A comprehensive overview of feature representation for biometric recognition

Article *in* Multimedia Tools and Applications · February 2020

DOI: 10.1007/s11042-018-6808-5

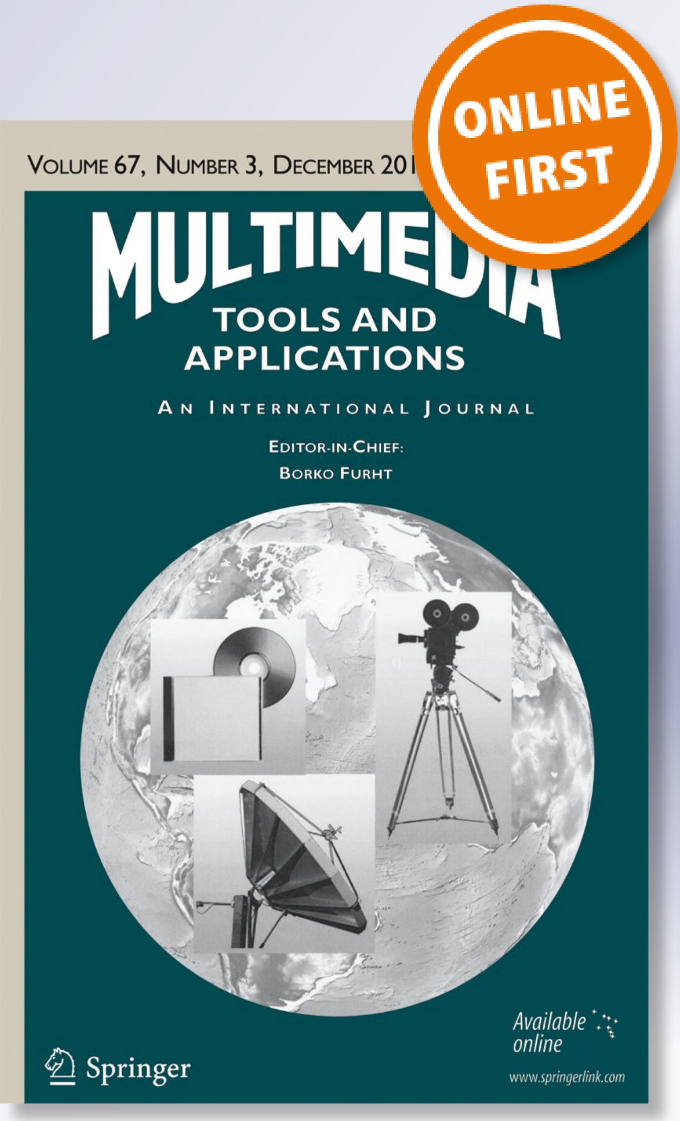
A comprehensive overview of feature representation for biometric recognition

Imad Rida, Noor Al-Maadeed, Somaya Al-Maadeed & Sambit Bakshi

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-018-6808-5



Springer

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-018-6808-5

A comprehensive overview of feature representation for biometric recognition

Imad Rida[1] · Noor Al-Maadeed[1] · Somaya Al-Maadeed[1] · Sambit Bakshi[2]

Received: 12 June 2018 / Revised: 18 September 2018 / Accepted: 23 October 2018 /
Published online: 29 October 2018

Abstract

The performance of any biometric recognition system heavily dependents on finding a good and suitable feature representation space where observations from different classes are well separated. Unfortunately, finding this proper representation is a challenging problem which has taken a huge interest in machine learning and computer vision communities. In the this paper we present a comprehensive overview of the different existing feature representation techniques. This is carried out by introducing simple and clear taxonomies as well as effective explanation of the prominent techniques. This is intended to guide the neophyte and provide researchers with state-of-the-art approaches in order to help advance the research topic in biometrics.

Keywords Biometrics · Feature representation · Dimensionality reduction · Feature selection · Decomposition learning

1 Introduction

Over the few past decades, biometric security is increasingly becoming an important tool to enhance security and brings greater convenience. Nowadays, biometric systems are widely

✉ Imad Rida
rida.imad@gmail.com

Noor Al-Maadeed
n.alali@qu.edu.qa

Somaya Al-Maadeed
s_alali@qu.edu.qa

Sambit Bakshi
sambitbaksi@gmail.com

[1] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

[2] Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela 769008, India

used by government agencies and private industries. Though a growing effort has been devoted in order to develop robust biometric recognition systems that can operate in various conditions, many problems still remain to be solved, including the design of techniques to handle varying illumination sources, occlusions and low quality images resulting from uncontrolled acquisition conditions [1, 70, 73, 78, 81]. Indeed, doing this efficiently and completely represents a continuous challenging problem which took the effort and attention of researchers.

Biometric recognition consists in determining the identity for a given person based on his/her physiological or behavioral characteristics [5, 76]. A large variety of biometric modalities including face, iris, gait, keystroke and palmprint have been studied providing different rates of robustness, accuracy and user acceptability [29, 79, 80]. Given the goal of recognizing individuals based on their features, the main task of an automated recognition system can be divided into three basic subtasks: the description subtask which generates features of a person using feature extraction techniques, mapping raw features into another discriminative space where different persons are well separated by feature representation techniques and finally the classification subtask which assigns the identity to new person based on those features and a trained classifier (see Fig. 1).

With the recent advances in machine learning techniques; the goal of this paper is to familiarize the biometrics researchers to the prominent methods and ways of machine learning for feature representation and its correct research methodology. Section 2 introduces the assumptions of a good representation. Section 3 presents the main dimensionality reduction techniques. Section 4 describes the prominent feature selection approaches. Section 5 reports decomposition learning. Section 6 explains the classification. Section 7 reports the discussions. Finally, Section 8 gives our conclusion.

2 Feature representation

The performance of any recognition system is heavily dependent on finding a good and suitable feature representation space which should satisfy the following assumptions [9]:

- Smoothness: in a high density region, if two points $\mathbf{x}_1$ and $\mathbf{x}_2$ are near $\mathbf{x}_1 \approx \mathbf{x}_2$, their outputs by a decision function f are more probable to be close $f(\mathbf{x}_1) \approx f(\mathbf{x}_2)$. This assumption implies also that in case two points are connected by a high density path, their outputs are also likely to be close also. On the other hand, if they are connected by a low density path, then their outputs don't need to be close.
- Cluster: the data tend to be organized in discrete clusters, and points in the same cluster are more likely to share the same class label. The cluster assumption does not mean the data from each class forms a single and unique compact cluster, but rather that we may not observe data from two different classes within the same cluster.
- Manifolds: curse of dimensionality represents a huge problem for many discriminative learning algorithms since the distances tend to be less meaningful and representative. The manifold assumption implies that, the initial data of high dimension reside in a manifold of lower dimension integrated in the ambient space to overcome the curse of dimensionality problem.
- Sparsity: a feature vector $\mathbf{x}$ is called sparse if most of its entries are zeros. Sparse representations are able to extract the hidden structure and provide a simple interpretation of the input data. Furthermore, it has been found that biological vision is based on sparse representations [68].

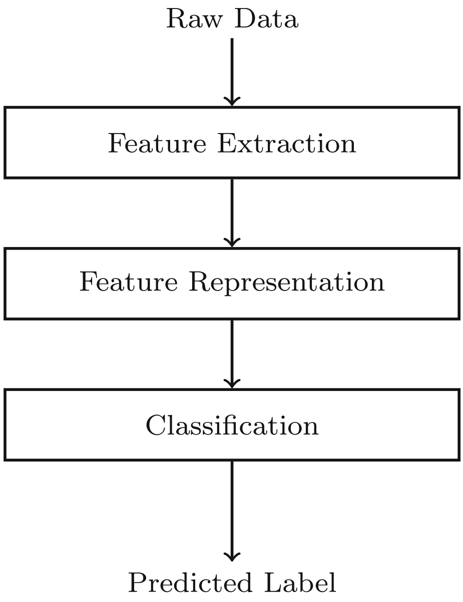


Fig. 1 Scheme of a conventional recognition system

- Temporal and spatial coherence: spatially nearby or consecutive (temporally close) observations tend to share the same value ($\mathbf{x}_t \approx \mathbf{x}_{t+1}$). The simultaneous temporal and spatial changes should be penalized.

We envision feature representations under three points of view: dimensionality reduction, feature selection and decomposition learning (see Fig. 2).

