Pattern Recognition 44 (2011) 650–661

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Signature extraction using mutual interdependencies

Heiko Claussen [a,b,*], Justinian Rosca [a], Robert Damper [b]

[a] Siemens Corporation, Corporate Research, 755 College Road East, Princeton, NJ 08540, USA
[b] University of Southampton, School of Electronics and Computer Science, Highfield, Southampton SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 20 March 2009
Received in revised form
25 August 2010
Accepted 18 September 2010

Keywords:
Algorithms
Signal processing
Pattern classification
Signal analysis
Speaker recognition
Face recognition

ABSTRACT

Recently, mutual interdependence analysis (MIA) has been successfully used to extract representations, or "mutual features", accounting for samples in the class. For example, a mutual feature is a face signature under varying illumination conditions or a speaker signature under varying channel conditions. A mutual feature is a linear regression that is equally correlated with all samples of the input class. Previous work discussed two equivalent definitions of this problem and a generalization of its solution called generalized MIA (GMIA). Moreover, it showed how mutual features can be computed and employed. This paper uses a parametrized version GMIA(λ) to pursue a deeper understanding of what GMIA features really represent. It defines a generative signal model that is used to interpret GMIA(λ) and visualize its difference to MIA, principal and independent component analysis. Finally, we analyze the effect of λ on the feature extraction performance of GMIA(λ) in two standard pattern recognition problems: illumination-independent face recognition and text-independent speaker verification.

1. Introduction

Statistical pattern recognition methods such as Fisher's linear discriminant analysis (FLDA) [9], canonical correlation analysis (CCA) [16] or ridge regression [25] aim to model or extract the essence of a dataset. The goal is to find a simplified data representation that retains the information that is necessary for subsequent tasks such as classification or prediction. Each method uses a different viewpoint and criteria to find this "optimal" representation. Furthermore, pattern recognition problems implicitly assume that the number of observations is usually much higher than the dimensionality of each observation. This allows one to study characteristics of the distributional observations and design proper discriminant functions for classification. For instance, FLDA is used to reduce the dimensionality of a dataset by projecting data points on a space that maximizes the ratio of the between- and within-class scatter of the training data. In this way, FLDA aims to find a simplified data representation that retains the discriminant characteristics for classification. On the other hand, CCA assumes one common source in two datasets. The dimensionality of the data is reduced by retaining the space that is spanned by pairs of projecting directions in which the datasets are maximally correlated. In contrast, ridge regression finds a linear combination of the inputs that best fits a desired response.

In this paper, we present alternative criteria to find an "optimal" dataset representation. We aim to extract an invariant representation of high-dimensional instances of a single class, where the number of input instances N is smaller than their dimensionality D. An invariant is a property or feature of the input data that does not change within its class. Approaches that have been designed for this purpose are mutual interdependence analysis (MIA) and generalized MIA (GMIA) [4–6]. We revisit both methods in Sections 2 and 3, respectively, and parametrize GMIA with λ, which subsumes MIA for $\lambda=0$. In Section 4, we introduce a generative model for GMIA(λ). On synthetic data, we demonstrate that GMIA(λ) extracts features unlike approaches such as PCA and ICA. Also we show how these features differ from the sample mean. Section 5 evaluates the discriminative quality of GMIA(λ) features for illumination-invariant face recognition on synthetic data. Section 6 analyses the effect of λ on real data for illumination-invariant face recognition and text-independent speaker verification. The document concludes with a summary and directions for future work.

2. Mutual interdependence analysis (MIA)

MIA was first introduced by the authors in Claussen et al. [4] to uniquely represent high-dimensional samples of a single class. The understanding of how this problem can be succinctly and

* Corresponding author at: Siemens Corporation, Corporate Research, 755 College Road East, Princeton, NJ 08540, USA.
E-mail addresses: Heiko.claussen@siemens.com (H. Claussen), Justinian.rosca@siemens.com (J. Rosca), rid@ecs.soton.ac.uk (R. Damper).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.012

elegantly stated has been evolved and generalized [6]. In this section we present an up to date statement of MIA.

2.1. Scatter-based definition of MIA

Throughout this paper, $\mathbf{x}_i^{(p)} \in \mathbb{R}^D$ denotes the ith input vector, $i=1 \ldots N^{(p)}$ in class p. Furthermore, we use $\mathbf{X}^{(p)} \subseteq \mathbf{X}$ to represent a matrix with columns $\mathbf{x}_i^{(p)}$ and $\mathbf{X}$ to denote the matrix with columns $\mathbf{x}_i$ of all K classes. Moreover, $\boldsymbol{\mu} = (1/N)\sum_{i=1}^{N} \mathbf{x}_i$, $\underline{\mathbf{1}}$ is a vector of ones and $\mathbf{I}$ represents the identity matrix.

Assume that we wish to find a class representation $\mathbf{w}^{(p)}$ of high-dimensional data vectors $\mathbf{x}_i^{(p)}$ $(D \geq N^{(p)})$. A common first step is to select features so as to reduce the dimensionality of the data. However, because of possible loss of information, this preprocessing is not always desirable. Therefore, we aim to find a class representation of similar or same dimensionality as the inputs.

The quality of such a representation can be evaluated by its correlation with the class instances. Our intuition is that a superior class representation is highly correlated and also has a small variance of the correlations over all instances in the class. The former condition ensures that most of the signal energy in the samples is captured. The latter condition is indicative of membership in a single class. Note that only vectors in the span of the class instances contribute to the cross-correlation value. Therefore, in the absence of prior knowledge, it is reasonable to constrain the search for a class representation $\mathbf{w}$ to the span of the training vectors $\mathbf{w} = \mathbf{X}^{(p)} \cdot \mathbf{c}$, where $\mathbf{c} \in \mathbb{R}^{N^{(p)}}$. This problem definition is the motivation for the MIA criterion proposed in Claussen et al. [4].

The MIA representation for class p is defined as a direction $\mathbf{w}_{\mathrm{MIA}}^{(p)} \in \mathbb{R}^D$ that minimizes the projection scatter of the class p inputs, under the linearity constraint to be in the span of $\mathbf{X}^{(p)}$:

$$\mathbf{w}_{\mathrm{MIA}}^{(p)} = \underset{\mathbf{w},\mathbf{w}=\mathbf{X}^{(p)}\cdot\mathbf{c}}{\operatorname{argmin}} \left(\mathbf{w}^T \cdot (\mathbf{X}^{(p)} - \boldsymbol{\mu}^{(p)} \cdot \underline{\mathbf{1}}^T) \cdot (\mathbf{X}^{(p)} - \boldsymbol{\mu}^{(p)} \cdot \underline{\mathbf{1}}^T)^T \cdot \mathbf{w}\right) \tag{1}$$

Note that the original space of the inputs spans the space of the mean subtracted inputs plus possibly one additional dimension. Indeed, the mean subtracted inputs, which are linear combinations of the original inputs, sum to zero. Mean subtraction cancels linear independence resulting in a 1D span reduction. The following two theorems describe the MIA solution.

Theorem 2.1. *The minimum of the criterion in Eq.* (1) *is zero if the inputs* $\mathbf{x}_i$ *are linearly independent.*

If inputs are linearly independent and span a space of dimensionality $N \leq D$, then the subspace of the mean subtracted inputs in Eq. (1) has dimensionality $N-1$. There exists an additional dimension in $\mathbb{R}^N$, orthogonal to this subspace. Thus, the scatter of the mean subtracted inputs can be made zero. The existence of a solution where the criterion in Eq. (1) becomes zero is indicative of an invariance property of the data.

Theorem 2.2. *The solution of Eq.* (1) *is unique (up to scaling) if the inputs* $\mathbf{x}_i$ *are linearly independent.*

By solving in the span of the original rather than mean subtracted inputs, a closed form solution of Eq. (1) can be found [4]:

$$\mathbf{w}_{\mathrm{MIA}}^{(p)} = \zeta \mathbf{X}^{(p)} \cdot (\mathbf{X}^{(p)^T} \cdot \mathbf{X}^{(p)})^{-1} \cdot \underline{\mathbf{1}} \quad \text{where } \zeta \text{ is a constant} \tag{2}$$

Consider that $(\mathbf{X}^{(p)^T} \cdot \mathbf{X}^{(p)})^{-1} \cdot \underline{\mathbf{1}}$ is a column vector. The structure of the solution shows that $\mathbf{w}$ is a data-dependent transformation representing a linear combination of the input observations.

The mathematical structure of this MIA solution has a striking similarity with linear regression. Indeed this result can be obtained as follows. Let us assume the regression problem $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta}$. We are looking for $\boldsymbol{\beta}$ such that the unknown regression $\mathbf{y}$ is equally correlated with all inputs $\mathbf{X}^T \cdot \mathbf{y} = \underline{\mathbf{1}}$. It can be shown that the solution to this problem is given by Eq. (2) with $\zeta = 1$ and $\mathbf{y} = \mathbf{w}$. In Section 3, we return to the discussion of similarities between the two problems. Eq. (2) computes a unique representation we call MIA with the property of invariant correlation with all samples in the input. This uniqueness indicates that MIA captures an inherent property of the input data.

2.2. CCA-based definition of MIA

The minimum variance criterion is also used in other data analysis approaches such as FLDA. However, this theory does not apply when analyzing data from one class. This motivated the comparison with CCA as a generalization of FLDA, and the discovery of an equivalent, CCA-based formulation of the MIA problem. We revisit this new definition following and extending Claussen et al. [6]. First, we review CCA and its FLDA equivalent formulation. Thereafter, we extend this formulation to address the MIA problem.

