Universidad de Oviedo

Programa de Doctorado en Matemáticas y Estadística (RD 99/2011)

IMAGE INDEXING OPTIMIZATION.

BIOMETRIC AND BIOMEDICAL APPLICATIONS

PhD Thesis

PhD Candidate:
Ana Cernea

Supervisor:
Juan Luis Fernández Martínez

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: OPTIMIZACIÓN DEL INDEXADO AUTOMÁTICO DE IMÁGENES DIGITALES. APLICACIONES BIOMÉTRICAS Y BIOMÉDICAS	Inglés: IMAGE INDEXING OPTIMIZATION BIOMETRIC AND BIOMEDICAL APPLICATIONS

2.- Autor	
Nombre: DOINA ANA CERNEA CORBEANU	DNI/Pasaporte/NIE:
Programa de Doctorado: Programa de Doctorado en Matemáticas y Estadística	
Órgano responsable: DEPARTAMENTO DE MATEMÁTICAS	

RESUMEN (en español)

La interpretación automática de imágenes es un reto de la inteligencia artificial, el reconocimiento de patrones y la visión por ordenador, con aplicaciones en seguridad (autentificación), biometría y biomedicina, por nombrar algunas de ellas. En esta tesis doctoral se desarrolla una metodología robusta para analizar y resolver este tipo de problemas, consistente en:

- **Analizar y comparar los principales métodos de descomposición espectral, utilizados en la reducción de dimensión en este tipo de problemas.**
- **Diseñar un algoritmo de aprendizaje por consenso, basado en diferentes clasificadores k-NN por vecino próximo, construidos a partir de diferentes tipos de atributos de carácter diverso, tales como, atributos espectrales, estadísticos y descriptores regionales, de las diferentes imágenes del conjunto de aprendizaje (o entrenamiento). El análisis de diversidad de los clasificadores posee una capital importancia en este diseño, para reducir la ambigüedad en la clasificación. Dicho clasificador pertenece originalmente a la categoría de aprendizaje no supervisado.**
- **Explorar y explotar el espacio de incertidumbre generado mediante la generalización de dicho algoritmo de consenso a la categoría de aprendizaje supervisado, con objeto de mejorar la precisión de la predicción, y proporcionar adicionalmente un conjunto de soluciones plausibles del problema de identificación, que constituyen una muestra representativa de la incertidumbre intrínseca de dicho problema biométrico.**

Esta metodología ha sido aplicada al problema de reconocimiento de caras y al análisis del grado histológico de cánceres de mama triplemente negativos, a partir de imágenes histológicas de los mismos. En ambos casos se han obtenido resultados muy prometedores. Futuras investigaciones serán dedicadas al estudio de otras técnicas biométricas (iris, huellas digitales, etc.), y también otras aplicaciones biomédicas, consistentes en el diseño de robots capaces de asesorar a los médicos en la toma de decisiones.

RESUMEN (en Inglés)

Automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision with applications in security authentication, biometry, and biomedicine, to name some of them. This PhD thesis develops a robust methodology to deal with these important problems, by:

- **Analyzing and comparing the main spectral decomposition methods used for dimensionality reduction.**
- **Designing an ensemble learning algorithm based on different nearest-neighbor classifiers that are built from diverse image attributes, concerning spectral, statistical and regional description features of the learning database. Diversity analysis of the classifiers has a paramount importance in this design to reduce ambiguity in the classification. This algorithm originally belongs to the category of unsupervised learning.**
- **Exploring and exploiting the uncertainty space generated by the supervised approach of the ensemble learning algorithm, in order to improve the accuracy of the predictions, and additionally providing a set of plausible solutions of the image identification problem that accounts for its intrinsic uncertainty.**

This methodology has been applied to the face recognition problem and to the analysis of the histological degree of Triple Negative Breast Cancers (TNBC), providing in both cases very promising results. Future research will be devoted to analyze other biometric techniques (iris recognition, fingerprint identification, etc), and also, other important biomedical applications, such as the design of biomedical robots able to asses medical doctors in their decisions.

SR. DIRECTOR DE DEPARTAMENTO DEL DEPARTAMENTO DE MATEMÁTICAS

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN MATEMÁTICAS Y ESTADÍSTICAS

IMAGE INDEXING OPTIMIZATION. BIOMETRIC AND BIOMEDICAL APPLICATIONS

By

Ana Cernea

To Silvia and Cristian.

Table of Contents

General Setup

The problem addressed in this PhD thesis is the automatic image interpretation and recognition which is an important challenge in artificial intelligence, pattern recognition and computer vision, with applications in face recognition and biomedicine.