3 Dimensionality reduction

Analyzing high-dimensional data is a difficult problem, since the high-dimensional spaces have geometrical properties that are very complex and hardly interpretable compared to low-dimensional ones. Furthermore, learning a good model needs enough data, while the number of learning data should grow exponentially with the dimension which causes the so called curse of dimensionality [98, 100]. Dimensionality reduction aims to find a transformation mapping the original data residing in a high-dimensional space into a lower one able to capture and preserve the intrinsic characteristics of the initial data. It helps in classification, visualization and compression since it has ability if well designed, to reduce the undesirable

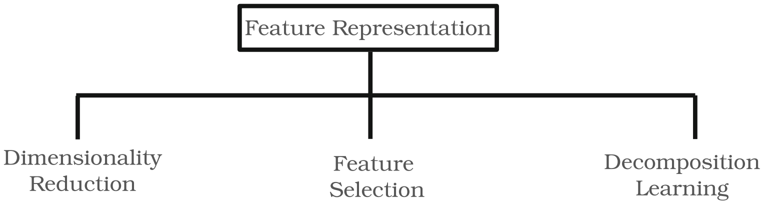


Fig. 2 Taxanomy of feature representation approaches

effects of high-dimensional spaces [39]. Related techniques can be broadly divided into two main groups, linear and non-linear. We briefly introduce hereafter the prominent linear and non-linear methods.

3.1 Linear dimensionality reduction

Given n d-dimensional samples $\{\mathbf{x}_i\}_{i=1}^{n}$ stored in matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and a dimensionality choice $r < d$, linear dimensionality reduction aims to find a linear matrix transformation $\mathbf{P} \in \mathbb{R}^{r\times d}$ by optimizing an objective function J such that the high-dimensional $\mathbf{X}$ is mapped into low-dimensional data $\mathbf{Z} = \mathbf{P}\mathbf{X} \in \mathbb{R}^{r\times n}$. Linear dimensionality reduction methods can be formulated as an optimization problem over a manifold matrix [19] as follows:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & J(\mathbf{M},\mathbf{X}) \\ \text{s.t} & \mathbf{M}\in\mathcal{M} \end{cases} \tag{1}$$

The objective function J and the manifold matrix $\mathbf{M}$ try to capture the desired and relevant characteristics. In some linear dimensionally techniques, the matrix $\mathbf{M}$ is imposed to be orthogonal, hence $\mathcal{M} = \{\mathbf{M} \in \mathbb{R}^{d\times r} : \mathbf{M}^T\mathbf{M} = \mathbf{I}\}$. In this particular case the manifold $\mathcal{M}$ is noted $\mathcal{O}^{d\times r}$.

The relation between the projection matrix $\mathbf{P}$ and $\mathbf{M}$ will change depending on the used method. Indeed, there are many techniques in linear dimensionality reduction such as, Principal Component Analysis (PCA) [65], Linear Discriminant Analysis (LDA) [30], Independent Component Analysis (ICA) [37] and Factor Analysis (FA) [88]. The objective function J differs according to desired properties or assumptions (supervised or not, gaussian assumption, statistical independence, etc) encoded by these techniques.

3.1.1 Principal component analysis

Principal Component Analysis (PCA) is an unsupervised linear dimensionality reduction technique initially formulated as the minimization of the residual errors between the original and the projected data [16, 54, 55, 65]:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & \left\|\mathbf{X}-\mathbf{M}\mathbf{M}^T\mathbf{X}\right\|_F^2 \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{2}$$

The problem 2 can be equivalently reformulated as variance maximization of projected data [10] leading to:

$$\begin{cases} \min\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & -\text{tr}\,(\mathbf{M}^T\mathbf{X}\mathbf{X}^T\mathbf{M}) \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{3}$$

The solution $\mathbf{M}$ corresponds to the r leading principal eigenvectors of $\mathbf{X}\mathbf{X}^T$ and we get the projection matrix $\mathbf{P} = \mathbf{M}^T$. The size of covariance matrix $\mathbf{X}\mathbf{X}^T$ is proportional to the dimensionality of the data which could lead to the tedious calculation of the eigenvectors when the initial data has very high-dimensionality. There have been some extensions of the PCA, such as Kernel PCA [85] a non-linear extension, probabilistic PCA [82, 94] and sparse PCA [20, 42, 113].

3.1.2 Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a supervised technique which aims to project the data in lower subspace where the data from different classes are well separated. In other terms, the LDA seeks to minimize the intra-class variations and to maximize the between-class variations. It is formulated by the following problem [28]:

$$\begin{cases} \max\limits_{\mathbf{M}\in\mathbb{R}^{d\times r}} & \dfrac{\text{tr }(\mathbf{M}^T\Sigma_B\mathbf{M})}{\text{tr }(\mathbf{M}^T\Sigma_W\mathbf{M})} \\ \\ \text{s.t} & \mathbf{M}\in\mathcal{O}^{d\times r} \end{cases} \tag{4}$$

with

$$\Sigma_W = \sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_{c_i})(\mathbf{x}_i-\boldsymbol{\mu}_{c_i})^T \qquad \Sigma_B = \sum_{i=1}^{n}(\boldsymbol{\mu}_{c_i}-\boldsymbol{\mu})(\boldsymbol{\mu}_{c_i}-\boldsymbol{\mu})^T \tag{5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{c_i}$ respectively represent the mean of the whole dataset and the mean of class c which the sample $\mathbf{x}_i$ belonging to. The projection matrix $\mathbf{P} = \mathbf{M}^T$.

3.1.3 Independent component analysis

Independent Component Analysis (ICA) is a linear higher-order method which does not impose the orthogonality constraint and with assumption that the components are as independent as possible. Compared to uncorrelatedness of linear PCA, the statistical independence represents a stronger condition to represent the data. ICA tries to find a matrix $\mathbf{P}\in\mathbb{R}^{r\times d}$ which is able to capture the independent sources $\mathbf{Z}\in\mathbb{R}^{r\times n}$ from the initial data $\mathbf{X}\in\mathbb{R}^{d\times n}$ where $\mathbf{Z} = \mathbf{PX}$. The majority of ICA implementations deal with dimension preserving case where the projection $\mathbf{P}$ is such that $d = r$ (in this case, the ICA is not seen as a dimensionality reduction method since it preserves the dimensionality of the initial data).

To use the ICA as dimensionality reduction method, an undercomplete version $r < d$ is needed. There are several works which tried to undercomplete the ICA using a preprocessing step [2, 21, 67, 103, 110]. A possible preprocessing is PCA, which reduces the dimensionality of the initial data to $r < d$, after that the conventional ICA is applied to the resulting data [41] which leads to a projection in a low-dimensionality space with statistical independence. Note also that there are also overcomplete versions of the ICA when $r > d$ [91] mainly applied to blind source separation task.

3.1.4 Factor analysis

Factor analysis (FA) is a generative model which assumes that the observed data have been produced from a set of latent unobserved variables (called here factors). FA can be seen as a more general case of Probabilistic PCA (PPCA) [19, 43] and addresses the following problem:

$$\min_{\mathbf{M}\in\mathbb{R}^{d\times r}} \log|\mathbf{M}\mathbf{M}^T+\mathbf{D}| + \mathrm{tr}\left((\mathbf{M}\mathbf{M}^T+\mathbf{D})^{-1}\mathbf{X}\mathbf{X}^T\right) \tag{6}$$

where $\mathbf{M}$ is the factor loading matrix and $\mathbf{D}$ is a diagonal matrix for the conditional data likelihood $\mathbf{x}_i|\mathbf{z}_i \sim \mathcal{N}(\mathbf{M}\mathbf{z}_i, \mathbf{D})$ representing the observation noise fit. The linear dimensionality reduction mapping of the initial data $\mathbf{X}$ is given by $\mathbf{Z} = \mathbf{P}\mathbf{X}$ where $\mathbf{P} = \mathbf{M}^T(\mathbf{M}\mathbf{M}^T + \mathbf{D})^{-1}$.

3.2 Nonlinear dimensionality reduction

Conventional linear dimensionality reduction techniques, such as PCA and ICA are designed to operate when the observed initial high-dimensionality data is embedded in a low-dimensional linear manifold. However, real world data have a very complex structure and reside generally on nonlinear manifolds [56]. Based on the latter reasons it has been demonstrated that traditional methods are not suitable to deal with such complex structure. Encouraged by the gaps and weakness of linear techniques, numerous nonlinear dimensionality reduction techniques have been introduced. These techniques can be broadly divided into two main groups: local and global. The local approach involves Locally Linear Embedding (LLE) [83] and Laplacian Eigenmaps (LE) [7]; when the global approach involves Isometric Feature Mapping (Isomap) [90] to name a few.

Local methods seek to preserve the local geometry of the observed data; in other terms, these methods try to preserve the neighborhood by mapping the nearby points in the initial high-dimensional manifold to nearby points in low-dimensional one. This is done by approximating each point on the manifold with a combination of its neighbors; and then based on resulting weights, a low-dimensional embedded manifold is constructed. Local approaches have good representational ability, for a larger range of manifolds, whose local geometry is close to Euclidean, furthermore they are computationally efficient [87]. In the other hand, global methods, attempt to preserve the geometry at all scales, mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points. The advantage of the global methods is the ability to give more general and faithful representation of global structure of the data [87].