If a common source $\mathbf{s} \in \mathbb{R}^N$ influences two datasets $\mathbf{X} \in \mathbb{R}^{D \times N}$ and $\mathbf{Z} \in \mathbb{R}^{K \times N}$, of possibly different dimensionality, CCA is used to extract this inherent similarity. The goal of CCA is to find two vectors to project the datasets such that their projection lengths are maximally correlated. Let $\mathbf{C}_{\mathbf{XZ}}$ denote the cross covariance matrix between the datasets $\mathbf{X}$ and $\mathbf{Z}$. Then the CCA problem is given by maximization of the objective function

$$J(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^T \cdot \mathbf{C}_{\mathbf{XZ}} \cdot \mathbf{b}}{\sqrt{\mathbf{a}^T \cdot \mathbf{C}_{\mathbf{XX}} \cdot \mathbf{a}} \cdot \sqrt{\mathbf{b}^T \cdot \mathbf{C}_{\mathbf{ZZ}} \cdot \mathbf{b}}} \tag{3}$$

over the vectors $\mathbf{a}$ and $\mathbf{b}$. The CCA problem can be solved by a singular value decomposition (SVD) of $\mathbf{C}_{\mathbf{XX}}^{-1/2} \cdot \mathbf{C}_{\mathbf{XZ}} \cdot \mathbf{C}_{\mathbf{ZZ}}^{-1/2}$ [19]. The solution is obtained by solving the two eigenvector problems:

$$\left(\mathbf{C}_{\mathbf{XX}}^{-1/2} \cdot \mathbf{C}_{\mathbf{XZ}} \cdot \mathbf{C}_{\mathbf{ZZ}}^{-1} \cdot \mathbf{C}_{\mathbf{ZX}} \cdot \mathbf{C}_{\mathbf{XX}}^{-1/2}\right) \cdot \mathbf{a} = \lambda \mathbf{a} \tag{4}$$

and

$$\left(\mathbf{C}_{\mathbf{ZZ}}^{-1/2} \cdot \mathbf{C}_{\mathbf{ZX}} \cdot \mathbf{C}_{\mathbf{XX}}^{-1} \cdot \mathbf{C}_{\mathbf{XZ}} \cdot \mathbf{C}_{\mathbf{ZZ}}^{-1/2}\right) \cdot \mathbf{b} = \lambda \mathbf{b} \tag{5}$$

We hypothesize that the maximally correlated projections $\mathbf{X}^T \cdot \mathbf{a}$ and $\mathbf{Z}^T \cdot \mathbf{b}$ represent an estimate of the common source.

Canonical correlation analysis can be used to extract classification relevant information from a set of inputs. Indeed, let $\mathbf{X}$ be the union of all data points and $\mathbf{Z}$ the table of corresponding class memberships, $k=1,\ldots,K$ and $i=1,\ldots,N$:

$$\mathbf{Z}_{ki} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbf{X}^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

All classification relevant information is represented by this classification table. Therefore, this information is retained in those input components of $\mathbf{X}$ that originate from a common virtual source with the classification table. It has been shown [2,19,14,1] that this special CCA approach is equivalent to FLDA.

CCA with $\mathbf{Z}$ given by the class membership can be modified to extract a representation of inputs from a single class, similar to MIA. One possible interpretation of CCA is from the point of view of the cosine angle between the (non-mean-subtracted) vectors $\mathbf{a}^T \cdot \mathbf{X}$ and $\mathbf{Z}^T \cdot \mathbf{b}$. The aim is to find a vector pair that results in a minimum angle. We will use a modified CCA criterion (MCCA) as follows. First, consider the original inputs rather than the mean subtracted covariance matrices; second, the class membership table for data from a single class collapses to a vector and $\mathbf{b}$ to a scalar, therefore $\mathbf{Z}^T \cdot \mathbf{b} = \underline{\mathbf{1}} \cdot \mathbf{b}$. Thus, criterion Eq. (3) becomes

independent of $\mathbf{b}$ resulting in

$$\hat{\mathbf{a}}_{\mathrm{MCCA}} = \underset{\mathbf{a}}{\operatorname{argmax}} \frac{\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}}}{\sqrt{\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a}}} \tag{6}$$

Theorem 2.3. *The MCCA criterion in Eq.* (6) *has MIA of Eq.* (2) *as solution.*

This criterion is maximized when the correlation of $\mathbf{a}$ with all inputs $\mathbf{x}_i^{(p)}$ is as uniform as possible. A solution to this problem can be found by

$$\frac{\partial J(\mathbf{a})}{\partial \mathbf{a}} = \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}} - \mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}} \cdot (\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a})^{-1} \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a} = 0 \tag{7}$$

Therefore, $\zeta \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}} = \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a}$ with $\zeta = \mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a} / (\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}})$. Furthermore,

$$\begin{aligned} \mathbf{a} &= \zeta (\mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T})^{-1} \cdot \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}} \\ &= \zeta (\mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T})^{-1} \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{X}^{(p)} \cdot (\mathbf{X}^{(p)^T} \cdot \mathbf{X}^{(p)})^{-1} \cdot \underline{\mathbf{1}} \\ &= \zeta \mathbf{X}^{(p)} \cdot (\mathbf{X}^{(p)^T} \cdot \mathbf{X}^{(p)})^{-1} \cdot \underline{\mathbf{1}} \end{aligned} \tag{8}$$

Note that ζ is a scalar that results in scale-independent solutions. As can easily be seen, the solution in Eq. (8) is identical to the MIA solution in Eq. (2). In the following, we rename $\hat{\mathbf{a}}_{\mathrm{MCCA}}$ to $\hat{\mathbf{a}}_{\mathrm{MIA}}$.

We review the properties of the MIA formulation in Eq. (6) [6]:

Corollary 2.4. *The MIA problem has no defined solution if the inputs are zero mean, i.e., if* $\mathbf{X}^{(p)} \cdot \underline{\mathbf{1}} = \underline{\mathbf{0}}$.

This is obvious from Eq. (6).

Corollary 2.5. *Any combination* $\hat{\mathbf{a}}_{\mathrm{MIA}} + \mathbf{b}$ *with* $\mathbf{b}$ *in the nullspace of* $\mathbf{X}^{(p)}$ *is also a solution to Eq.* (6).

This means that only the component of $\mathbf{a}$ that is in the span of $\mathbf{X}^{(p)}$ contributes to the criterion in Eq. (6).

Corollary 2.6. *The solution of Eq.* (6) *is not unique if the* $N^{(p)}$ *inputs* $\mathbf{X}^{(p)}$ *do not span the D-dimensional space* $\mathbb{R}^D$.

This follows from Corollary 2.5. A unique solution can be found by further constraining Eq. (6). One such constraint is that $\mathbf{a}$ be a linear combination of the inputs $\mathbf{X}^{(p)}$:

$$\hat{\mathbf{a}}_{\mathrm{MIA}} = \underset{\mathbf{a}, \mathbf{a} = \mathbf{X}^{(p)} \cdot \mathbf{c}}{\operatorname{argmax}} \frac{\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \underline{\mathbf{1}}}{\sqrt{\mathbf{a}^T \cdot \mathbf{X}^{(p)} \cdot \mathbf{X}^{(p)^T} \cdot \mathbf{a}}} \tag{9}$$

Corollary 2.7. *The MIA solution reduces to the mean of the inputs in the special case when the covariance matrix* $\mathbf{C_{XX}}$ *has one eigenvalue* λ *of multiplicity D, i.e.,* $\mathbf{C_{XX}} = \lambda \mathbf{I}$.

Indeed, Eq. (9) can be rewritten as

$$\hat{\mathbf{a}}_{\mathrm{MIA}} = \underset{\mathbf{a}, \mathbf{a} = \mathbf{X}^{(p)} \cdot \mathbf{c}}{\operatorname{argmax}} \frac{\mathbf{a}^T \cdot \boldsymbol{\mu}^{(p)}}{\sqrt{\mathbf{a}^T \cdot \mathbf{C}_{\mathbf{XX}}^{(p)} \cdot \mathbf{a} + (\mathbf{a}^T \cdot \boldsymbol{\mu}^{(p)})^2}} \tag{10}$$

After normalizing with $\mathbf{a} = \mathbf{X}^{(p)} \cdot \mathbf{c} / \|\mathbf{X}^{(p)} \cdot \mathbf{c}\|$ and using the spectral decomposition theorem ([20], p. 317), it can be shown that $\mathbf{a}^T \cdot \mathbf{C}_{\mathbf{XX}}^{(p)} \cdot \mathbf{a}$ is invariant to $\mathbf{a}$ given equal eigenvalues of $\mathbf{C}_{\mathbf{XX}}^{(p)}$. The function under Eq. (10) is monotonically increasing in $\mathbf{a}^T \cdot \boldsymbol{\mu}^{(p)}$. Therefore, the optimum is obtained when $\mathbf{a}^T \cdot \boldsymbol{\mu}^{(p)} / \|\mathbf{a}\|$ is maximum resulting in $\hat{\mathbf{a}}_{\mathrm{MIA}} = \boldsymbol{\mu}^{(p)}$.

Sample data from one class results in a unique direction that is a characteristic feature of the data. MIA may capture information that is powerful enough to distinguish instances from different classes.

3. Generalized mutual interdependence analysis (GMIA)

Real world data are generally noisy. Claussen et al. [6] analyzed MIA's sensitivity to noise and extended its model to capture this effect in the data with a Bayesian MIA interpretation. This section provides a complete presentation of this analysis. It reviews the Bayesian general linear model, shows assumptions that distinguish MIA from linear regression, and generalizes MIA for utilization of uncertainties and prior knowledge.

