset, most methods lose information through the common preprocessing step of mean subtraction. High-dimensional input samples are generally linearly independent, thus mean subtraction can reduce the span of the data and lose information.We would like to extract invariants/features through data-dependent transformations of these raw inputs.

In this paper, we propose Mutual Interdependence Analysis (MIA) [6] for robust feature extraction. In section 2 we define the MIA problem, state its solution and discuss its properties. Sections 3 and 4 show the application of MIA to face identification and text-independent speaker recognition problems respectively. Our approach is effective in learning "mutual faces" and "mutual speaker signatures" from the data and achieves competitive error rates on challenging data.

## 2. MUTUAL INTERDEPENDENCE ANALYSIS (MIA)

Throughout the paper, we use $x_i(t_j)$, with $i = 1, \ldots, N$ and $j = 1, \ldots, D$, to denote $N$ real-valued inputs of dimensionality $D$. In our case, $D$ is typically much larger than $N$. Also, we denote $\mathbf{X}$ to be the matrix whose columns are $\mathbf{x}_i$. For example, $\mathbf{x}_i$ is an image representing the face or a speech segment from one person $p$, and $\mathbf{X}^{(p)} \subseteq \mathbf{X}$ represents the set of such samples for the person. In that case $\mathbf{X}$ denotes a concatenation of the matrices $\mathbf{X}^{(p)}$ of all classes $p$ in a database.

[*]Further affiliation of the author: School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK, hc05r@ecs.soton.ac.uk

Consider the scatter of the data:

$$\tilde{S}(\mathbf{X}|\mathbf{w}) = \sum_{i=1}^{N} \left(\mathbf{w}^T \cdot \mathbf{x}_i - \mathbf{w}^T \cdot \mu\right)^2 = \mathbf{w}^T \cdot \mathbf{S} \cdot \mathbf{w}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$ is the $D$-dimensional sample mean, and $\mathbf{S}$ is the scatter matrix of the data.

The basic MIA idea is to look for invariant input data features (i.e. directions to project data) in the data-dependent space $\left\{\sum_{k=1}^{N} c_k \cdot \mathbf{x}_k : \mathbf{c} \in I\!R^N\right\}$. For any input set $\mathbf{X}^{(p)}$, the MIA problem can be formulated as the search for an optimal direction $\mathbf{w}_{\mathrm{MIA}}^{(p)}$ that solves the constrained optimization:

$$\hat{\mathbf{w}}_{\mathrm{MIA}}^{(p)} = \underset{\mathbf{w},\|\mathbf{w}\|=1,\mathbf{w}=\sum_{k=1}^{N} c_k \cdot \mathbf{x}_k}{\arg\min} \tilde{S}(\mathbf{X}^{(p)}|\mathbf{w}) \ . \qquad (1)$$

Note the differences from formulations of the classical problems PCA, MCA and LDA:

$$\begin{aligned}
\hat{\mathbf{w}}_{\mathrm{PCA}} &= \underset{\mathbf{w},\|\mathbf{w}\|=1}{\arg\max}\, \tilde{S}(\mathbf{X}|\mathbf{w}) \\
\hat{\mathbf{w}}_{\mathrm{MCA}}^{(p)} &= \underset{\mathbf{w},\|\mathbf{w}\|=1,\tilde{S}(\mathbf{X}^{(p)}|\mathbf{w})>0}{\arg\min}\, \tilde{S}(\mathbf{X}^{(p)}|\mathbf{w}) \\
\hat{\mathbf{w}}_{\mathrm{LDA}} &= \underset{\mathbf{w},\|\mathbf{w}\|=1}{\arg\max} \frac{\tilde{S}(\mathbf{X}|\mathbf{w})}{\sum_p \tilde{S}(\mathbf{X}^{(p)}|\mathbf{w})}
\end{aligned}$$

Our optimization problem is unique due to its constrained formulation. We will show that $\hat{\mathbf{w}}_{\mathrm{MIA}}$ "optimally" represents the data samples for a class as one aggregate sample.