There have been some works which tried to incorporate strengths of the local methods in the global methods such as Conformal Isomap (C-Isomap) [87]. C-Isomap extends Isomap to be capable to learn the structure of curved manifolds. As a result it is computationally efficient (equals to or better than the existing local approaches such LLE and LE) with good stability and theoretical tractability characteristics of the methods belonging to global approach [87]. In the following we introduce the main concepts of several widely used nonlinear techniques.

3.2.1 Isomap

It attempts to preserve the geometric properties of the data. It was introduced to deal with the problem of classical scaling methods which consider two high-dimensional data points lying in curved manifold as close points whereas they are not really close [96]. Isomap method has three main steps, the first one consists on constructing a neighborhood graph G where each data point $\{\mathbf{x}_i\}_{i=1}^n$ is connected with its neighbors $\{\mathbf{x}_j\}_{j=1}^k$ in the high-dimensional dataset $\mathbf{X} \in \mathbb{R}^{d\times n}$. In second step, Isomap estimates the geodesic distances between all pairs of data points by computing their shortest path in the graph G using Dijikstra's [23] or Floyd's [31] shortest path algorithm. The third and ultimate step consists on applying classical Multidimensional Scaling (MDS) [95] to resulting geodesic distance matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$. It consists in solving the following optimization problem:

$$\min_{\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left({d_{ij}}^2 - \|\mathbf{z}_i - \mathbf{z}_j\|^2 \right) \tag{7}$$

where d_{ij} represents the geodesic distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{z}_i$ and $\mathbf{z}_j$ stand for the low-dimensional representation of $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. It has been shown that the solution of the problem is $\mathbf{Z} = \mathbf{U}\Sigma^{\frac{1}{2}}$ issued from the spectral decomposition of the Gram matrix $\mathbf{K}$ which is the double centering of the geodesic distance matrix $\mathbf{D}$.

3.2.2 Locally linear embedding

Locally Linear Embedding (LLE) is a method which aims to preserve the local characteristics and properties of the data. Compared to the methods belonging to global approach such as Isomap, the LLE is less sensitive to short-circuiting problem which happens when the local neighborhood connections shortcut across the manifold [96]. LLE captures the local properties of the manifold around each data point $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ by expressing $\mathbf{x}_i$ as a linear combination of its k neighbors $\{\mathbf{x}_{ij}\}_{j=1}^k$ with coefficients $\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^n$. Here $\mathbf{x}_{ij}$ represents the j^{th} neighbor of $\mathbf{x}_i$. By doing so, the manifold is assumed to be locally linear which implies that the weights $\mathbf{w}_i$ of $\mathbf{x}_i$ are invariant to different transformations such as translation and rotation, etc. Formally the weights $\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^n$ are first estimated by solving:

$$\begin{cases} \min\limits_{\{\mathbf{w}_i \in \mathbb{R}^k\}_{i=1}^n} \sum\limits_{i=1}^{n} \left\| \mathbf{x}_i - \sum\limits_{j=1}^{k} w_{ij} \mathbf{x}_{ij} \right\|^2 \\ \text{s.t} \quad \sum\limits_{j=1}^{k} w_{ij} = 1 \qquad \forall i = 1, \cdots, n \end{cases}$$

We shall notice that the weights $w_{ij} = 0$ for all samples $\mathbf{x}_j$ not belonging to the k-neighborhood of $\mathbf{x}_i$. Based on the transformation invariance property, the weights $\mathbf{w}_i = [w_{i1}, \cdots, w_{ik}]$ that construct the initial data in high-dimensional space based on its neighbors are also able to reconstruct $\mathbf{z}_i$ from its neighbors in low-dimensional space. Finding the new representation $\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n$ where $r < d$ is formulated by the following minimization problem:

$$\begin{cases} \min\limits_{\{\mathbf{z}_i \in \mathbb{R}^r\}_{i=1}^n} \sum\limits_{i=1}^{n} \left\| \mathbf{z}_i - \sum\limits_{j=1}^{k} w_{ij} \mathbf{z}_{ij} \right\|^2 \\ \text{s.t} \quad \|\mathbf{z}_i\|^2 = 1 \qquad \forall i = 1, \cdots, n \end{cases} \tag{8}$$

[83] established that the reduced dimension solutions $\{\mathbf{z}_i\}_{i=1}^n$ are obtained by calculating the eigenvectors corresponding to r smallest nonzero eigenvalues of $(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ where $\mathbf{I} \in \mathbb{R}^{n \times n}$ and $\mathbf{W} \in \mathbb{R}^{n \times n}$ a matrix with entries equal to the weight w_{ij} when i and j are connected in the neighborhood graph and 0 otherwise. Note that there have been some extensions of the LLE such as Orthogonal Neighborhood Preserving Projections [46] and Neighborhood Preserving Embeddings [36].

3.2.3 Laplacian eigenmaps

Laplacian Eigenmaps (LE) aims to find a low-dimensional representation by preserving local properties of the high-dimensional data based on pairwise distances between neighbors. For the latter, LE tries to minimize a cost function based on the sum of the distances between each data point in the low-dimensional space $\{\mathbf{z}_i\}_{i=1}^n$ and its k nearest neighbors $\{\mathbf{z}_j\}_{j=1}^k$. The distance between each data point and its first nearest neighbor contributes more in the cost function than the second and so on. This is made possible by constructing a weighting matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$, where its entries w_{ij} corresponds to the distance between data point $\mathbf{x}_i$ and its k-nearest neighbor using the Gaussian kernel function given by:

$$\begin{cases} w_{ij} = e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}} & \text{if } \mathbf{x}_j \text{ is in the } k\text{-neighborhood of } \mathbf{x}_i \\ w_{ij} = 0 & \text{otherwise} \end{cases} \quad (9)$$

where σ is the bandwith of the Gaussian. The computation of the low-dimensional representation $\mathbf{z}_i$ is obtained through the following optimization problem:

$$\min_{\{\mathbf{z}_i\in\mathbb{R}^r\}_{i=1}^n} \sum_{i=1}^n \sum_{j=1}^n \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2 w_{ij} \quad (10)$$

In the cost function, large values of w_{ij} means that the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have small distance in the high-dimensional space. In other words, nearby points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the high-dimensional space are mapped into low-dimensional space $\mathbf{z}_i$ and $\mathbf{z}_j$ with the lowest distance possible. Defining $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n]$, the problem in formula (10) can be reformulated as an eigenproblem [96] as follows:

$$\begin{cases} \min\limits_{\mathbf{Z}\in\mathbb{R}^{r\times n}} \quad 2\mathbf{Z}\mathbf{L}\mathbf{Z}^T \\ \text{s.t} \quad \mathbf{Z}\mathbf{D}\mathbf{Z}^T = \mathbf{I} \end{cases} \quad (11)$$

where the equality constraint removes an arbitrary scaling factor in low-dimensional space, $\mathbf{D}$ is a diagonal matrix with entries $\mathbf{D}_{ii} = \sum_{j=1}^n w_{ij}$ and $\mathbf{L}$ is the graph Laplacian given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The solution of the problem is the r eigenvectors corresponding to the r smallest nonzero eigenvalues of generalized eigenvalue problem:

$$\mathbf{L}\mathbf{v} = \lambda\mathbf{D}\mathbf{v} \quad (12)$$

4 Feature selection

Feature selection aims to select a relevant feature subset $\mathcal{S}$ from the original initial set $\mathcal{I}$ ($\mathcal{S} \subset \mathcal{I}$) which is efficiently able to describe the intrinsic characteristics of the input data by reducing the impact of the noise and irrelevant features [15, 17, 57, 58, 72]. In fact dependent features do not give extra information about the data belonging to a class (e.g. when two features are highly correlated, a single one is sufficient to describe the characteristics of the class). In other words, the total information of the data can be captured only from few unique features able to express the discriminative characteristics of each class leading to the reduction of the data dimension [14, 74]. As such feature selection can be seen as an instance of dimension reduction preserving the original variables. Removing irrelevant features requires an efficient feature criterion which measures the relevance of each feature so as to be able to select a feature subset from 2^d possible subsets where d is the cardinality of $\mathcal{I}$. There are three main approaches used in features selection, filter, wrapper and embedded methods [34].

4.1 Filter

Filter methods include non-learning techniques exclusively. Features are ranked according to scores that depend on their relevance according to pre-defined criterion. They are mainly applied before the classification step, to filter out the irrelevant features (for instance features with scores below a threshold are discarded). The notion of feature relevance remains an open question; several definitions have been introduced based on the context of the problem [34, 45, 48]. In our thesis and since we are in classification context, we adopt the definition that presents an irrelevant feature as the independent one of the class label. In other words, a feature is considered irrelevant if it has no information about the class label [49]. In some cases features which have no dependency or correlation with classes serve as noise and eliminating them might lead to improvement in the classification accuracy.