3.1. The Bayesian general linear model

In the following, let $\mathbf{y} \in \mathbb{R}^D$, $\mathbf{X} \in \mathbb{R}^{D \times N}$, $\mathbf{n} \in \mathbb{R}^D$ and $\boldsymbol{\beta} \in \mathbb{R}^N$ represent the observations, the matrix of known inputs, a noise vector and the weight parameters of interest, respectively. The general linear model is defined as

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{n} \tag{11}$$

The Bayesian estimation finds the expectation of the random variable $\boldsymbol{\beta}$ given its a priori known or estimated distribution, the signal model and observed data $\mathbf{y}$. As discussed in Kay ([18], p. 325), the expected value $E\{\boldsymbol{\beta}|\mathbf{y}\}$ from the conditional probability $p(\boldsymbol{\beta}|\mathbf{y})$ can be introduced as a biased estimator of $\boldsymbol{\beta}$. If $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$ and $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \mathbf{C}_\beta)$ are independent Gaussian variables, the joint probability density function (PDF) $p(\mathbf{y}, \boldsymbol{\beta})$ as well as the conditional PDF $p(\boldsymbol{\beta}|\mathbf{y})$ are Gaussian. Considering our prior assumptions, $p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{C_y})$ and $p(\mathbf{y}, \boldsymbol{\beta}) = \mathcal{N}([\begin{smallmatrix}\boldsymbol{\mu}_\mathbf{y} \\ \boldsymbol{\mu}_\beta\end{smallmatrix}], [\begin{smallmatrix}\mathbf{C_y} & \mathbf{C}_{\mathbf{y}\beta} \\ \mathbf{C}_{\beta\mathbf{y}} & \mathbf{C}_\beta\end{smallmatrix}])$. Using this, the conditional probability $p(\boldsymbol{\beta}|\mathbf{y})$ can be computed as follows:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}) &= \frac{p(\mathbf{y}, \boldsymbol{\beta})}{p(\mathbf{y})} \\ &= \frac{\dfrac{1}{\sqrt{(2\pi)^{D+N} \left| \begin{bmatrix} \mathbf{C_y} & \mathbf{C}_{\mathbf{y}\beta} \\ \mathbf{C}_{\beta\mathbf{y}} & \mathbf{C}_\beta \end{bmatrix} \right|}} \exp\left[-\frac{1}{2} \begin{bmatrix} \mathbf{y} - \boldsymbol{\mu}_\mathbf{y} \\ \boldsymbol{\beta} - \boldsymbol{\mu}_\beta \end{bmatrix}^T \cdot \begin{bmatrix} \mathbf{C_y} & \mathbf{C}_{\mathbf{y}\beta} \\ \mathbf{C}_{\beta\mathbf{y}} & \mathbf{C}_\beta \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{y} - \boldsymbol{\mu}_\mathbf{y} \\ \boldsymbol{\beta} - \boldsymbol{\mu}_\beta \end{bmatrix} \right]}{\dfrac{1}{\sqrt{(2\pi)^D |\mathbf{C_y}|}} \exp\left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^T \cdot \mathbf{C_y}^{-1} \cdot (\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}) \right]} \end{aligned}$$

After a few mathematical transformations ([18], p. 326), the posterior expectation of $\boldsymbol{\beta}$ given $\mathbf{y}$ is found to become

$$E\{\boldsymbol{\beta}|\mathbf{y}\} = \boldsymbol{\mu}_\beta + \mathbf{C}_\beta \cdot \mathbf{X}^T \cdot (\mathbf{X} \cdot \mathbf{C}_\beta \cdot \mathbf{X}^T + \mathbf{C_n})^{-1} \cdot (\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\mu}_\beta) \tag{12}$$

$$= \boldsymbol{\mu}_\beta + (\mathbf{X}^T \cdot \mathbf{C_n}^{-1} \cdot \mathbf{X} + \mathbf{C}_\beta^{-1})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{C_n}^{-1} \cdot (\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\mu}_\beta) \tag{13}$$

Ridge regression is a generalization of the least squares solution to the regression problem. It follows from Eq. (13) by further assuming $\boldsymbol{\mu}_\beta = \underline{\mathbf{0}}$, $\mathbf{C}_\beta = \sigma_\beta^2 \mathbf{I}$ and $\mathbf{C_n} = \sigma_\mathbf{n}^2 \mathbf{I}$:

$$\boldsymbol{\beta}_{\mathrm{RIDGE}} = \left(\mathbf{X}^T \cdot \mathbf{X} + \frac{\sigma_\mathbf{n}^2}{\sigma_\beta^2} \mathbf{I} \right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} \tag{14}$$

Eq. (14) is useful when $\mathbf{X}^T \cdot \mathbf{X}$ is not full rank or when we have numerical instability in the computation of the inverse. During training, ridge regression assumes an availability of the desired output $\mathbf{y}$ to aid the estimation of a weighting vector $\boldsymbol{\beta}$. Thereafter, $\boldsymbol{\beta}$ is used to predict future outcomes of $\mathbf{y}$.

3.2. A Bayesian view on MIA

Next, we discuss the Bayesian interpretation of MIA to account for uncertainties in the inputs. Consider the following model:

$$\mathbf{r} = \mathbf{X}^T \cdot \mathbf{w} + \mathbf{n} \tag{15}$$

The intended meaning of $\mathbf{r}$ is the vector of observed projections of inputs $\mathbf{x}$ on $\mathbf{w}$, while $\mathbf{n}$ is measurement noise, e.g., $\mathbf{n} \sim \mathcal{N}(\underline{\mathbf{0}}, \mathbf{C_n})$.

We assume $\mathbf{w}$ to be a random variable. Our goal is to estimate $\mathbf{w}\sim\mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\mathbf{C}_\mathbf{w})$ assuming that $\mathbf{w}$ and $\mathbf{n}$ are statistically independent. Ideally, the data $\mathbf{r}=\zeta\underline{\mathbf{1}}$ follow from the variance minimization objective if no noise is present and the variance of projections is zero, which is exactly the MIA criterion (as expressed in Theorem 2.1). We define a generalized MIA criterion (GMIA) applying the derivation for Eqs. (12) and (13) to model Eq. (15):

$$\mathbf{w}_{\mathrm{GMIA}}=\boldsymbol{\mu}_\mathbf{w}+\mathbf{C}_\mathbf{w}\cdot\mathbf{X}\cdot(\mathbf{X}^T\cdot\mathbf{C}_\mathbf{w}\cdot\mathbf{X}+\mathbf{C}_\mathbf{n})^{-1}\cdot(\mathbf{r}-\mathbf{X}^T\cdot\boldsymbol{\mu}_\mathbf{w}) \tag{16}$$

$$=\boldsymbol{\mu}_\mathbf{w}+(\mathbf{X}\cdot\mathbf{C}_\mathbf{n}^{-1}\cdot\mathbf{X}^T+\mathbf{C}_\mathbf{w}^{-1})^{-1}\cdot\mathbf{X}\cdot\mathbf{C}_\mathbf{n}^{-1}\cdot(\mathbf{r}-\mathbf{X}^T\cdot\boldsymbol{\mu}_\mathbf{w}) \tag{17}$$

The GMIA solution, interpreted as a direction in a high-dimensional space $\mathbb{R}^D$, aims to minimize the difference between the observed projections $\mathbf{r}$ considering prior information on the noise distribution. It is an update of the prior mean $\boldsymbol{\mu}_\mathbf{w}$ by the current misfit $\mathbf{r}-\mathbf{X}^T\cdot\boldsymbol{\mu}_\mathbf{w}$ times an input data $\mathbf{X}$ and prior covariance dependent weighting matrix. Eqs. (16) and (17) suggest various properties of MIA and will enable us to analyze the relationship between the mean of the dataset and the solution $\mathbf{w}_{\mathrm{GMIA}}$. In general, it is desirable that the MIA representation is robust to small variations in $\mathbf{X}$ (e.g., due to noise). Eq. (16) indicates that small variations in $\mathbf{X}$ do not have a large effect on the GMIA result. Indeed $\mathbf{w}_{\mathrm{GMIA}}$ is an invariant property of the class of inputs. Furthermore, Eqs. (16) and (17) allow us to integrate additional prior knowledge such as smoothness of $\mathbf{w}_{\mathrm{GMIA}}$ through the prior $\mathbf{C}_\mathbf{w}$, correlation of consecutive instances $\mathbf{x}_i$ through the prior $\mathbf{C}_\mathbf{n}$, etc. Moreover, we can use the GMIA formulation to define an iterative procedure that tackles datasets with large N and D. In such cases it might be unfeasible to compute the matrix inverse. By using subsets of the input data, one can iteratively compute $\boldsymbol{\mu}_\mathbf{w}$ as a GMIA representation of the whole dataset from smaller subsets.

Throughout the remainder of the document, the GMIA parameters are $\mathbf{C}_\mathbf{w}=\mathbf{I}$, $\mathbf{C}_\mathbf{n}=\lambda\mathbf{I}$ and $\boldsymbol{\mu}_\mathbf{w}=\underline{\mathbf{0}}$. We refer to this parameterization by

$$\mathrm{GMIA}(\lambda)=\zeta\mathbf{X}\cdot(\mathbf{X}^T\cdot\mathbf{X}+\lambda\mathbf{I})^{-1}\cdot\underline{\mathbf{1}}\quad\text{where }\zeta\text{ is a constant}$$

When $\lambda\to\infty$, the GMIA solution represents the mean of the inputs. Indeed, the inverse $(\mathbf{X}^T\cdot\mathbf{X}+\lambda\mathbf{I})^{-1}\to(1/\lambda)\mathbf{I}$ simplifying the solution to $\mathbf{w}_{\mathrm{GMIA}}\to(\zeta/\lambda)\mathbf{X}\cdot\underline{\mathbf{1}}$. Furthermore, MIA (solution to Eq. (2)) is equivalent to GMIA(λ) when $\lambda=0$. In the rest of the paper, we denote MIA by GMIA(0) to emphasize their common theoretical foundation.

4. Generative signal model for GMIA

So far we have discussed two equivalent definitions of GMIA(0) and a generalization of the criterion by following Claussen et al. [4,6]. Furthermore, Claussen et al. [4–6] show how a mutual feature can be computed using mutual interdependencies in data (sounds and images) of the same class. Nonetheless, we aim for a deeper understanding of what GMIA features really represent, which lacks in previously published materials. This section defines a generative signal model that will allow us to create synthetic data in order to interpret GMIA and visualize its differences to GMIA(0), principal component analysis (PCA) [21], independent component analysis (ICA) [17] and the sample mean. This way we can compare the feature extraction results to the true feature desired.

Assume the following generative model for input data $\mathbf{x}$:

$$\begin{aligned}\mathbf{x}_1&=\alpha_1\mathbf{s}+\mathbf{f}_1+\mathbf{n}_1\\\mathbf{x}_2&=\alpha_2\mathbf{s}+\mathbf{f}_2+\mathbf{n}_2\\&\vdots\\\mathbf{x}_N&=\alpha_N\mathbf{s}+\mathbf{f}_N+\mathbf{n}_N\end{aligned} \tag{18}$$

where $\mathbf{s}$ is a common, invariant component or feature we aim to extract from the inputs, α_i, $i=1,\ldots,N$ are scalars (typically all close to 1), $\mathbf{f}_i$, $i=1,\ldots,N$ are combinations of basis functions from a given orthogonal dictionary such that any two are orthogonal and $\mathbf{n}_i$, $i=1,\ldots,N$ are Gaussian noises. We will show that GMIA estimates the invariant component $\mathbf{s}$, inherent in the inputs $\mathbf{x}$.