### 2.1. Solution to MIA

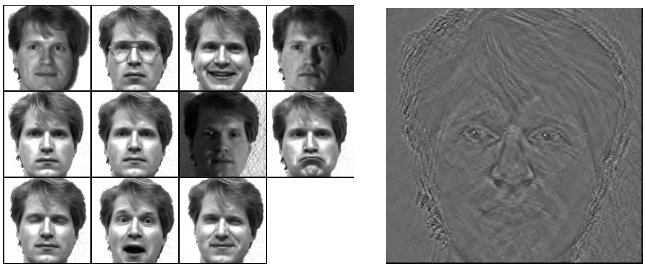
We sketch an equivalent formulation of the MIA problem and its solution. Let us denote $\mathbf{y}_i = \mathbf{x}_i - \mu$. If for simplicity $\mathbf{X}^{(p)} = \mathbf{X}$, (1) can also be written as:

$$\hat{\mathbf{w}}_{\mathrm{MIA}} = \underset{\mathbf{w},\|\mathbf{w}\|=1,\mathbf{w}=\sum_{k=1}^{N} c_k \cdot \mathbf{x}_k}{\arg\min} \left\|\mathbf{w}^T \cdot \mathbf{Y}\right\|^2 \ , \qquad (2)$$

where $\mathbf{Y} = \mathbf{X} - \mu \cdot \underline{\mathbf{1}}^T$ and $\mu = \frac{1}{N}\mathbf{X} \cdot \underline{\mathbf{1}}$. It follows that $\mathbf{Y} = \mathbf{X} \cdot \mathbf{P}$ with $\mathbf{P} = \mathbf{I} - \frac{1}{N}\underline{\underline{\mathbf{1}}}$. Obviously, $\sum_{i=1}^{N} \mathbf{y}_i = \underline{\mathbf{0}}$. Thus, the nullspace $\mathcal{NULL}(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is non trivial. All vectors $\mathbf{w} \in \mathcal{NULL}(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ will minimize $\tilde{S}(\mathbf{X}|\mathbf{w})$. The following theorem shows that the problem given by (2) has exactly one solution.

**Theorem 1** *Assume $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are linearly independent. Then, there exists $\mathbf{w} \neq \underline{\mathbf{0}}$ in $\mathcal{NULL}(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ such that $\mathbf{w}$ is in the span of the inputs $\mathbf{x}_i$, $i = 1, \ldots, N$.*

The proof of this theorem as well as the derivation of the solution $\hat{\mathbf{w}}_{\mathrm{MIA}} = \zeta\, \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \underline{\mathbf{1}}$, with $\zeta$ an arbitrary scalar, is given in [6]. It follows that $\frac{\hat{\mathbf{w}}_{\mathrm{MIA}}}{\|\hat{\mathbf{w}}_{\mathrm{MIA}}\|}$ is a unique solution to (2). It can be easily seen that common, additive components in all data samples will not affect the scatter matrix $\mathbf{S}$ and therefore will not affect the MIA solution. It turns out that the MIA result $\hat{\mathbf{w}}_{\mathrm{MIA}}$ is equally correlated with all inputs [6]. Should input data be seen as a linear combination of a hidden, invariant part $\mathbf{w}$ and variant components $\mathbf{f}_i$ orthogonal to $\mathbf{w}$, $\mathbf{x}_i = \mathbf{w} + \mathbf{f}_i$, then MIA captures the invariant part $\mathbf{w}$. Hence, the MIA result represents invariant information present in all samples as one aggregate sample.



(a) Yale face database class (b) Mutual image information

**Fig. 1**. (a) Image set of one individual in the Yale database. The set contains 11 images of the person taken with various facial expressions and illuminations, with or without glasses. (b) MIA result, or mutual face estimated from all images of the set.