Several criterions have been introduced such as, Pearson correlation coefficients [34] and Mutual Information (MI) [6, 45, 50] which are able to estimate the dependency between a feature and a target (the target can be for instance the class label). The advantage of methods belonging to filter approaches is that they are computationally efficient and avoid overfitting since they do not rely on learning algorithms [34, 50]. However filter methods have also some drawbacks, such as, MI and correlation-based methods which are not able to estimate the correlation between features leading sometimes to correlated features within the same feature subset [40, 53]. Furthermore filter methods are usually not optimal since they do not account for the mechanism of the learning algorithm [3].

4.2 Wrapper

Wrapper methods used a learning algorithm as a black-box. Given the original feature set, all possible subsets obtained by search algorithms are evaluated with a classifier. The prediction performance serves as the selection criterion, and the subset that performs the best is retained. Sadly, evaluating 2^d is an NP-hard problem and can become intractable and computationally intensive when the number of features is very large [45, 64]. Based on that, some simplified algorithms such as Genetic Algorithm (GA) [32] and Particle Swarm Optimization (PSO) [44] have been introduced; they can make a good trade off between computational cost and performance. Methods belonging to wrapper approaches can be broadly divided into, Sequential Selection Algorithms and Heuristic Search Algorithms [14].

In sequential selection algorithms we can find Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). The first one starts with an empty set and adds one feature at time which gives the maximum classification accuracy. The process is repeated until the number of required features is reached. The second one follows the same steps, however instead of starting with empty set, it starts with the full set and, instead of adding a feature, it removes it. In the heuristic search algorithms we can find algorithms such as GA [32] and its variants such as CHCGA [26] and PSO [44]. The heuristic algorithms have been introduced to avoid exhaustive search and cope against the problem of the greedy methods which do not examine all possible subsets and hence do not guarantee finding an optimal subset.

4.3 Embedded

Embedded methods, as the name suggests, embed feature selection into the learning algorithm. They seek to reduce the computation complexity time needed to evaluate the different feature subsets in order to select an optimal one as in the wrapper methods [14]. Embedded methods have been successfully used in linear problems, by including convex and concave regularization terms [89]. Recently, there have been also some works to extend feature selection methods to group feature selection in both linear and nonlinear models [61]. For sake of simplicity, we suppose that our decision function is linear and applied on $\mathbf{x} \in \mathbb{R}^d$. The definition is given by:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b \tag{13}$$

with $\mathbf{w} \in \mathbf{R}^d$ and $b \in \mathbb{R}$ is the bias. Embedded methods typically attempt to solve the learning problem:

$$\min_{\mathbf{w},b} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda\Omega(\mathbf{w}) \tag{14}$$

where y_i is the label associated with $\mathbf{x}_i$, $\Omega(\mathbf{w})$ is the regularization term and $\lambda > 0$ the regularization parameter. The first term in previous equation expresses data fitting error. Regularization aims to select features and also to avoid the overtraining [59]. This generally leads to better performances of the learned decision function [66].

The regularization $\Omega(\mathbf{w})$ tends to promote peculiar characteristics such as sparsity on $\mathbf{w}$ [101]. Norms and quasi-norms ℓ_p represent one of the most used regularization terms, they are given by:

$$\Omega_p(\mathbf{w}) = \|\mathbf{w}\|_p = \left(\sum_{i=1}^{d} |\mathbf{w}_i|^p\right)^{\frac{1}{p}} \tag{15}$$

with $0 < p \leq \infty$ and $\Omega_p(\mathbf{w})$ is considered as norm for $p \geq 1$. The regularization can be broadly categorized as standard and structured.

4.3.1 Standard regularization

- ℓ_0 - ”pseudo norm”: it counts the number of non zero coefficients in the vector $\mathbf{w}$.
- Convex relaxation: it promotes sparsity on the vector $\mathbf{w}$ using convex regularizers which generally lead to easier optimization problem.
  - Norm ℓ_2: also called Euclidean norm because it is inducted from the dot product. In the case of the linear regression [35], the square of the ℓ_2 regularization is called ridge regression. Notice that sparsity is attained in practice for high values of the regularization parameter λ.
  - Norm ℓ_1: it is known in the linear regression as LASSO (Least Absolute Shrinkage and Selection Operator) [92].

  - Fused Lasso: it penalizes ℓ_1-norm of the difference between two successive coefficients of $\mathbf{w}$ which leads to sparsity of the coefficients difference [93]:

$$\Omega(\mathbf{w}) = \sum_{i=1}^{d-1} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|_1 \tag{16}$$

- Non-convex relaxation: promotes sparsity more strongly than convex regularizers, but it suffers the difficulties brought by local optimums.
  - ℓ_p with $0 < p < 1$: when the sparsity obtained by ℓ_1 is not sufficient and more sparsity is needed, the ℓ_p with $0 < p < 1$ could be applied.
  - Log-sum: introduced in [104] for sparse SVM classification, it is given by:

$$\Omega_\epsilon(\mathbf{w}) = \sum_{i=1}^{d} \log(\epsilon + |\mathbf{w}_i|) \tag{17}$$

  - Minimax concave penalty (MCP): introduced in the context of linear regression [111], it is given by:

$$\Omega_{\lambda,\gamma}(\mathbf{w}) = \begin{cases} \lambda|\mathbf{w}_i| - \dfrac{|\mathbf{w}_i|^2}{2\gamma} & \text{if } |\mathbf{w}_i| \le \gamma\lambda \\ \dfrac{\gamma\lambda^2}{2} & \text{if } |\mathbf{w}_i| > \gamma\lambda \end{cases} \tag{18}$$

Table 1, Figs. 3 and 4 compare the properties of the different regularizers introduced above.

4.3.2 Structured regularization

In some cases, it is interesting to introduce sparsity by group of features based on the previous regularizers [99]. For a linear decision function, the weights of $\mathbf{w} \in \mathbb{R}^d$ can be decomposed intro groups (overlapping or not) $g \in \mathcal{G}$. For instance, when $d = 3$, the partition $\mathcal{G} = \{(1, 2), (3)\}$ contains two groups, the first one includes two variables (1 and 2)

Table 1 Properties of several regularization terms

	Standard regularization		
	Regularity	Convexity	Non convexity
• ℓ_2	✓	✓	–
• ℓ_1	–	✓	–
• Fused lasso	–	✓	–
• $\underset{0<p<1}{\ell_p}$	–	–	✓
• Log-sum	✓	–	✓
• MCP	✓	–	✓
• ℓ_0	–	–	✓

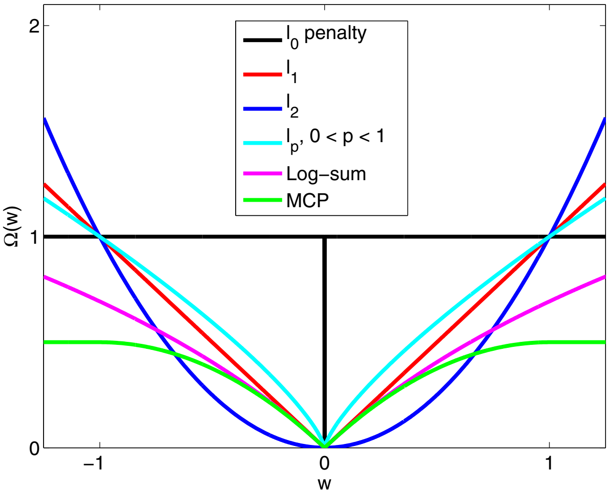


Fig. 3 Comparison of several unstructured regularization terms ($\epsilon = 1$ and $\gamma = 1$ are respectively the parameters of the log-sum and MCP regularizations)

when the second includes only the variable 3. The group regularization applied to the coefficients of **w** based on the mixed norm ℓ_p - ℓ_q is as follows:

$$\Omega_{p,q}(\mathbf{w}) = \sum_{g \in \mathcal{G}} \left(\left\| \mathbf{w}_g \right\|_q \right)^p \tag{19}$$

where $\mathbf{w}_g$ corresponds to the sub-vector of **w** corresponding to variables of the group g.

In the structured regularization, we can find the mixture ℓ_1 - ℓ_2 (also called the group Lasso), it represents the most known mixture of norms which applies the norm ℓ_1 to the sum of the ℓ_2 of each group leading to sparsity on the groups [4, 109]. There are some variants

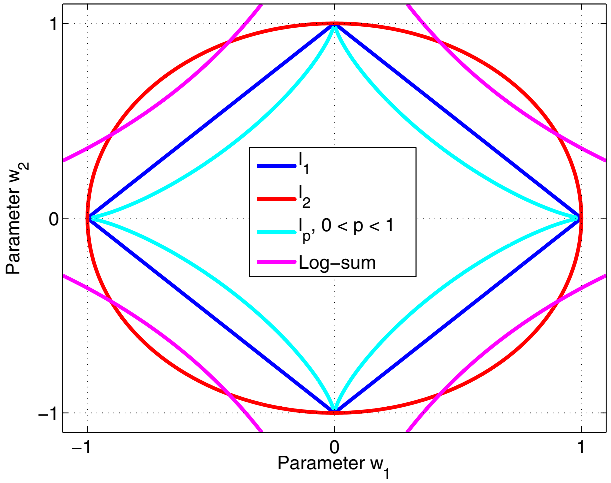


Fig. 4 Illustration 2D for several regularization terms

such as ℓ_p - ℓ_q where $0 < p < 1$ able to promote more sparsity. In the family of structured regularization, we can also find the group fused Lasso [12, 75, 77], which penalizes ℓ_1-norm of the difference between two successive groups of variable which leads to sparsity of groups difference. It is given by:

$$\Omega(\mathbf{W}) = \sum_{i=1}^{d-1} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|_1 \tag{20}$$

where $\mathbf{w}_i$ is the i^{th} group corresponding to the i^{th} row of $\mathbf{W}$.