Let us make this model precise. As before, D and N denote the dimensionality and the number of observations. Additionally, K is the size of a dictionary $\mathbf{B}$ of orthogonal basis functions. Let $\mathbf{B}=[\mathbf{b}_1,\ldots,\mathbf{b}_K]$ with $\mathbf{b}_k\in\mathbb{R}^D$. Each basis vector $\mathbf{b}_k$ is generated as a weighted mixture of maximally J elements of the Fourier basis which are not reused ensuring orthogonality of $\mathbf{B}$. The actual number of mixed elements is chosen uniformly at random, $J_k\in\mathbb{N}$ and $J_k\sim\mathcal{U}(1,J)$. For $\mathbf{b}_k$, the weights of each Fourier basis element i are given by $w_{jk}\sim\mathcal{N}(0,1)$, $j=1,\ldots J_k$. For $i=1,\ldots,D$ (analogous to a time dimension) the basis functions are generated as

$$b_k(i)=\frac{\sum_{j=1}^{J_k}w_{jk}\sin\left(\frac{2\pi i\alpha_{jk}}{D}+\beta_{jk}\frac{\pi}{2}\right)}{\sqrt{\frac{D}{2}\sum_{j=1}^{J_k}w_{jk}^2}}$$

with

$$\alpha_{jk}\in\left[1,\ldots,\frac{D}{2}\right],\quad\beta_{jk}\in[0,1],\quad[\alpha_{jk},\beta_{jk}]\neq[\alpha_{lp},\beta_{lp}],\ \forall j\neq l\quad\text{or}\quad k\neq p$$

In the following, one of the basis functions $\mathbf{b}_k$ is randomly selected to be the common component $\mathbf{s}\in[\mathbf{b}_1,\ldots,\mathbf{b}_K]$. The common component is excluded from the basis used to generate uncorrelated additive functions $\mathbf{f}_n$, $n=1,\ldots,N$. Thus only $K-1$ basis functions can be combined to generate the additive functions $\mathbf{f}_n\in\mathbb{R}^D$. The actual number of basis functions J_n is randomly chosen, i.e., similarly to J_k, with $J=K-1$. The randomly correlated additive components are given by

$$f_n(i)=\frac{\sum_{j=1}^{J_n}w_{jn}c_{jn}(i)}{\sqrt{\sum_{j=1}^{J_n}w_{jn}^2}}$$

with

$$\mathbf{c}_{jn}\in[\mathbf{b}_1,\ldots,\mathbf{b}_K],\quad\mathbf{c}_{jn}\neq\mathbf{s},\ \forall j,n,\quad\mathbf{c}_{jn}\neq\mathbf{c}_{lp},\ \forall j\neq l\quad\text{and}\quad n=p$$

Note that $\|\mathbf{s}\|=\|\mathbf{f}_n\|=\|\mathbf{n}_n\|=1,\forall n=1,\ldots,N$. To control the mean and variance of the norms of common, additive and noise component in the inputs, each component is multiplied by the random variable $a_1\sim\mathcal{N}(m_1,\sigma_1^2)$, $a_2\sim\mathcal{N}(m_2,\sigma_2^2)$ and $a_3\sim\mathcal{N}(m_3,\sigma_3^2)$, respectively. Finally, the synthetic inputs are generated as

$$\mathbf{x}_n=a_1\mathbf{s}+a_2\mathbf{f}_n+a_3\mathbf{n}_n \tag{19}$$

with $\sum_{i=1}^{D}x_n(i)\approx0$. The parameters of the artificial data generation model are chosen as $D=1000$, $K=10$, $J=10$ and $N=20$. The parameters of the distributions for a_1, a_2 and a_3 are dependent on the particular experiment and are defined correspondingly.

The GMIA solution is compared in Fig. 1 (rightmost plot in top row) to the mean of the inputs as well as PCA and ICA results. Note that the parametrization of GMIA(λ) represents the variance of the noise in model (18). The mixing model parameters are chosen as $m_1=1$, $m_2=10$, $m_3=0$, $\sigma_1=0.05$, $\sigma_2=0.05$ and $\sigma_3=0.05$.

We hand selected PC10, the 10th principal component and IC1, the first independent component, due to their maximal correlation with the common component. Over all compared methods, GMIA extracts a signature that is maximally correlated to $\mathbf{s}$. All other methods fail to extract a signature as similar to the common component as GMIA.

We now analyze and compare in more detail GMIA(0), GMIA(λ) and the sample mean, by representing graphically results in a large number of randomly created synthetic problems, matching

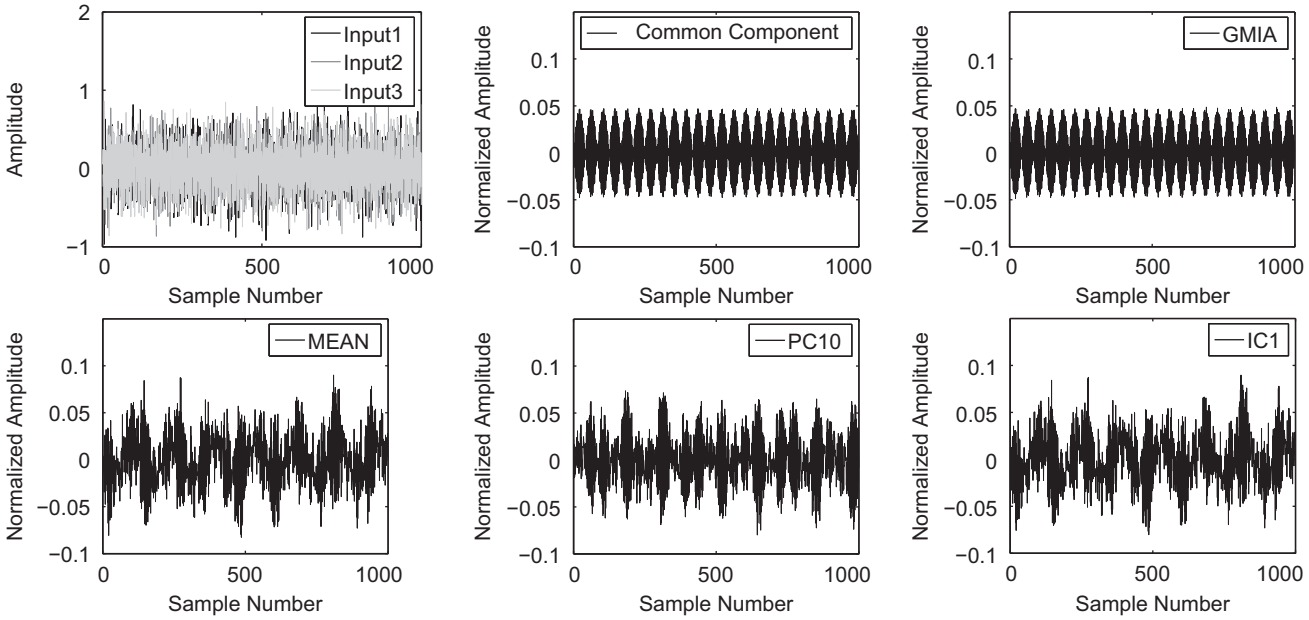


Fig. 1. Comparison of results using various ubiquitous signal processing methods. Top left plot shows, for simplicity, only the first three inputs. The plots of principal and independent component analysis show particular components that maximally correlate with the common component **s**. The GMIA solution turns out to represent the common component, as it is maximally correlated to it.

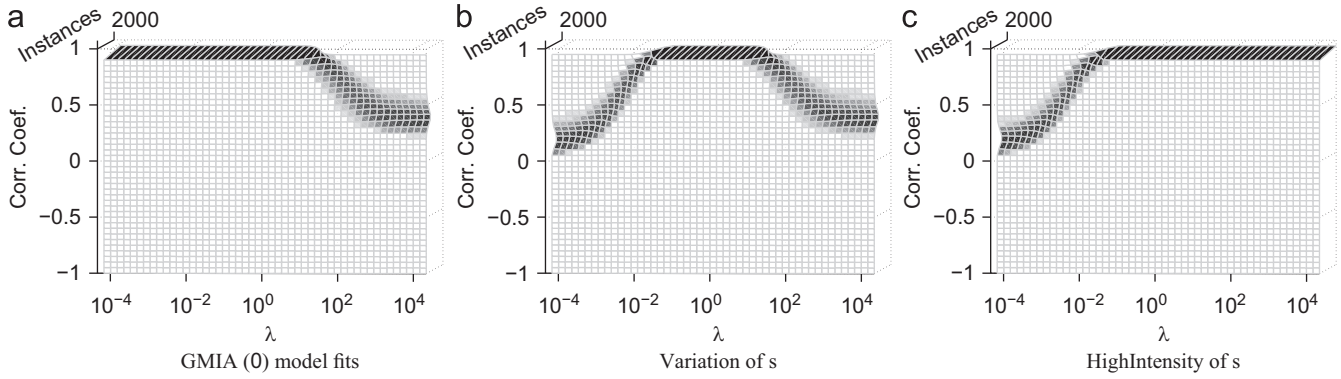


Fig. 2. Statistical behavior of GMIA. Correlation of $\mathbf{w}_{\mathrm{GMIA}}$ and **s** for various λ values. Right vertical regions ($\lambda \to 10^4$) correspond to $\mathbf{w}_{\mathrm{GMIA}} \approx \boldsymbol{\mu}$, the mean of the inputs. (a) The common component intensity is invariant over the inputs and contributes little to their intensities. $\mathbf{w}_{\mathrm{GMIA}(0)}$ best represents the common component. (b) The common component intensity varies over the inputs with $\sigma_1 = 0.05$ and contributes little to their intensities. In this case, GMIA is preferable to GMIA(0) and the mean to learn a feature $\mathbf{w}_{\mathrm{GMIA}}$ that is best correlated with the common component. (c) The common component represents most of the energy in the inputs. In this case, the mean best represents the common component.

model (18). Each point in Fig. 2 represents an experiment for a given value of λ (x-axis). The y-axis indicates the correlation of the GMIA solution with **s**, the true common component. The intensity of the point represents the number of experiments, in a series of random experiments, where we obtain this specific correlation value for the given λ. Overall, we performed 1000 random experiments with randomly generated inputs using various values of λ.