## 3. APPLICATION TO FACE RECOGNITION

State-of-the-art face recognition approaches suffer from a number of outstanding problems, including sensitivity to multiple illumination sources and diffuse light conditions. In this section, we show that MIA can be used to extract illumination invariant "mutual faces" for face recognition.

We tested the MIA-based mutual face approach on the Yale face database [7]. The image set of one individual is given, for illustration, in Fig. 1(a). As discussed in [8], the reflected light intensity $I$ of each image pixel can be modeled as a sum of an ambient light component and directional light source reflections. Let $I_a$ and $I_p$ be the ambient/directional light source intensities. Also, let $k_a$, $k_d$, $\bar{\mathbf{N}}$ and $\bar{\mathbf{L}}$ be ambient/diffuse reflection coefficients, surface normal of the object, and the direction of the light source respectively. Hence,

$$I = I_a k_a + I_p k_d (\bar{\mathbf{N}} \cdot \bar{\mathbf{L}}) \ .$$

More complex illumination models including multiple directional light sources can be captured by the additive superposition of the ambient and reflective components for each light source [8] (see Equation 16.20).

We claim that MIA can extract an illumination-invariant mutual image, perhaps including $I_a k_a$, from a set of aligned images of the same object (face) under various illumination conditions.

(a) Eigenface input (b) Fisherface input (c) MIA input

**Fig. 2**. Examples of training instances used in (a) Eigenfaces, (b) Fisherfaces and (c) MIA: (a) Mean-subtracted face obtained as difference between a face instance and the mean of all images in the database. (b) Mean-subtracted face obtained as difference between a face instance and the mean image of all instances for the same person. (c) "Centered" face image, obtained by subtraction of the mean column value from each image column.

### 3.1. Face recognition using mutual faces

In the following, mutual faces were used in a simple appearance-based face recognition experiment. Prominent methods of this widely researched area include the Eigenface [9] and Fisherface [7] approaches. Most approaches use mean image subtraction for preprocessing, which reduces the image space dimensionality compared to the original image set. Therefore, this step cancels potentially useful image information. In contrast, MIA uses centered images ($\mathbf{x}_i^T \cdot \underline{\mathbf{1}} = 0 \quad \forall i$) as inputs. Figure 2 illustrates the difference between a mean-face-subtracted input instance in the Eigenface/Fisherface approach and the centered MIA input.

The procedure to extract the mutual face from the face set of one person can be defined as follows: First, images are 2D Fourier transformed. Second, each row of the images is centered and windowed. Thereafter, MIA is performed separately on each set of rows. After normalizing and reassembling the rows the procedure is repeated with the columns of the original images. Thus, two mutual faces are generated, added, and the result is transformed through the 2D inverse FFT. Face identification is performed using cropped and centered images. The measure of similarity between a test image and the MIA representation of a person is the mean cosine distance of the corresponding centered lines and columns. The resulting scores $s_1$ and $s_2$ are fused using $s = \sqrt{{s_1}^2 + {s_2}^2}$.

Mutual faces are learned on all but a single test image using the "leave-one-out" method discussed in [10]. The left-out image is one of the three illumination variant cases of the Yale database (centered light, left light and right light). This approach leads to an identification error rate (IER) of 2.2%. Overall, in exhaustive leave-one-out tests, the mutual face method results in an error rate of 7.4%. Recognition performance for unknown illumination is comparable or beyond various reported results obtained with similar data (Table 1). The MIA approach can be used to enhance both feature- and appearance-based methods, only requires minimal training, and appears insensitive to multiple illumination sources and diffuse light conditions. A complete analysis will be reported separately.

| Method | Evaluation | IER [%] | Comments |
|---|---|---|---|
| MIA | leave-one-out | 7.4 | Full face test |
| | | 2.2 | Only illumination |
| Fisherface [7] | leave-one-out | 7.7 | Cropped face test |
| | | 0.6 | Full face test |
| Eigenface [7] | leave-one-out | 24.4 | Cropped face test |
| | | 19.4 | Full face test |
| Kernel PCA [11] | leave-one-out | 26.0 | Cropped face test |
| Minimax Probability Machine [12] | $k$-fold cross validation | 21.2 | Cropped face test |
| | | 10.1 | Without illumination |

**Table 1**. Comparison of the identification error rate (IER) of MIA with other methods using the Yale database. Full faces include some background compared to cropped images.