5 Decomposition learning

Instead of selecting most relevant features or learning a mapping of the data in low dimensional another trend of feature representation attempts to find a sparse decomposition of the data over a learned dictionary. The problem of sparse decomposition has known growing interest. A very interesting task in this field is dictionary learning which attempts usually to design a dictionary capable to capture all or most information of the signal with a linear combination of a small number of elementary signals called dictionary atoms. Different from conventional predefined dictionaries such as wavelet basis, wavelet packet basis, Gabor atoms or Discrete Cosine Basis, dictionary learning allows more representation flexibility and efficiency in reconstruction and classification [22, 106]. Searching for the sparse representation of a signal over a dictionary is achieved by optimizing an objective function that consists of two terms: one that measures the reconstruction error and the other that measures the sparsity of the representation. Dictionary learning has been applied for different applications, such as image denoising [24, 60], inpainting [25, 60], clustering [18, 105] and classification [62].

It has been shown that the conventional dictionary learning algorithm is rather adapted for signal construction than classification [47]. Therefore, researchers introduced novel approaches more adapted for signal classification by taking the class label in consideration. Dictionary-based classification can be broadly divided into two main groups [47]:

- Discriminative dictionaries, such as Meta-face learning [107] and dictionary learning with structured incoherence [69].
- Discriminative coefficients, such as supervised dictionary learning [62], discriminative K-SVD [112], label consistent K-SVD [38] or fisher discriminant dictionary learning [108].

5.1 Conventional dictionary learning

Let n d-dimensional signals $\{\mathbf{x}_i\}_{i=1}^n$ stored in $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_i \cdots \mathbf{x}_n\ ] \in \mathbb{R}^{d\times n}$. The conventional learning approach attempts to find a dictionary (possibly overcomplete) of K atoms $\mathbf{D} = [\mathbf{d}_1 \cdots \mathbf{d}_k \cdots \mathbf{d}_K] \in \mathbb{R}^{d\times K}$ and the sparse coefficients $\mathbf{A} \in \mathbb{R}^{K\times n}$ corresponding to the representation of $\mathbf{X}$ over $\mathbf{D}$ by minimizing the following objective function:

$$\begin{cases} \min\limits_{\substack{\mathbf{D}\in\mathbb{R}^{d\times K}\\ \mathbf{A}\in\mathbb{R}^{K\times n}}} \|\mathbf{X} - \mathbf{DA}\|_F^2 + \lambda \|\mathbf{A}\|_1 \\ \\ \text{s.t} \quad \|\mathbf{d}_k\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \end{cases} \tag{21}$$

where $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_i \cdots \mathbf{a}_n]$ with $\mathbf{a}_i \in \mathbb{R}^K$ represents the coefficients of the representation of $\mathbf{x}_i$ over $\mathbf{D}$ and $\|\mathbf{A}\|_1 = \sum\limits_{i=1}^{n} \|\mathbf{a}_i\|_1$ a term promoting sparsity of each decomposition.

5.2 Discriminative dictionary

Let $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}\}_{i=1}^n$ where $\mathcal{Y} = \{1, \cdots, C\}$ is the label set. A method introduced in this context is dictionary learning with structured incoherence [69]. It attempts to learn a dictionary per class while enforcing incoherence in order to make dictionaries from different class as different as possible. The resulting optimization problem is:

$$\begin{cases} \min\limits_{\substack{\{\mathbf{D}_c\}_{c=1}^C \in \mathbb{R}^{d\times K} \\ \{\mathbf{A}_c\}_{c=1}^C \in \mathbb{R}^{K\times n}}} \sum_{c=1}^C \left\{ \|\mathbf{X}_c - \mathbf{D}_c\mathbf{A}_c\|_F^2 + \lambda \|\mathbf{A}_c\|_1 \right\} + \eta \sum\limits_{c=1}^{C} \sum_{\substack{j=1 \\ j\neq c}}^C \left\|\mathbf{D}_c^T\mathbf{D}_j\right\|_F^2 \\ \text{s.t} \quad \left\|\mathbf{d}_k^c\right\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \quad \forall c = 1, \cdots, C \end{cases} \tag{22}$$

where $\mathbf{X}_c$, $\mathbf{D}_c$ and $\mathbf{A}_c$ respectively correspond to the data from class c, the corresponding learned dictionary and the coefficients of representing $\mathbf{X}_c$ over $\mathbf{D}_c$. The first term in (22) represents the classical dictionary learning expression; the second term promotes orthogonality of learned $\mathbf{D}_c$ hence inducing their incoherence.

5.3 Discriminative coefficients

The most prominent method in the context of discriminative coefficients is the supervised dictionary method introduced in [62]. They incorporated a classification cost based on the logistic loss function:

$$\begin{cases} \min\limits_{\substack{\mathbf{D}\in\mathbb{R}^{d\times K} \\ \mathbf{A}\in\mathbb{R}^{K\times n} \\ \mathbf{w}\in\mathbb{R}^K \\ b\in\mathbb{R}}} \sum\limits_{i=1}^{n} (L(y_i f(\mathbf{x}_i, \mathbf{a}_i, \mathbf{w})) + \lambda_0 \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda_1 \|\mathbf{a}_i\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 \\ \text{s.t} \quad \|\mathbf{d}_k\|_2^2 \leq 1 \quad \forall k = 1, \cdots, K \end{cases} \tag{23}$$

where L represents the logistic loss function (Section 6.2) and $f(\mathbf{x}, \mathbf{a}, \mathbf{w}) = \mathbf{w}^T\mathbf{a} + b$ is a linear classification function depending on the learned decomposition coefficients $\mathbf{a}$ for the sample $\mathbf{x}$.

6 Classification

Classification methods can be broadly organized in two main groups: generative and discriminative approaches. The Generative classifiers learn a model of the joint probability $p(\mathbf{x}, y)$, of the inputs $\mathbf{x}$ and the label y, and make their predictions using Bayes rule to calculate $p(y|\mathbf{x})$, and then picking the most likely label y [11]. Discriminative classifiers model the posterior $p(y|\mathbf{x})$ directly, or learn a direct map from inputs $\mathbf{x}$ to the class label. There are several compelling reasons for using discriminative rather than generative classifiers, one of which, succinctly articulated by [97] is that ”one should solve the classification problem directly and never solve a more general problem as an intermediate step such as modeling $p(\mathbf{x}|y)$”. Indeed leaving aside computational issues and other matters, the prevailing consensus seems to be that discriminative classifiers are efficient alternatives to generative approaches. Indeed, the discriminative methods require few parameters to be determined ; they are not prone to a mis-specification of the joint distribution $p(\mathbf{x}, y)$.

Let suppose $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ where each sample $(\mathbf{x}, y)$ is drawn from an unknown joint distribution $\mathbb{P}(X, Y)$. The goal of classification is to find a decision function $f : \mathcal{X} \rightarrow \mathbb{R}$ capable to predict the correctly the label y' of a given observation $\mathbf{x}'$.

6.1 Regularized risk minimization

Learning a decision function could be based on a fixed structure such as k nearest neighbors, or by expressing the learning as an optimization problem. For this sake, a loss function L which measures the error between the predicted and real label is defined. Usually one seeks this function equals to 0 if the real and predicted labels are similar and greater than 0 otherwise. Theoretically, the best possible decision function is the one which minimizes the expected prediction error:

$$R(f) = \mathbb{E}[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(\mathbf{x}))\mathbb{P}(\mathbf{x}, y)\, \mathrm{d}y\mathrm{d}\mathbf{x} \tag{24}$$

Unfortunately, in practice $R(f)$ can not be minimized since the distribution $\mathbb{P}(X, Y)$ is unknown. However, an approximation called empirical risk, can be computed by averaging the loss function on the training set:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) \tag{25}$$

Minimizing $\hat{R}(f)$ with respect to f does not guarantee to obtain a function with good generalization properties (as overfitting can occur). Indeed, the minimization of empirical risk suffers from a lack of generalization and stability. Furthermore it has been demonstrated that the generalization and stability are linked, a stable problem implies generalization and vice versa [13, 63, 71]. To make the problem stable, a regularization term $\Omega(.)$ is added leading to the minimization of the structural risk [27, 97]. Usually one addresses the regularized empirical risk minimization:

$$\min_f \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda\Omega(f) \tag{26}$$

The first term is the classical empirical risk and the second is similar to the one introduced in (15) to (19). We refer the reader to [61] to have a broad overview of the usual regularizers.