For all test cases in Fig. 2, the weight of the additive noise is chosen as $a_3 \sim \mathcal{N}(0,0.0025)$. We experiment with three cases: (a) inputs contain equally a common component; (b) inputs contain approximately a common component; (c) inputs are approximately equal.

In Fig. 2(a), the remaining mixing model parameters are chosen as $m_1=1$, $m_2=10$, $\sigma_1=0$ and $\sigma_2=0.05$. This situation fits the GMIA(0) assumption of an equally present component with an energy one order of magnitude smaller than the residual $\mathbf{f}_i+\mathbf{n}_i$. The results show that the common component is best extracted by GMIA(0). In Fig. 2(b), $m_1=1$, $m_2=10$, $\sigma_1=0.05$ and $\sigma_2=0.05$. This situation relaxes the strictly equal presence of the common component. Clearly, the simple GMIA(0) result and the mean do not represent **s**. However, for some λ, GMIA succeeds in extracting the common component. Fig. 2(c) illustrates the case $m_1=10$, $m_2=1$, $\sigma_1=0.05$ and $\sigma_2=0.05$. Here, all inputs are similar to the common component and therefore well represented by a signal plus noise model. The mean of the inputs is a good solution to this problem.

In summary, GMIA(0) can extract an invariant, or mutual feature, **s** from a dataset whenever it fits the model in Eq. (18) and $\alpha_i=1$ for all $i=1,\ldots,N$ inputs. This even holds when the energy of **s** is significantly smaller that the energy of the other additive components in the model. In the more general case of noisy data and $\sigma_1 \neq 0$, the choice of λ will trade-off the expected value and variance of the fit between the feature and the data across experiments. Moreover, we show that the computed feature $\mathbf{w}_{\mathrm{GMIA}}$ is radically different from the mean of the data for cases

like (a) and (b) in Fig. 2. The invariant feature **s** may have a physical interpretation of its own, depending on the problem and it is powerful in determining the class membership, as we will see in Sections 5 and 6 below.

5. Illumination invariant 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. We tested the robustness to illumination scenarios of a GMIA(0)-based mutual face approach in Claussen et al. [5]. In this problem, we have called the presumed common invariant feature "mutual face". However, Section 4 showed that λ is problem dependent for an effective extraction of the common component. Can GMIA(λ), with $\lambda > 0$, extract a more discriminative illumination invariant face representation than GMIA(0)? In the following, we analyze the suitability of GMIA(0) versus GMIA(λ) for illumination invariant face recognition.

Following our generative signal model in Section 4 we define a realistic synthetic model that allows the artificial generation of differently illuminated faces. Thus, a large number of test cases can be generated, which facilitates a statistical analysis of GMIA for face recognition. Let the face be a Lambertian object ([11], p. 723), where the object image has light reflected such that the surface is observed equally bright from different angles of the observer. Then, one can assume a face image $\mathbf{H}$ to be a linear combination of images from an image basis $\mathbf{H}_n$ with $n=1,\ldots,K$ [29]:

$$\mathbf{H} = \sum_{n=1}^{K} \alpha_n \mathbf{H}_n \tag{20}$$

where the α_n's are image weights. An appropriate set of basis images, to study illumination effects, is the YaleB database [12]. This database contains 65 differently illuminated faces from 10 people and for 9 different camera angles to view a face. Each illuminated face image is obtained for a single light source at some unique but distinct position. Here, we use only the frontal face direction but at various light source positions. The frontal illuminated faces are excluded from the basis and used as test images. Moreover, the images with ambient lighting conditions are excluded. The set of basis functions for the first person, *A*, of the YaleB database is illustrated in Fig. 3. Additionally, the test image $\mathbf{H}_0^A$ of this person is shown in Fig. 4(a).

Next, 20 images are synthetically generated as inputs to GMIA. Each of these images is a combination of $J=5$ randomly selected images $\mathbf{H}_i$ from the basis set $\mathbf{H}_n$. The basis images are combined according to Eq. (20) using weights $\alpha \sim \mathcal{U}(0,1)$. To retain the image scaling: $\mathbf{H} = \left(\sum_{i=1}^{J} \alpha_i \mathbf{H}_i\right) / \sum_{i=1}^{J} \alpha_i$.

An "invariant" face signature is extracted to represent each person using GMIA(0). This process, illustrated in Fig. 8 later, is defined as follows. First, images are 2D Fourier transformed and filtered. Thereafter, GMIA is separately computed for rows and columns resulting in $D=250$ and $N=20$. In a final step, GMIA representations for rows and columns are processed by an inverse 2D Fourier transform and added to obtain a face signature of the person. This signature is called a mutual face and is, e.g., denoted $\mathbf{H}^A_{\mathrm{GMIA}(0)}$ for person *A*. Fig. 4(b) illustrates a GMIA(0) representation that is generated using the previously described procedure. Note that GMIA(λ) images $\mathbf{H}^A_{\mathrm{GMIA}(\lambda)}$ are extracted accordingly.

A measure is defined to evaluate the similarity between test and GMIA images for the purpose of face recognition. First, the images are filtered on their boundary. Second, the mean correlation scores of both images are computed separately

Fig. 3. Frontal images of the first person from the Yale face database B excluding the ambient and test image. The test image is illuminated frontally.

for rows (ς_1) and columns (ς_2). A combined score is generated as: $\varsigma = \sqrt{(\varsigma_1^2+\varsigma_2^2)/2}$. Thus, the score is upper-bounded by the value one.

Now we test if GMIA is able to capture illumination invariant facial features and can aid face recognition. Fig. 5 illustrates the results of synthetic GMIA experiments with various illumination conditions. In particular, we show similarity scores between GMIA(λ) representations of 50 randomly generated input sets from person A and the test images from both A and other persons $B \neq A$. GMIA(0) results in an invariant image representation (all 50 scores are almost equal). Note that there is a λ–dependent trade-off between the score value and the variance. For all cases of λ, the person A scores higher than person B. Fig. 5(b) shows the training database from Fig. 3 sorted by the score with the GMIA(0) representation (mutual face) of the same person. The score becomes lower line after line from the top left to the bottom right. The mutual face achieves the highest scores with evenly illuminated images, i.e., where the illumination does not distort the image.

Results indicate that the GMIA(0) feature is more robust to variations in illumination than the one using GMIA(λ) while their discrimination power to other classes appears comparable. Following, we verify on a more realistic face recognition application that GMIA(0) achieves similar results to GMIA(λ).

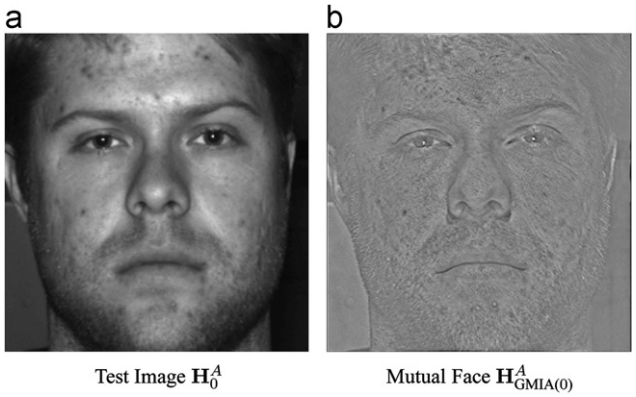


Fig. 4. Images used for testing. (a) Frontal illuminated image of the first person from the Yale face database B. (b) Mutual face that is extracted from 20 randomly generated inputs. Each input is a combination of five randomly selected images of a person.

6. Applications of GMIA

GMIA(0) has already been tested on challenging real world applications such as illumination robust face recognition and text-independent speaker verification [5,6]. In this section, we evaluate the effect of λ on the feature extraction and classification in both domains. First, we repeat the mutual face approach in Claussen et al. [5] on the Yale face database [3] with GMIA(λ) and compare the result for an optimized value of λ to the previous GMIA(0)

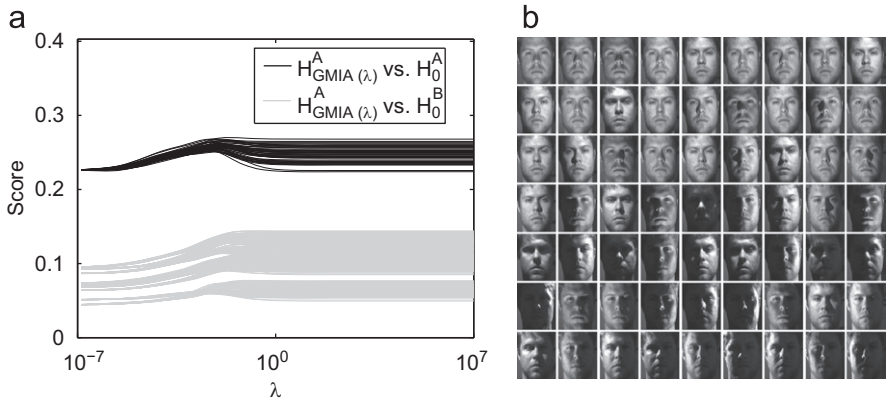


Fig. 5. Results of synthetic GMIA experiments with various illumination conditions. (a) Similarity scores of GMIA(λ) representation and the test image of the same and different people from the YaleB database in 50 random experiments. (b) Images of the YaleB database, ordered from high to low by their similarity score with the mutual face $\mathbf{H}^A_{\mathrm{GMIA}(0)}$. The score becomes lower line after line from the top left to the bottom right.