## 4. APPLICATION TO TEXT-INDEPENDENT SPEAKER VERIFICATION

In this section, we apply MIA to the problem of extracting signatures from speech data for the purpose of text-independent speaker verification. This problem is challenging when we need to verify the identity of a person but can not control the way data are acquired (e.g. recording equipment, environment, etc.). For comparability with [13], we used the "test" portion of the NTIMIT database [14]. This database contains noisy data from 168 speakers (112 males and 56 females) that we partitioned 50-50 for training and testing. The data were preprocessed by silence removal, low-pass filtering and normalization of each recording.

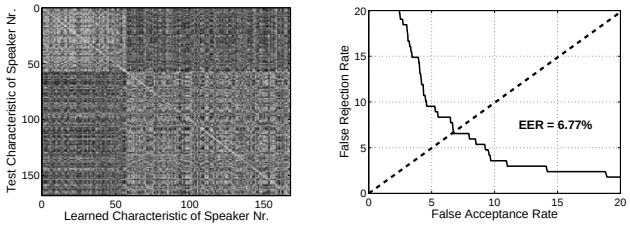
A speech signal can be modeled as an excitation that is convolved with a linear dynamic filter which represents the vocal tract. The excitation signal can be modeled for voiced speech as a periodic signal and for unvoiced speech as random noise. It is common to analyze the voiced and unvoiced speech separately [15]. In this paper, only the voiced speech is used for speaker recognition. Let $\mathbf{E}^{(p)}$, $\mathbf{H}^{(p)}$ and $\mathbf{V}^{(p)}$ be the spectral representations of the excitation, vocal tract filter and the voiced signal parts of person $p$ respectively. Moreover, let $\mathbf{M}$ represent speaker-independent signal parts in the spectral domain (e.g. recording equipment, environment, etc.). Therefore, the data can be modeled as: $\mathbf{V}^{(p)} = \mathbf{E}^{(p)} \cdot \mathbf{H}^{(p)} \cdot \mathbf{M}$. By cepstral deconvolution, the model is represented as a linear combination of its basis functions:

$$\mathbf{x}_i^{(p)} = \log \mathbf{V}_i^{(p)} = \log \mathbf{E}_i^{(p)} + \log \mathbf{H}^{(p)} + \log \mathbf{M}_i$$

This additive model suggests that we could use MIA to extract a function that represents the speaker's signature $\log \mathbf{H}^{(p)}$. In practice, we consider high dimensionality inputs: $\mathbf{x}_i^{(p)}$ are speech segments of one second, in order to achieve high spectral accuracy.

| Method | Identification [%] | ERR [%] | Database | Comments |
|---|---|---|---|---|
| MIA | 56 | 6.8 | NTIMIT(168) | 50-50 partitioning for training and testing |
| GMM [13] | 69 | 7.2 | NTIMIT(168) | Similar/dissimilar speakers excluded |
| GMM [16] | N/A | 9.6 | NTIMIT(168) | |
| GMM [17] | N/A | 12.4<br>8.8 | NTIMIT(168)<br>NTIMIT(630) | |
| Phoneme GMM [18] | N/A | 15.7 | NTIMIT(438) | Only male speakers used |

**Table 2**. Comparison of the MIA results with Gaussian Mixture Model (GMM) results using NTIMIT. Note that MIA achieves a competitive EER while scoring below "state-of-the-art" identification rates.