This thesis is included in the PhD Programme in Mathematics and Statistics (RD 99/2011) of the University of Zaragoza, the University of La Laguna, the University of Oviedo, the University of the Basque Country and the Public University of Navarra. In the elaboration of the manuscript of this PhD thesis we followed the requirements fixed by Article 26 of the Regulations of PhD Studies, agreement of June 17 of 2013 (BOPA 146 / 25-VI-2013) about the nature of the PhD thesis, that literally states: *"1. La tesis doctoral consistirá en un trabajo original de investigación elaborado por el doctorando en cualquier campo del conocimiento. La tesis debe capacitar al estudiante de doctorado para el trabajo autónomo en el ámbito de la I+D+i. 2. En su elaboración, habrán de ser tenidas en cuenta las siguientes normas mínimas: a) La memoria que recoge la labor realizada en la tesis doctoral se redactará en español. No obstante, la Comisión de Doctorado podrá autorizar su redacción en otro idioma oficial de la Unión Europea, previo informe de la Comisión Académica del Programa de Doctorado, y siempre que se garantice que los miembros del Tribunal estén en condiciones de juzgarla. En este caso, la memoria deberá contener el resumen y las conclusiones en español. En las mismas condiciones, y de acuerdo con el artículo 6.2 de los Estatutos de la Universidad de Oviedo, la redacción podrá hacerse en lengua asturiana. b) En la cubierta de la memoria figurará Universidad de Oviedo, junto con el escudo institucional, el nombre*

del Programa de Doctorado, el título de la tesis y el nombre del autor. c) Los datos anteriores aparecerán también en la portada, y en las páginas siguientes figurará la autorización de la Comisión Académica del Programa de Doctorado, del tutor y del director del trabajo para la presentación de la tesis."

The present work belongs to the line of research dedicated to Biometric Applications, of the research group for Inverse Problems, Optimization and Learning, directed by Professor Juan Luis Fernández Martínez from the Mathematics Department of Oviedo University.

Before deciding to approach this challenging and interesting problem, my MSc project, treated on the automatic labeling of multimedia contents for the Semantic Web, under the supervision of Dr. José Emilio Labra Gayo from the Informatics Department of Oviedo University. The Semantic Web problem can be regarded as a particular case of the methodology that is shown in this PhD thesis. Future research work will be devoted to the application of our methodology to this problem.

The main results of this research were published (or in revision) in INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE that had an Impact Factor of 0,562 in 2012. The complete information on this journal is:

- Category Name: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE.
- Quartile in Category: $Q4$.
- Journal Rank in Category: 94/115.

The selection of this journal was based on the fact that, although this interesting problem belongs to the branch of Applied Mathematics, biometric problems are addressed mainly in Computer Science, Pattern Recognition and Artificial Intelligence.

Besides, we also presented two communications to International Conferences, and finally, a book chapter edited by Springer.

In each case, the article is presented in original form and it is preceded by an explanation of the methodology that is used, and the main original results that are achieved, focussing in each case on some specific topics and concepts used in the research, that needed a more detailed description out of the cited articles.

Consequently, the structure of this thesis follows the publications that constitute the main steps of our research, and which determine the body of this manuscript.

Acknowledgements

I would like to thank to Juan Luis Fernández Martínez, my supervisor who dedicated a lot of time and energy in the achievement of this PhD thesis. He also guided me through the shortest and most simple way in the art of mathematical modelling.

I would also like to thank to all our collaborators, specially to Anatomical Pathology Service from the Hospital Universitario Central de Asturias (Oviedo, Spain) for providing all the facilities to acquire and treat the TBNC samples that were used in this PhD thesis, and to Emilio Suárez Díaz for his support and help to the edition of this Latex manuscript.

In addition, I would like to thank Professor Benjamin Dugnol who introduced me to Digital Signal Processing when I first joined the Mathematics Department of Oviedo University.

Resumen

La interpretación automática de imágenes es un reto de la inteligencia artificial, el reconocimiento de patrones y la visión por ordenador, con aplicaciones en seguridad (autentificación), biometría y biomedicina, por nombrar algunas de ellas. En esta tesis doctoral se desarrolla una metodología robusta para analizar y resolver este tipo de problemas, consitente en:

1. Analizar y comparar los principales métodos de descomposición espectral, utilizados en la reducción de dimensión en este tipo de problemas.

2. Diseñar un algoritmo de aprendizaje por consenso, basado en diferentes clasificadores k-NN por vecino próximo, construidos a partir de diferentes tipos de atributos de carácter diverso, tales como, atributos espectrales, estadísticos y descriptores regionales, de las diferentes imágenes del conjunto de aprendizaje (o entrenamiento). El análisis de diversidad de los clasificadores posee una capital importancia en este diseño, para reducir la ambiguedad en la clasificación. Dicho clasificador pertenece originalmente a la categoría de aprendizaje no supervisado.

3. Explorar y explotar el espacio de incertidumbre generado mediante la generalización de dicho algoritmo de consenso a la categoría de aprendizaje supervisado, con objeto de mejorar la precisión de la predicción, y proporcionar adicionalmente un conjunto de soluciones plausibles del problema de identificación, que constituyen una muestra representativa de la incertidumbre intrínseca de dicho problema biométrico.