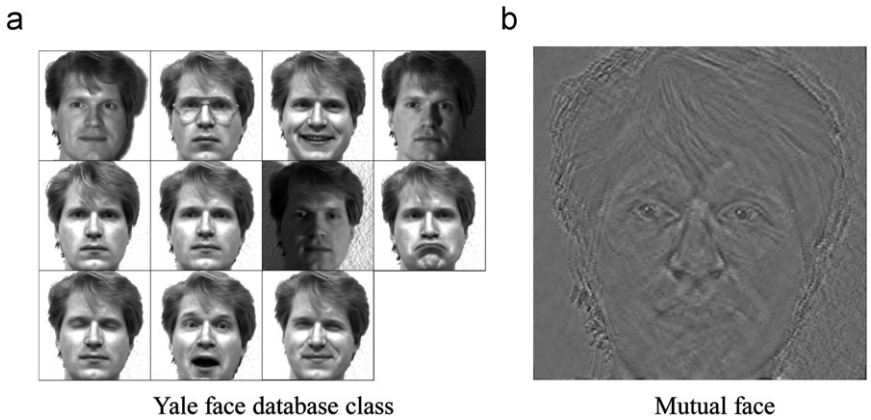


Fig. 6. (a) Image set of one individual in the Yale database. The set contains 11 images of the person taken with various facial expressions and illuminations, with or without glasses. (b) GMIA(0) result, or mutual face estimated from all images of the set.

result. We aim to verify that both the results are similar for this illumination-invariant face recognition approach as indicated in Section 5. Thereafter, we analyze how λ and the data segmentation affect the result of a GMIA-based text-independent speaker verification system. We include preprocessing and evaluation steps to enable the reproducibility of the results.

6.1. Experiments on the Yale database

Next, we compare the performance of GMIA(0) and GMIA(λ) on the Yale database. The difference to the YaleB database, used in Section 5, is that this earlier version includes misalignment, different facial expressions and slight variations in scaling and camera angles. By allowing these variations, the algorithm can be tested in a more realistic face recognition scenario. The image set of one individual is given, for illustration, in Fig. 6(a). As discussed in Foley et al. [11], 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, $\overline{\mathbf{n}}$ and $\overline{\mathbf{l}}$ be ambient/diffuse reflection coefficients, surface normal of the object, and the direction of the light source, respectively. Hence,

$$I = I_a k_a + I_p k_d(\overline{\mathbf{n}} \cdot \overline{\mathbf{l}})$$

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 ([11], see Eq. 16.20).

We claim that GMIA(0) 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. 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 [26] and Fisherface [3] approaches. Most use mean image subtraction for preprocessing, which reduces the image space dimensionality compared to the original image set. Therefore, this step cancels potentially discriminant image information. In contrast, GMIA uses centered images ($\mathbf{x}_i^T \cdot \underline{\mathbf{1}} = 0\ \forall i$) as inputs. Fig. 7 illustrates the difference between a mean-face-subtracted input instance in the Eigenface/Fisherface approach and the centered GMIA input.

The procedure to extract the mutual face from the face set of one person is discussed in Section 5 and illustrated in Fig. 8. Face identification is performed using cropped and centered images. Moreover, the measure of similarity between a test image and the GMIA representation of a person is defined in Section 5 above.

Mutual faces are learned on all but a single test image using the "leave-one-out" method discussed in Duda and Hart ([8], p. 75). In exhaustive leave-one-out tests, the performance of the GMIA(0) and GMIA(λ) based approach are 7.4% and 7.3%, respectively. This verifies the hypothesis in Section 5 that the illumination-invariant face recognition performances of GMIA(0) and GMIA(λ) are similar. Recognition performance for unknown illumination is comparable or beyond various reported results obtained on the same data (Table 1). The GMIA(0) approach can be used to enhance both feature- and appearance-based methods, only requires minimal training due to its closed form solution, and appears insensitive to multiple illumination sources and diffuse light conditions.

6.2. Text-independent speaker verification

In the following, we apply GMIA to the problem of extracting signatures from speech data for the purpose of text-independent speaker verification. The signal quality and background noise are

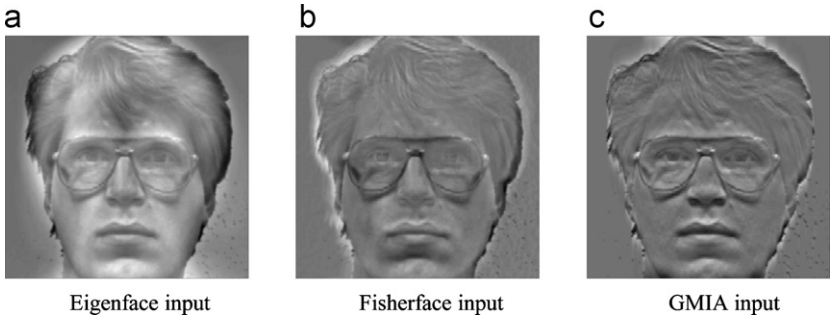


Fig. 7. Examples of training instances used in (a) Eigenfaces, (b) Fisherfaces and (c) GMIA: (a) Mean-subtracted face obtained as difference between a face instance and the mean of all images in the database. (b) Mean-subtracted face obtained as difference between a face instance and the mean image of all instances for the same person. (c) "Centered" face image, obtained by subtraction of the mean column value from each image column.

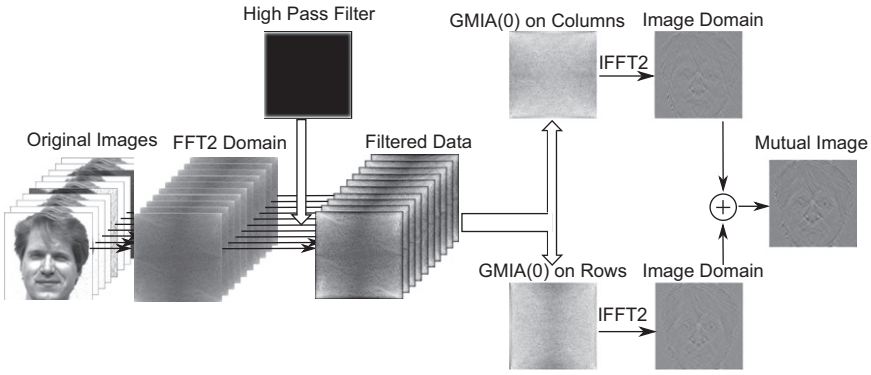


Fig. 8. Extraction process of the mutual image representation.

major challenges in automated speaker verification. For example, telephone signals are nonlinearly distorted by the channel. As noted in Schmidt–Nielsen and Crystal [24], humans are robust to such changes in environmental conditions. The goal of GMIA is to extract a signature that mutually represents the speaker in recordings from different nonlinear channels. Therefore, this feature represents the speaker but is invariant to the channels. Intuitively, this signature should provide a robust feature for speaker verification in unknown channel conditions.

As shown in Section 4, there exists an application dependent trade-off in the selection of λ to enable an accurate extraction of a common component in the data. For example, we demonstrate in Sections 5 and 6.1 that GMIA(0) provides a discriminative, illumination-invariant representation of a set of face images. Similarly, the goal of this section is to find a suitable λ for text-independent speaker verification and analyze its sensitivity to preprocessing and data segmentation. The sample mean, which is equivalent to GMIA(λ)$|_{\lambda\to\infty}$, is used as a baseline to evaluate the choice of λ. We start the analysis by a discussion of the test setup and system to enable reproducibility of the results.

We use various portions of the NTIMIT database [10] to test the effect of λ and compare our results to other methods. The NTIMIT database contains speech from 630 speakers that is nonlinearly distorted by real telephone channels. Each speaker is represented by 10 utterances that are subdivided into three content types: Type one represents two dialect sentences that are the same for all speakers in the database, type two contains five sentences per speaker that are in common with seven other speakers and type three includes three unique sentences. We use a mix of all content types for training and testing.

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 ([7], p. 50) to ensure that only one of those excitation types is present in each instance. A comparison of the waveform structures from voiced and unvoiced sounds is shown in Fig. 9. In this section, we analyze the speaker verification performance on both the original data and voiced speech. 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, $\mathbf{m}$ represents 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, for each instance i

$$\mathbf{x}_i^{(p)}=\log\mathbf{v}_i^{(p)}=\log\mathbf{e}_i^{(p)}+\log\mathbf{h}^{(p)}+\log\mathbf{m}_i \tag{21}$$

This additive model suggests that we can use GMIA to extract a signature that represents the speaker's vocal tract $\log\mathbf{h}^{(p)}$. Several preprocessing steps are necessary to transform the raw data such that the additive model holds.

Table 1
Comparison of the identification error rate (IER) of GMIA with other methods using the Yale database.

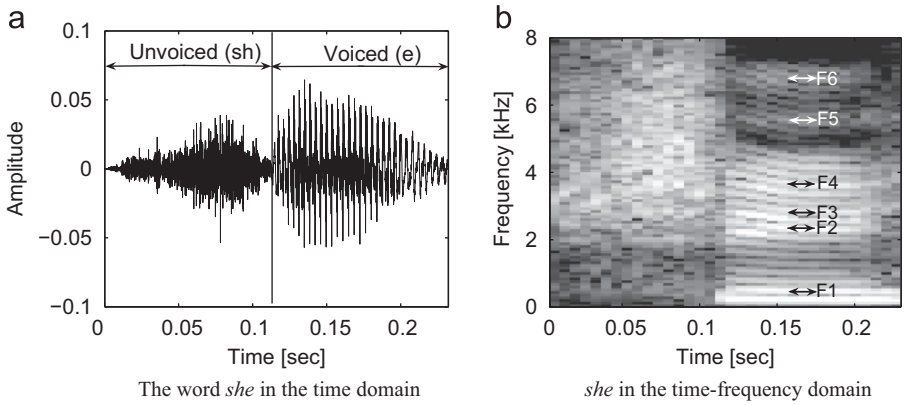
Method	IER (%)	Evaluation	Comments
GMIA (in this paper)	7.3	Leave-one-out	Cropped face test
GMIA(0) [5]	7.4		
Minimax probability machine [15]	21.2	k-fold cross validation	
Kernel PCA [28]	26.0	Leave-one-out	
Fisherface [3][a]	7.3		
Eigenface [3][b]	24.4		
Eigenface w/o 1–3 [3][b,c]	15.3		
GMIA(0) [5]	2.2	Leave-one-out	Only illumination
Minimax Probability Machine [15]	10.1	k-fold cross validation	Without illumination
Fisherface [3][a]	0.6	Leave-one-out	Full face test
Eigenface [3][b]	19.4		
Eigenface w/o 1–3 [3][b,c]	10.8		

Full faces include some background compared to cropped images.