(a) Similarity scores (b) FA vs. FR percentage

**Fig. 3**. Results of MIA-based text independent speaker verification on NTIMIT. (a) Matrix of similarity scores between different signatures. Bright gray stands for high and dark gray for low similarity between signatures. (b) False rejection (FR) versus false acceptance (FA) rate.

### 4.1. MIA-based text-independent speaker verification

The data are partitioned with non-overlapping, nearly rectangular window functions of one second lengths and exponential tails. The input functions are centered. For each person, we extract a voice signature $\mathbf{w}_{\mathrm{MIA}}^{(p)}$. Thereafter, each extracted signature is down-sampled to 128 points. The mean signature is subtracted from all signatures to focus on the evaluation of differences. The Euclidean distance between the test and training signatures is used as a measure of similarity. A matrix that represents the similarities between all signatures in the database is illustrated in Fig. 3(a).

The false acceptance rate (FA) versus false rejection rate (FR) is computed in an exhaustive test. For various thresholds, their values are illustrated in Fig. 3(b). The equal error rate (EER), where FA equals FR, is used to compare results in Table 2. The EER of this MIA-based text independent speaker recognition system was 6.8%. The best speaker verification results on the NTIMIT database that we are aware of were published in [13] for a similar experiment. The method uses Gaussian mixture speaker models and results in EER's between 7.19% and 8.68%. Note that, in contrast to the Gaussian mixture model, MIA extracts a signature of 128 samples length per speaker.

## 5. CONCLUSION

We propose Mutual Interdependence Analysis (MIA) for robust extraction of "mutual features". We showed that MIA has a unique solution utilizing its high-dimensional, linearly-independent inputs as basis. The MIA result is equally correlated with all inputs, thus representing the inherent, hidden invariance in the dataset. We demonstrate the effective application of MIA for face identification and text independent speaker verification problems with competitive error rates on challenging data. Current work is investigating further the precise advantages and disadvantages of MIA relative to competitor methods such as PCA, MCA and canonical correlation analysis.

## 6. REFERENCES


[1] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 2nd edition, 2002.

[2] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer-Verlag, 2nd edition, 2006.

[3] L. Xu, E. Oja, and C. Y. Suen, "Modified Hebbian learning for curve and surface fitting," *Neural Networks*, vol. 5, pp. 441–457, 1992.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[5] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[6] H. Claussen, J. Rosca, and R. Damper, "Mutual interdependence analysis," in *Independent Component Analysis and Blind Signal Separation*, Heidelberg, Germany, 2007, pp. 446–453, Springer-Verlag.

[7] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[8] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics (2nd ed. in C): Principles and Practice*, Addison-Wesley Longman Publishing, Boston, MA, 1997.

[9] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[10] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Willey & Sons, New York, 1973.

[11] M. Yang, N. Ahuja, and D. J. Kriegman, "Face recognition using kernel eigenfaces," in *International Conference on Image Processing*, Vancouver, Canada, 2000, vol. 1, pp. 37–40.

[12] C. Hoi and M. R. Lyu, "Robust face recognition using minimax probability machine," in *International Conference on Multimedia and Expo*, Taipei, Taiwan, 2004, pp. 1175–1178.

[13] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[14] W. M. Fisher, G. R. Doddington, K. M. Goudie-Marshall, C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT," CDROM, 1993.

[15] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, Signal Processing and Communications. Marcel Dekker, Inc., 2003.

[16] C. Sanderson, "Speech processing & text-independent automatic person verification," Tech. Rep. 08, IDIAP, Martigny, Switzerland, 2002.

[17] B. Wildermoth and K.K. Paliwal, "GMM based speaker recognition on readily available databases," in *Microelectronic Engineering Research Conference*, Brisbane, Australia, 2003.

[18] D. Gutman and Y. Bistritz, "Speaker verification using phoneme-adapted gaussian mixture models," in *European Signal Processing Conference*, Toulouse, France, 2002, vol. 3, pp. 85–88.