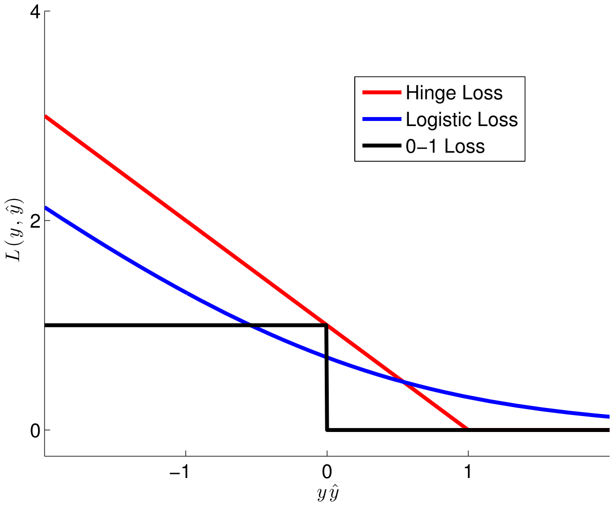


Fig. 5 Visualization of the loss functions

6.2 Loss function

There are numerous loss functions $L(y, \hat{y})$ measuring the error of prediction $\hat{y}$ of y. A large part of binary classification methods are based on learning a function capable to predict the class label using the sign of the predicted value. In this case the quantity used in the loss function is the product $y\hat{y}$. In the following we review a few most common loss functions shown in Fig. 5.

6.2.1 The 0-1 loss

It returns 0 if the class is well predicted and 1 otherwise. This cost is non differentiable and non-convex. Furthermore, the complexity of the resulting optimization problem is combinatorial which makes it very difficult to use in practice. It is given by:

$$L(y, \hat{y}) = (1 - sgn(y\hat{y}))/2 \tag{27}$$

6.2.2 The hinge loss

It is the cost used in the Support Vector Machines (SVM). Unlike the previous loss function, this cost is not necessarily equal to 0 when the class is well predicted. Hinge loss is equal to 0 only if $y\hat{y}$ is greater than 1, which means in other terms that $\hat{y}$ is predicted with some margin.

The hinge function is convex, however it needs a regularization term to make the problem strictly convex and ensure the uniqueness of the solution [86]. Its expression is:

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y}) \tag{28}$$

6.2.3 The logistic loss

It permits to learn probabilistic classifiers; the decision could be made based on the estimation of class conditional probability. In the binary classification case with $\mathcal{Y} = \{-1, 1\}$ this probability is:

$$\hat{P}(Y = y|X = \mathbf{x}) = \frac{1}{1 + \exp(-yf(\mathbf{x}))} \tag{29}$$

The logistic loss has the particularity to be strictly convex with value equals to 0 when $y\hat{y} = \infty$; it is given by:

$$L(y, f(x)) = \log(1 + \exp(-yf(x)) \tag{30}$$

7 Discussions

Obtaining good recognition performances relies mainly on finding suitable feature representations where observations from different classes are well separated. For the latter, huge efforts have been devoted to find adequate feature spaces which could offer these properties. Several approaches have been introduced, such as dimensionality reduction, feature selection and decomposition learning.

Dimensionality reduction is a common preprocessing step for classification. Learning a classifier on low-dimensional space is fast (despite learning the dimensionality reduction itself may be costly). Furthermore, dimensionality reduction can help learn a better classifier, particularly when the data do have an intrinsic low-dimensional structure at small scale since dimensionality reduction has a regularizing effect that can help avoid overfitting. This can be explained by the ability of dimensionality reduction to attenuate the impact of noise that perturbs the samples along the manifold.

The majority of supervised dimensionality reduction techniques usually encourage to learn a mapping $\mathbf{F}$ to push apart inputs having different labels. For classification, once the data is mapped into the low-dimensional space, a classifier g is learned on the pairs $(\mathbf{F}(\mathbf{x}_i), y_i)$. This clearly shows that $\mathbf{F}$ and g are separately learned and this gives an insight to jointly learn them for improved performances [8, 102].

From our point of view, despite the positive points of dimensionality reduction and feature selection techniques, we believe that the methods based on learning feature representations such as dictionary learning are more able to represent the data for classification purpose, since they have more flexibility to model the problem while introducing classification in the formalized problem and sparsity to avoid the overfitting.

We shall notice that a growing and intensive body of research, with the goal of end-to-end recognition system from feature extraction, representation and classification, is displayed by Deep Learning [9, 52]. The involved approaches proceed by giving raw signal as input features and by stacking more than the usual two neural layers. Each low level layer encodes specific properties of the signals as primitives that are gradually combined by successive higher level layers in order to produce representative and hopefully discriminative representations of the signals.

Among the deep learning models we can cite: i) Convolutional Neural Networks (CNN) [51], suited to represent invariance property; ii) Deep Boltzman Machine (DBM) [84] that can provide a generative model of the data; and iii) (Bidirectional) Long-Short Term Memory (BLSTM) [33] adapted for a recurrent representation, taking into account the temporal nature of the data. To be effective deep models require a huge amount of data, due to their complex structure coupled with their computing power to exhibit striking performances. When one lacks training data (as in the case of most of biometric applications), the conventional techniques provide valuable alternatives.

8 Conclusion

The performance of the biometric recognition systems strongly dependents on learning a proper suitable feature representation space where samples from different classes are well presented and separated. In this paper, we have presented a comprehensive study of the existing state-of-the-art feature representation techniques. This was carried out by introducing simple and clear taxonomies as well as well as simple and effective explanation of the prominent techniques. This is intended to guide the neophyte and provide researchers with state-of-the-art approaches in order to help advance the research topic in biometrics.

Acknowledgments This publication was made possible using a grant from the Qatar National Research Fund through National Priority Research Program (NPRP) # NPRP 8-140-2-065. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund or Qatar University.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Al Maadeed S, Jiang X, Rida I, Bouridane A (2018) Palmprint identification using sparse and dense hybrid representation. Multimedia Tools and Applications:1–15. https://doi.org/10.1007/s11042-018-5655-8
2. Amari S (1999) Natural gradient learning for over-and under-complete bases in ICA. Neural Comput 11(8):1875–1883. https://doi.org/10.1162/089976699300015990
3. Archibald R, Fann G (2007) Feature selection and classification of hyperspectral images with support vector machines. IEEE Geosci Remote Sens Lett 4(4):674–677. https://doi.org/10.1109/LGRS.2007.905116
4. Bach F (2008) Consistency of the group lasso and multiple kernel learning. J Mach Learn Res 9:1179–1225
5. Bakshi S, Tuglular T (2013) Security through human-factors and biometrics. In: Proceedings of the 6th International Conference on Security of Information and Networks. ACM, pp 463–463. https://doi.org/10.1145/2523514.2523597
6. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550. https://doi.org/10.1109/72.298224
7. Belkin M, Niyogi P (2001) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: NIPS, vol 14, pp 585–591
8. Bellet A, Habrard A, Sebban M (2013) A survey on metric learning for feature vectors and structured data. arXiv:1306.6709
9. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50
10. Bishop CM (2006) Pattern recognition. Machine Learning
11. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin
12. Bleakley K, Vert JP (2011) The group fused lasso for multiple change-point detection. arXiv:1106.4199

13. Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res 2:499–526
14. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
15. Chang X, Nie F, Yang Y, Huang H (2014) A convex formulation for semi-supervised multi-label feature selection. In: AAAI, pp 1171–1177
16. Chang X, Nie F, Yang Y, Zhang C, Huang H (2016) Convex sparse PCA for unsupervised feature learning. ACM Trans Knowl Discov Data (TKDD) 11(1):3. https://doi.org/10.1145/29105854
17. Chen X, Yuan G, Wang W, Nie F, Chang X, Huang JZ (2018) Local adaptive projection framework for feature selection of labeled and unlabeled data. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2018.2830186
18. Cheng B, Yang J, Yan S, Fu Y, Huang TS (2010) Learning with ℓ_1-graph for image analysis. IEEE Trans Image Process 19(4):858–866. https://doi.org/10.1109/TIP.2009.2038764
19. Cunningham JP, Ghahramani Z (2015) Linear dimensionality reduction: survey, insights, and generalizations. J Mach Learn Res 16:2859–2900
20. d'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GR (2007) A direct formulation for sparse PCA using semidefinite programming. SIAM Rev 49(3):434–448. https://doi.org/10.1137/050645506
21. De Ridder D, Duin RP, Kittler J (2002) Texture description by independent components. In: Structural, Syntactic, and Statistical Pattern Recognition. Springer, pp 587–596. https://doi.org/10.1007/3-540-70659-3_61
22. Diamant I, Klang E, Amitai M, Konen E, Goldberger J, Greenspan H (2017) Task-driven dictionary learning based on mutual information for medical image classification. IEEE Trans Biomed Eng 64(6):1380–1392. https://doi.org/10.1109/TBME.2016.2605627
23. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1(1):269–271
24. Elad M, Aharon M (2006) Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process 15(12):3736–3745. https://doi.org/10.1109/TIP.2006.881969
25. Elad M, Figueiredo MA, Ma Y (2010) On the role of sparse and redundant representations in image processing. Proc IEEE 98(6):972–982. https://doi.org/10.1109/JPROC.2009.2037655
26. Eshelman LJ (2014) The CHC adaptive search algorithm: How to have safe search when engaging. Found Genet Algorithm 1991 (FOGA 1(1):265. https://doi.org/10.1016/B978-0-08-050684-5.50020-3
27. Evgeniou T, Poggio T, Pontil M, Verri A (2002) Regularization and statistical learning theory for data analysis. Comput Stat Data Anal 38(4):421–432. https://doi.org/10.1016/S0167-9473(01)00069-X
28. Fan M, Chang X, Tao D (2017) Structure regularized unsupervised discriminant feature analysis. In: AAAI, pp 1870–1876
29. Fei L, Teng S, Wu J, Rida I (2017) Enhanced minutiae extraction for high-resolution palmprint recognition. Int J Image Graph 17(04):1750,020. https://doi.org/10.1142/S0219467817500206
30. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7(2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x 4
31. Floyd RW (1962) Algorithm 97: shortest path. Commun ACM 5(6):345
32. Goldberg DE et al (1989) Genetic algorithms in search optimization and machine learning, vol 412. Addison-wesley, Reading. ISBN: 0201157675
33. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5):602–610. https://doi.org/10.1016/j.neunet.2005.06.042
34. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
35. Hastie T, Tibshirani R, Friedman J (2001) Springer series in statistics. The elements of statistical learning: Data mining, inference and prediction. https://doi.org/10.1007/978-0-387-84858-7
36. He X, Cai D, Yan S, Zhang HJ (2005) Neighborhood preserving embedding. In: IEEE International conference on computer vision, vol 2, pp 1208–1213. https://doi.org/10.1109/ICCV.2005.167
37. Hyvärinen A, Karhunen J, Oja E (2004) Independent component analysis, vol 46. Wiley, New York
38. Jiang Z, Lin Z, Davis LS (2011) Learning a discriminative dictionary for sparse coding via label consistent k-svd. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1697–1704. https://doi.org/10.1109/CVPR.2011.5995354
39. Jimenez LO, Landgrebe DA (1998) Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Trans Syst Man Cybern Part C: Appl Rev 28(1):39–54. https://doi.org/10.1109/5326.661089
40. John GH, Kohavi R, Pfleger K et al (1994) Irrelevant features and the subset selection problem. In: Machine learning: proceedings of the eleventh international conference, pp 121–129. https://doi.org/10.1016/B978-1-55860-335-6.50023-4

41. Joho M, Mathis H, Lambert RH (2000) Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture. In: Proceedings of International conference on independent component analysis and blind signal separation. Helsinki, pp 81–86
42. Journée M., Nesterov Y, Richtárik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. J Mach Learn Res 11:517–553
43. Kao YH, Van Roy B (2013) Learning a factor model via regularized PCA. Mach Learn 91(3):279–303. https://doi.org/10.1007/s10994-013-5345-8
44. Kennedy J (2011) Particle swarm optimization. In: Encyclopedia of machine learning. Springer, pp 760–766. https://doi.org/10.1007/978-0-387-30164-8
45. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1):273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
46. Kokiopoulou E, Saad Y (2007) Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique. IEEE Trans Pattern Anal Mach Intell 29(12):2143–2156. https://doi.org/10.1109/TPAMI.2007.1131
47. Kong S, Wang D (2012) A brief summary of dictionary learning based approach for classification (revised). arXiv:1205.6544
48. Langley P et al (1994) Selection of relevant features in machine learning. Defense Technical Information Center
49. Law MH, Figueiredo MA, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Trans Pattern Anal Mach Intell 26(9):1154–1166. https://doi.org/10.1109/TPAMI.2004.71
50. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, De Schaetzen V, Duque R, Bersini H, Nowe A (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinforma (TCBB) 9(4):1106–1119. https://doi.org/10.1109/TCBB.2012.33
51. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
52. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
53. Liu H, Setiono R et al (1996) A probabilistic approach to feature selection-a filter solution. In: ICML. Citeseer, vol 96, pp 319–327
54. Luo M, Nie F, Chang X, Yang Y, Hauptmann A, Zheng Q (2016) Avoiding optimal mean robust pca/2dpca with non-greedy ℓ_1 norm maximization. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence
55. Luo M, Nie F, Chang X, Yang Y, Hauptmann AG, Zheng Q (2017) Avoiding optimal mean $\ell_{2,1}$, norm maximization-based robust pca for reconstruction. Neural computation 29(4):1124–1150. https://doi.org/10.1162/NECO_a_00937
56. Luo M, Chang X, Li Z, Nie L, Hauptmann A, Zheng Q (2017) Simple to complex cross-modal learning to rank. Comput Vis Image Underst 163:67–77. https://doi.org/10.1016/j.cviu.2017.07.001
57. Luo M, Chang X, Nie L, Yang Y, Hauptmann A, Zheng Q (2018) An adaptive semisupervised feature analysis for video semantic recognition. IEEE Trans Cybern 48(2):648–660. https://doi.org/10.1109/TCYB.2017.2647904
58. Luo M, Nie F, Chang X, Yang Y, Hauptmann A, Zheng Q (2018) Adaptive unsupervised feature selection with structure regularization. IEEE Trans Neural Netw Learn Syst 29(4):944–956. https://doi.org/10.1109/TNNLS.2017.2650978
59. Ma Z, Chang X, Xu Z, Sebe N, Hauptmann A (2017) Joint attributes and event analysis for multimedia event detection. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2017.2709308
60. Mairal J, Elad M, Sapiro G (2008) Sparse representation for color image restoration. IEEE Trans Image Process 17(1):53–69. https://doi.org/10.1109/TIP.2007.911828
61. Mairal J, Bach F, Ponce J (2014) Sparse modeling for image and vision processing. arXiv:1411.3230
62. Mairal J, Ponce J, Sapiro G, Zisserman A, Bach F (2009) Supervised dictionary learning. In: Advances in neural information processing systems, pp 1033–1040
63. Mukherjee S, Rifkin R, Poggio T (2003) Regression and classification with regularization. In: Nonlinear estimation and classification. Springer, pp 111–128. https://doi.org/10.1007/978-0-387-21579-2_7
64. Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. IEEE Trans Comput 100(9):917–922. https://doi.org/10.1109/TC.1977.1674939
65. Pearson K (1901) Liii. on lines and planes of closest fit to systems of points in space. The London. Edinb Dublin Philos Mag J Sci 2(11):559–572. https://doi.org/10.1080/14786440109462720