[a] Classification was performed using 15 FLDA directions.
[b] Classification was performed using 30 principal components.
[c] The first three principal components that represent eigenvectors with maximal eigenvalues were disregarded.

6.2.1. Data preprocessing

In contrast to Claussen et al. [4], each of the utterances is preprocessed separately to prevent cross interference. First, silence and background noise are excluded from the wave data. To achieve this, the logarithmic absolute kurtosis values for 20 ms, half overlapping data intervals are compared against an empirical threshold. If the values of more than two consecutive intervals fall below this threshold, all but the first and last interval are cut. The two retained intervals are exponentially smoothed



The word *she* in the time domain

she in the time-frequency domain

Fig. 9. Structure of voiced versus unvoiced sounds. (a) The unvoiced part / ʃ / of the word *she* appears like amplitude modulated noise. The voiced part /i/ has a clear periodic structure. (b) The time frequency representation of the same waveform unveils the formants (F1–F6) of the voiced /i/. In contrast, the unvoiced sounds are smoothly structured over the whole frequency range lacking the horizontal line-structure of the voiced sounds. Note that there is not always such a clear boundary between the voiced and unvoiced sounds as in this example.

preventing discontinuities at the cutting ends. Second, the unvoiced speech segments are eliminated using a short-time autocorrelation (STAC) like approach. Let $w(k)$ represent a window function with nonzero elements for $k=0\ldots K-1$. The STAC, which is commonly used for voiced/unvoiced speech separation, is defined as ([7], p. 35)

$$STAC_n(i)=\sum_{m=-\infty}^{\infty} x(m)\,w(n-m)\,x(m-i)\,w(n-m+i)$$

The range of the summation is limited by the window $w(k)$. Furthermore, $STAC$ is even, $STAC_n(i)=STAC_n(-i)$, and tends toward zero for $|i|\to K$. The disadvantage of this method is its inherent filter effect that makes it necessary to use long windows ([7], p. 46). However, short windows are important to ensure accurate voiced/unvoiced segmentation. Thus, we employ a different windowing procedure that reduces this effect and prevents the convergence toward zero. In the following, we use the Hann window:

$$w(k)=\begin{cases} 0.5\left(1-\cos\left(\dfrac{2\pi k}{K-1}\right)\right) & \text{for } 0\le k\le K-1 \\ 0 & \text{otherwise}\end{cases}$$

The modified short-time autocorrelation (MSTAC) function is given by

$$MSTAC_n(i)=\sum_{m=-\infty}^{\infty} x(m)\,w(m-n)\,x(m+i)\,w(m-n)$$

We compute this result for $i=-K/2\ldots K/2$ and steps in n of size $K/2$. Note that in contrast to the STAC, these results are not necessarily even. However, quasi-periodic signals $x(m)$, e.g., voiced sounds, unveil their periodicity in this domain. The voiced and unvoiced segments are separated using an empirical decision function that compares the low and high frequency energies of each segment. That is, the input segment is assumed to be voiced if the low frequency energies (0–680 Hz) outweigh the high frequencies (680–3400 Hz) and vice versa.

The NTIMIT utterances are band limited by the telephone channels used. Thus, to increase the signal-to-noise ratio, the voiced speech is downsampled to 6.8 kHz. The data are processed with various window sizes to show data segmentation effects. Each utterance is segmented separately to comply with the data model in Eq. (21). An overlap is introduced if more than half of a segment would be disregarded at the end of an utterance. This step limits the loss of signal energy for short utterances and long window sizes. We partition the utterances alternating in a training and testing set to balance the text type composition.

6.2.2. Feature extraction

The segmented voiced speech $\mathbf{x}^{(p)}$ is nonlinearly transformed to fit the linear model in Eq. (18). Throughout this article, we have used correlation coefficients as a measure of similarity between two vectors. This measure is sensitive to outliers. Also, low signal values result in large negative peaks in the logarithmic domain. A nonlinear filter and offset are used, before the logarithmic transformation, to reduce the effect of these signal distortions. First, the inputs are transferred to the absolute of their Fourier representation. Second, each sample is reassigned with the maximum of its original and its direct neighboring sample values. Third, an offset is added to limit the sensitivity to low signal intensities that are affected by noise. The resulting signals are transferred to the logarithmic domain.

Speech has a speaker-independent characteristic with maximum energy in the lower frequencies. As we aim to extract signatures to distinguish speakers, it is sensible to disregard information that is common between them. Also, by disregarding this information, we prevent the effect illustrated in Fig. 2(c). To achieve this, the mean of the original inputs of all speakers is decorrelated from them. The new inputs are then used to compute the final GMIA signatures for each speaker. The procedure used to extract GMIA speaker signatures is illustrated in Fig. 10.

Note that the GMIA result is a weighted sum of the high-dimensional inputs. For example, a window size of 250 ms and 10 s of speech data result in $D=1700$ and $N=40$. In the nonlinear logarithmic space, it is not meaningful to subtract two features from each other. Therefore, the parameter λ is chosen as the smallest value that ensures positive weights. Note that in the limit

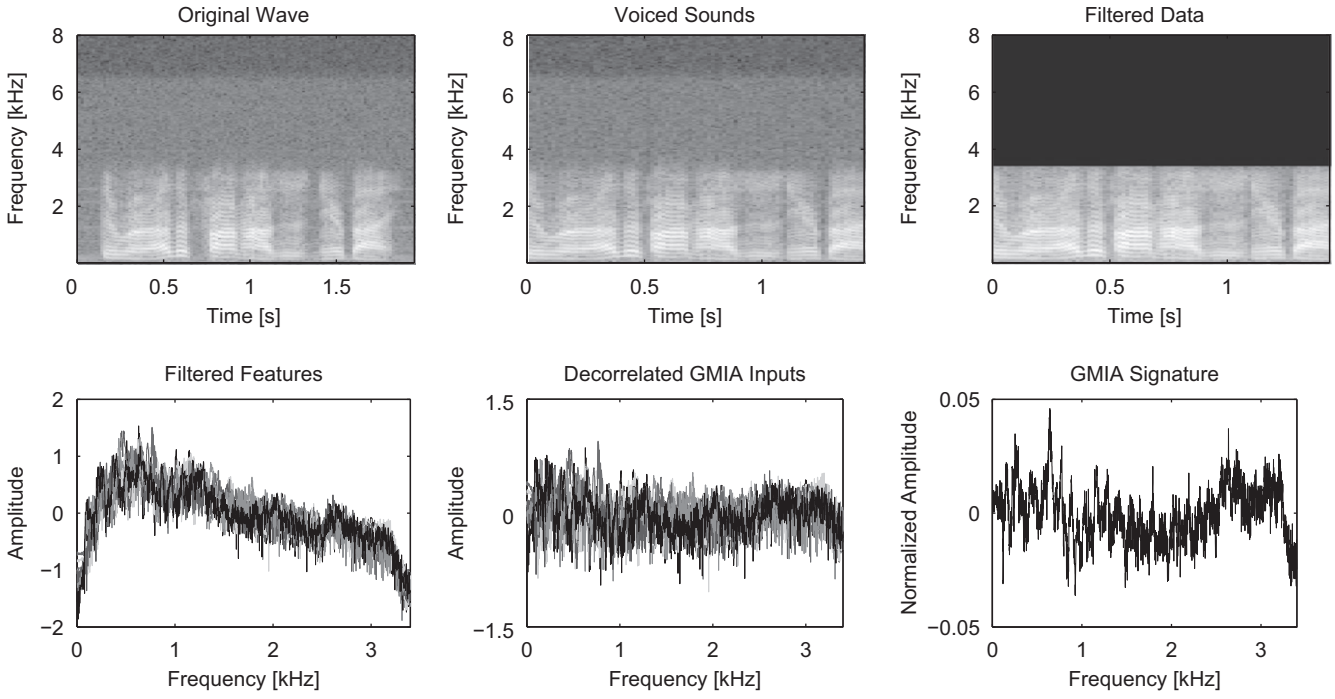


Fig. 10. Processing chain for text-independent speaker verification using GMIA.

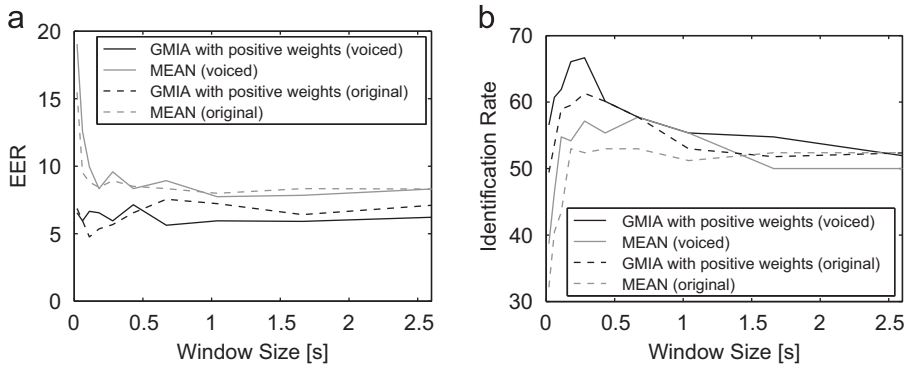


Fig. 11. Comparison of speaker verification results using GMIA and mean features. Optimal performance is achieved for window lengths between 100 and 500 ms. Note that the performance drops sharply for shorter window sizes. (a) GMIA clearly outperforms the mean based feature. (b) The result of the mean feature is more affected than GMIA if only voiced speech is used.

($\lambda \to \infty$), all weights are equal and positive. The similarity value of the test data and the learned signatures is given as the negative sum of square distances between the correspondent signatures. The possible range of the GMIA distance is $[-4,0]$ because $\|\mathbf{w}_{\mathrm{GMIA}}\| = 1$.