66. Platt J et al (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 10(3):61–74
67. Porrill J, Stone JV (1998) Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, Citeseer
68. Poultney C, Chopra S, Cun YL et al (2006) Efficient learning of sparse representations with an energy-based model. In: Advances in neural information processing systems, pp 1137–1144
69. Ramirez I, Sprechmann P, Sapiro G (2010) Classification and clustering via dictionary learning with structured incoherence and shared features. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3501–3508. https://doi.org/10.1109/CVPR.2010.5539964
70. Rida I, Almaadeed S, Bouridane A (2014) Improved gait recognition based on gait energy images. In: 26Th international conference on microelectronics (ICM). IEEE, pp 40–43. https://doi.org/10.1109/ICM.2014.7071801
71. Rida I, Herault R, Gasso G (2014) Supervised music chord recognition. In: 2014 13th international conference on Machine learning and applications (ICMLA). IEEE, pp 336–341. https://doi.org/10.1109/ICMLA.2014.60
72. Rida I, Al Maadeed S, Bouridane A (2015) Unsupervised feature selection method for improved human gait recognition. In: 23Rd european signal processing conference (EUSIPCO). IEEE, pp 1128–1132. https://doi.org/10.1109/EUSIPCO.2015.7362559
73. Rida I, Bouridane A, Marcialis GL, Tuveri P (2015) Improved human gait recognition. In: Image Analysis and Processing (ICIAP). Springer, pp 119–129. https://doi.org/10.1007/978-3-319-23234-8_12
74. Rida I, Boubchir L, Al-Maadeed N, Al-Maadeed S, Bouridane A (2016) Robust model-free gait recognition by statistical dependency feature selection and globality-locality preserving projections. In: 39Th international conference on telecommunications and signal processing (TSP). IEEE, pp 652–655. https://doi.org/10.1109/TSP.2016.7760963
75. Rida I, Jiang X, Marcialis GL (2016) Human body part selection by group lasso of motion for model-free gait recognition. IEEE Signal Process Lett 23(1):154–158. https://doi.org/10.1109/LSP.2015.2507200
76. Rida I, Almaadeed S, Bouridane A (2016) Gait recognition based on modified phase-only correlation. SIViP 10(3):463–470. https://doi.org/10.1007/s11760-015-0766-4
77. Rida I, Al Maadeed N, Marcialis GL, Bouridane A, Herault R, Gasso G (2017) Improved model-free gait recognition based on human body part. In: Biometric Security and Privacy. Springer, pp 141–161. https://doi.org/10.1007/978-3-319-47301-7_6
78. Rida I, Al-maadeed N, Al-maadeed S (2018) Robust gait recognition: a comprehensive survey. IET Biometrics. https://doi.org/10.1049/iet-bmt.2018.5063
79. Rida I, Al-Maadeed S, Mahmood A, Bouridane A, Bakshi S (2018) Palmprint identification using an ensemble of sparse representations. IEEE Access 6:3241–3248. https://doi.org/10.1109/ACCESS.2017.2787666
80. Rida I, Herault R, Marcialis GL, Gasso G (2018) Palmprint recognition with an efficient data driven ensemble classifier. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2018.04.033
81. Rida I, Maadeed SA, Jiang X, Lunke F, Bensrhair A (2018) An ensemble learning method based on random subspace sampling for palmprint identification. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2047–2051. https://doi.org/10.1109/ICASSP.2018.8462051
82. Roweis S (1998) EM algorithms for PCA and SPCA. Advances in neural information processing systems:626–632
83. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326. https://doi.org/10.1126/science.290.5500.2323
84. Salakhutdinov R, Hinton GE (2009) Deep boltzmann machines. In: AISTATS, vol 1, p 3
85. Scholkopf B, Smola A, Müller KR (1999) Kernel principal component analysis. In: Advances in Kernel Methods-Support Vector Learning
86. Scholkopf B, Smola AJ (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge. ISBN: 0262194759
87. Silva VD, Tenenbaum JB (2002) Global versus local methods in nonlinear dimensionality reduction. In: Advances in neural information processing systems, pp 705–712
88. Spearman C (1904) General intelligence, objectively determined and measured. Amer J Psychol 15(2):201–292. https://doi.org/10.2307/1412107
89. Subrahmanya N, Shin YC (2010) Sparse multiple kernel learning for signal processing applications. IEEE Trans Pattern Anal Mach Intell 32(5):788–798. https://doi.org/10.1109/TPAMI.2009.98
90. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323. https://doi.org/10.1126/science.290.5500.2319

91. Theis FJ, Lang EW, Puntonet CG (2004) A geometric algorithm for overcomplete linear ICA. Neurocomputing 56:381–398. https://doi.org/10.1016/j.neucom.2003.09.008
92. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc. Series B (Methodological) 58:267–288
93. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J Royal Stat Soc Ser B (Stat Methodol) 67(1):91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x
94. Tipping M, Bishop CM (1999) Probabilistic principal component analysis. J Royal Stat Soc Ser B (Stat Methodol) 61(3):611–622. https://doi.org/10.1111/1467-9868.00196
95. Torgerson WS (1952) Multidimensional scaling: i. theory and method. Psychometrika 17(4):401–419. https://doi.org/10.1007/BF02288916
96. Van Der Maaten L, Postma E, Van den Herik J (2009) Dimensionality reduction: a comparative. J Mach Learn Res 10:66–71
97. Vapnik V (1995) The nature of statistical learning theory. Springer Science & Business Media, Berlin
98. Verleysen M, François D (2005) The curse of dimensionality in data mining and time series prediction. In: Computational Intelligence and Bioinspired Systems. Springer, pp 758–770. https://doi.org/10.1007/11494669_93
99. Wang S, Chang X, Li X, Long G, Yao L, Sheng QZ (2016) Diagnosis code assignment using sparsity-based disease correlation embedding. IEEE Trans Knowl Data Eng 28(12):3191–3202. https://doi.org/10.1109/TKDE.2016.2605687
100. Wang R, Nie F, Hong R, Chang X, Yang X, Yu W (2017) Fast and orthogonal locality preserving projections for dimensionality reduction. IEEE Trans Image Process 26(10):5019–5030. https://doi.org/10.1109/TIP.2017.2726188
101. Wang S, Li X, Yao L, Sheng QZ, Long G et al (2017) Learning multiple diagnosis codes for ICU patients with local disease correlation mining. ACM Trans Knowl Discov Data (TKDD) 11(3):31. https://doi.org/10.1145/3003729
102. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10(Feb):207–244
103. Welling M, Zemel RS, Hinton GE (2004) Probabilistic sequential independent components analysis. IEEE Trans Neural Netw 15(4):838–849. https://doi.org/10.1109/TNN.2004.828765
104. Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero norm with linear models and kernel methods. J Mach Learn Res 3:1439–1461
105. Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S (2010) Sparse representation for computer vision and pattern recognition. Proc IEEE 98(6):1031–1044. https://doi.org/10.1109/JPROC.2010.2044470
106. Wu L, Wang Y, Pan S (2016) Exploiting attribute correlations: a novel trace lasso-based weakly supervised dictionary learning method. IEEE Trans Cybern 47(12):4497–4508. https://doi.org/10.1109/TCYB.2016.2612686
107. Yang M, Zhang L, Yang J, Zhang D (2010) Metaface learning for sparse representation based face recognition. In: IEEE International conference on image processing (ICIP), pp 1601–1604. https://doi.org/10.1109/ICIP.2010.5652363
108. Yang M, Zhang L, Feng X, Zhang D (2011) Fisher discrimination dictionary learning for sparse representation. In: IEEE International conference on computer vision (ICCV), pp 543–550. https://doi.org/10.1109/ICCV.2011.6126286
109. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J Royal Stat Soc Ser B (Stat Methodol) 68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x
110. Zhang LQ, Cichocki A, Amari S (1999) Natural gradient algorithm for blind separation of overdetermined mixture with additive noise. IEEE Signal Process Lett 6(11):293–295. https://doi.org/10.1109/97.796292
111. Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942
112. Zhang Q, Li B (2010) Discriminative k-svd for dictionary learning in face recognition. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2691–2698. https://doi.org/10.1109/CVPR.2010.5539989
113. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15(2):265–286. https://doi.org/10.1198/106186006X113430

Imad Rida received the master's degree in biometric security from the University of Paris Est Creteil in 2012, and Ph.D. degree in computer science from Normandie University in 2017. He was a Visiting Researcher at Qatar University. He is currently an Assistant Professor at the Department of Information Systems Architectures, INSA Rouen Normandie. He is keenly interested in exploring the applications of machine learning techniques in biometrics. His main areas of research are machine learning, computer vision, and pattern recognition.

Noor Al-Maadeed is an assistant professor at the computer science and Engineering Department at Qatar University. In 2014 she receive her PhD in computer science and engineering from Brunel University. She graduate from the Qatar Leadership Centre, first batch 2013. She has been a Lecturer of Computer Engineering in Qatar University since 2001. Her areas of research are speech signal detection, speaker identification and audio/visual speaker recognition. Honours and Awards Qatar Education Excellence Day Platinum Award - New PhD Holders Category 2014-2015.

Somaya Al-Maadeed received the Ph.D. degree in computer science from Nottingham, U.K., in 2004. She enjoys excellent collaboration with national and international institutions and industry. She was a Visiting Academic at Northumbria University, U.K. She is currently the Head of the Computer Science Department at Qatar University (QU). She is also the coordinator of the Computer Vision Research Group at QU. She published extensively in computer vision and pattern recognition. She organized several workshops and competitions related to biometrics and computer vision. She was selected as a participant in Current and Future Executive Leaders Program at Qatar Leadership Centre (2012-2013).

Sambit Bakshi received the Ph.D. degree in computer science, in 2015. He is currently with the Centre for Computer Vision and Pattern Recognition, National Institute of Technology Rourkela, India. He also serves as an assistant professor with the Department of Computer Science and Engineering, National Institute of Technology Rourkela. His research interest includes visual surveillance and biometric security. He serves as an associate editor of IEEE Access (2016 −), Plos One (2017 −), Innovations in Systems and Software Engineering - A NASA Journal (2016 −), International Journal of Biometrics (2013 −), and Expert Systems, Wiley (2018 −). He is a technical committee member of the IEEE Computer Society Technical Committee on Pattern Analysis and Machine Intelligence. He received the prestigious Innovative Student Projects Award in 2011 from the Indian National Academy of Engineering for his master’s thesis. He has more than 50 publications in journals, reports, and conferences. He is a member of the IEEE.