6.2.3. Speaker verification performance evaluation

Let P, CA, WA, IR, FAR, FRR and EER denote the number of speakers in the database, number of correctly accepted speakers, number of wrongly accepted speakers, identification rate, false acceptance rate, false rejection rate and equal error rate, respectively. The IR, FAR and FRR rates are given by

$$\mathrm{IR} = 100\frac{\mathrm{CA}}{\mathrm{P}}(\%), \quad \mathrm{FRR} = 100\left(\frac{\mathrm{P}-\mathrm{CA}}{\mathrm{P}}\right)(\%), \quad \mathrm{FAR} = 100\left(\frac{\mathrm{WA}}{\mathrm{P}(\mathrm{P}-1)}\right)(\%)$$

In the speaker identification problem, the identity of the speaker with the highest score is assigned to the current input. On the other hand, in speaker verification, a speaker is accepted if the score between its own and the claimed identity signature exceeds the one with a background speaker model by more than a defined threshold. In the following, this background model is taken simply as the signature of a speaker in the database that achieves the highest score with the claimant's input. Thus, multiple speakers from the database could be accepted for a single claimed identity. The error rates are computed using all possible combinations of claimant and speaker identities in the database. For simplicity, we do not simulate an open set where unknown impostors are present. Clearly, the threshold has a direct effect on the FRR and FAR. The point where both error ratios are equal, called equal error rate (EER), is a prominent evaluation criterion for verification methods.

6.2.4. Experimental results

Fig. 11(a) illustrates the EER results of the speaker verification approach discussed above on the NTIMIT test portion of 168 speakers. We experimented with various window sizes. As shown in Fig. 11(b), the performance is optimal for windows between 100 and 500 ms and drops sharply for shorter lengths. The results of unprocessed speech are compared to the ones using only voiced speech. In all cases, GMIA is contrasted to the mean input feature.

Table 2 presents EER results of GMIA against previous approaches of the authors and other representative results from the literature. The identification rates of the algorithms are included for comparison with previous results in the literature. Our assumption of differently distorted inputs results in the chosen data partitioning where the utterances are alternatively separated in a training and testing set. Note that this partitioning—and therefore the results—are not exactly comparable to the standard work of, e.g., Reynolds [22].

Table 2
GMIA(0) and GMIA performance comparison using various NTIMIT database segments. "GMM" indicates the standard Gaussian mixture model approach [22].

Method	EER (%)	Identification (%)	NTIMIT database section
GMIA (in this paper)	6.0	67	Test section with 168 speakers
GMIA [6]	6.0	52	
GMIA(0) [6]	6.9	48	
GMIA(0) [5]	6.8	56	
GMM [27]	12.4	N/A	
GMM [23]	9.6	N/A	
GMM [22]	7.2	69	
GMIA (in this paper)	5.7	47	Selection of all 438 male speakers
GMIA [6]	6.9	39	
GMIA(0) [6]	8.4	35	
Phoneme GMM [13]	15.7	N/A	
GMIA (in this paper)	5.1	44	Full database of 630 speakers
GMIA [6]	6.5	37	
GMIA(0) [6]	7.5	32	
GMM [27]	8.8	N/A	

7. Conclusion

The Bayesian estimation perspective on the mutual interdependence analysis problem allows for a parameterized formulation called GMIA(λ). When the parameter $\lambda = 0$, GMIA(0) is equivalent to the original definition of MIA. The goal of GMIA(0) is to compute a unique characteristic or invariant feature of a high-dimensional dataset that can be used in pattern recognition problems. By definition, the GMIA(0) representation is a linear combination of class examples that has an equal correlation with all training samples in the class.

This paper defines a generative signal model for GMIA(λ) and analyses the effect of λ on its feature extraction performance. This allows us to evaluate and successfully apply GMIA(λ) in two problems: illumination-independent face recognition and text-independent speaker verification. GMIA-based methods are rather general, nonetheless they extract discriminant features resulting in competitive classification performance. Given that

the GMIA solution depends on the Gram matrix of the data, future work will investigate computational tractability in large dimensions and statistical properties of GMIA for a large number of inputs.

References

[1] F.R. Bach, M.I. Jordan, A probabilistic interpretation of canonical correlation analysis, Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[2] M.S. Bartlett, Further aspects of the theory of multiple regression, Proceedings of the Cambridge Philosophical Society 34 (1938) 33–40.

[3] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[4] H. Claussen, J. Rosca, R. Damper, Mutual interdependence analysis, in: Independent Component Analysis and Blind Signal Separation, Springer-Verlag, Heidelberg, Germany2007, pp. 446–453.

[5] H. Claussen, J. Rosca, R. Damper, Mutual features for robust identification and verification, in: International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008, pp. 1849–1852.

[6] H. Claussen, J. Rosca, R. Damper, Generalized mutual interdependence analysis, in: International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 3317–3320.

[7] L. Deng, D. O'Shaughnessy, Speech processing: a dynamic and optimization-oriented approach, Signal Processing and Communications, Marcel Dekker, Inc, 2003.

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

[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.

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

[11] J.D. Foley, A. van Dam, S.K. Feiner, J.F. Hughes, Computer Graphics: Principles and Practice, second ed., Addison-Wesley Longman Publishing, Boston, MA, 1997.

[12] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 643–660.

[13] D. Gutman, Y. Bistritz, Speaker verification using phoneme-adapted Gaussian mixture models, in: European Signal Processing Conference, vol. 3, Toulouse, France, 2002, pp. 85–88.

[14] T. Hastie, R. Tibshirani, A. Buja, Flexible discriminant analysis by optimal scoring, Journal of the American Statistical Association 89 (428) (1994) 1255–1270.

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

[16] H. Hotelling, Relation between two sets of variates, Biometrika 28 (1936) 322–377.

[17] C. Jutten, J. Herault, Blind separation of sources, part 1: an adaptive algorithm based on neuromimetic architecture, Signal Processing 24 (1) (1991) 1–10.

[18] S.M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.

[19] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, Padstow, Cornwall, UK, 1979.

[20] T.K. Moon, W.C. Stirling, Mathematical Methods and Algorithms for Signal Processing, Prentice-Hall, Upper Saddle River, NJ, 2000.

[21] K. Pearson, On lines and planes of closest fit to points in space, Philosophical Magazine 2 (1901) 559–572.

[22] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Communication 17 (1–2) (1995) 91–108.

[23] C. Sanderson, Speech processing and text-independent automatic person verification. Technical Report 08, IDIAP, Martigny, Switzerland, 2002.

[24] A. Schmidt-Nielsen, T.H. Crystal, Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data, Digital Signal Processing 10 (2000) 249–266.

[25] A. Tikhonov, On the stability of inverse problems, Doklady Akademii Nauk SSSR 39 (5) (1943) 195–198.

[26] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[27] B. Wildermoth, K. Paliwal, GMM based speaker recognition on readily available databases, in: Microelectronic Engineering Research Conference, Brisbane, Australia, pagination unknown, 2003.

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

[29] S. Zhou, G. Aggarwal, R. Chellappa, D. Jacobs, Appearance characterization of linear Lambertian objects, generalized photometric stereo, and illumination-invariant face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2) (2007) 230–245.

Heiko Claussen is a Research Scientist at Siemens Corporation, Corporate Research in Princeton, USA. He obtained his Dipl.- Ing. (FH), M.Eng. and Ph.D. degrees in Electrical Engineering from the University of Applied Sciences in Kempten, Germany, the University of Ulster, UK, and the University of Southampton, UK, respectively. His research interests include feature extraction, pattern recognition and machine learning in domains like audio and image processing. He received the best graduate of the year award in Electrical Engineering from the University of Applied Sciences in Kempten in 2005, and graduated with distinction at the University of Ulster, where he received the first prize for his final project "Canceling Cardiovascular Noise from Speech" from the IEE.

Justinian Rosca is a Program Manager in Audio, Signal Processing and Wireless Communications at Siemens Corporation, Corporate Research in Princeton, USA. He is also Affiliate Professor, Department of Electrical Engineering of University of Washington, Seattle, USA. He received the Dipl. Eng. degree in Computers and Control Engineering from Bucharest Polytechnic University in 1984, the M.S. and Ph.D. degrees in Computer Science from University of Rochester in 1992 and 1997, respectively. Dr. Rosca is conducting research in signal processing and radio management, with an emphasis on topics involving acquisition, management and processing of data with uncertainties, such as statistical audio or wireless processing, signal separation, wireless management, adaptive principles in stochastic search and optimization, and probabilistic inference in artificial intelligence. Dr. Rosca has more that two dozen US and international patents awarded, and more than 80 reviewed publications. He coauthored a book on solved problems in higher mathematics and coedited the Proceedings of ICA 2006. He is presently on the editorial board of the Journal of Signal Processing Systems and the Journal for Genetic Programming and Evolvable Hardware both from Springer and serves as a member of committees of various conferences in the areas of machine learning and signal processing. Dr. Rosca chaired numerous sessions at international conferences, and events such the International Conference on Independent Component Analysis as program chair in 2006, and the Sparse Representations in Signal Processing workshop at Neural Information Processing Systems Conference in 2003. He gave tutorials or invited talks on stochastic search techniques, signal processing and radio management at multiple international events within the last 10 years. Dr Rosca is a member of AAAI and IEEE.

Robert Damper was born in Tunbridge Wells, England, in 1948. He obtained his M.Sc. in Biophysics in 1973 and Ph.D. in Electrical Engineering in 1979, both from the University of London. He also holds the Diploma of Imperial College, London, in Electrical Engineering. He was appointed as a Lecturer in Electrical Engineering at the University of Abertay Dundee in 1976, Lecturer in Electronics at the University of Southampton in 1980, Senior Lecturer in Electronics and Computer Science in 1989, Reader in 1998 and Professor in 2003. He has wide research interests including speech science and technology, neural computing, cognitive modeling, pattern recognition and intelligent systems engineering. Prof. Damper has published over 300 research articles and authored the undergraduate text "Introduction to Discrete-Time Signals and Systems". He is a chartered engineer and a fellow of the UK Institution of Engineering and Technology, a chartered physicist and a fellow of the UK Institute of Physics, a senior member of the IEEE and an honorary (foreign) member of the Yugoslav Engineering Academy.