Esta metodología ha sido aplicada al problema de reconocimiento de caras y al análisis del grado histológico de cánceres de mama triplemnente negativos, a partir de imágenes histológicas de los mismos. En ambos casos se han obtenido resultados muy prometedores. Futuras investigaciones serán dedicadas al estudio de otras técnicas biométricas (iris, huellas digitales, etc), y también otras aplicaciones biomédicas, consistentes en el diseño de robots capaces de asesorar a los médicos en la toma de decisiones.

Abstract

Automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision with applications in security authentication, biometry, and biomedicine, to name some of them. This PhD thesis develops a robust methodology to deal with these important problems, by:

1. Analyzing and comparing the main spectral decomposition methods used for dimensionality reduction.

2. Designing an ensemble learning algorithm based on different nearest-neighbor classifiers that are built from diverse image attributes, concerning spectral, statistical and regional description features of the learning database. Diversity analysis of the classifiers has a paramount importance in this design to reduce ambiguity in the classification. This algorithm originally belongs to the category of unsupervised learning.

3. Exploring and exploiting the uncertainty space generated by the supervised approach of the ensemble learning algorithm, in order to improve the accuracy of the predictions, and additionally providing a set of plausible solutions of the image identification problem, that accounts for its intrinsic uncertainty.

This methodology has been applied to the face recognition problem and to the analysis of the histological degree of Triple Negative Breast Cancers (TNBC), providing in both cases very promising results. Future research will be devoted to analyze other biometric techniques (iris recognition, fingerprint identification, etc), and also, other important biomedical applications, such as the design of biomedical robots able to asses medical doctors in their decisions.

Introduction

Designing systems for the automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision with multiple applications in many fields such as biometry, biomedicine, information security and law enforcement, to name some of them. The methodology developed in this PhD thesis is based on biometric and biomedical applications, in particular, on the face recognition problem and basal images of TNBC classification.

Our choice of biometry among the vast fields of application of image processing is due to its great importance and interest in science and technology. A large manifold of industry areas such as information security (data access, user authentication), security authentication (persons authentication in national IDs, Passports, border controls, video surveillance and suspect tracking) [4] and biomedical applications [24] are based on biometric techniques.

Biometrics is the automatic person identification using physiological or behavioral characteristics, such as fingerprint, face, iris, retina, palm-print, etc. Among all the biometric techniques, fingerprint recognition is still the most popular method and is successfully used in many authentication applications [18]. Nevertheless, face recognition has the benefit of being a passive, non intrusive system for verifying personal identity and therefore it responds to the increasing requirement for reliable personal identification in computerized access control. A survey of popular face recognition techniques can be found in [33].

1. Problem Definition

The automatic image recognition problem consists in classifying a given probe image I providing a database of training images. Given a database of training images (the

learning database)

$$Bd = \left\{ I_k \in \mathcal{M}_{(n,m,c)} \left(\mathbb{N} \right) : k = 1, \ldots, N \right\}, \tag{0.0.1}$$

organized into q classes:

$$C = \left\{ C_k \in \left\{ 1, 2, ..., q \right\}, k = 1, \ldots, N \right\}, \tag{0.0.2}$$

and a new incoming image $I \notin Bd$, the problem consists in estimating its class $C_I^* \in C$, according to the different poses of I that are at disposal in Bd.

In this definition $\mathcal{M}_{(n,m,c)}$ is the space of RGB color images (if $c = 3$), or gray-scale images (if $c = 1$) of size $m \times n$, and the learning database typically contains N_p poses for each class $C_i \in C$ $(i = 1, ..., q)$.

To perform the classification, a learning algorithm, L^* is defined as follows: L^* is a functional defined over the set of images $\mathcal{M}_{(n,m,c)}$ into the set of classes C:

$$L^* : \mathcal{M}_{(n,m,c)} \to C, \quad L^*(I_j) = C_j. \tag{0.0.3}$$

L^* is based on a nearest-neighbor criterion, that is, finding the image $I_k \in Bd$ such as:

$$d\left(I, I_k\right) = \min_{I_j \in Bd} d\left(I, I_j\right), \tag{0.0.4}$$

where d is a suitable distance (or norm) defined over $\mathcal{M}_{(n,m,c)}$. Typically, these distances are defined over the different attribute spaces, as follows:

$$d_k\left(I_i, I_j\right) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \tag{0.0.5}$$

where $\mathbf{v}_i^k \in \mathbb{R}^{s_k}$ represents the feature vector of length s_k corresponding to image I_i according to the attribute k, and p is the norm (usually $p = 1$ or $p = 2$ depending on the method) defined over the k-th attribute vector space ($\mathbb{R}^{s_k}$). Typically, the L_2 norm ($p = 2$) is used and it is related to the Euclidean scalar product. In this case it is also possible to maximize the cosine of the angle between these two images according to their attributes, as follows:

$$cos_k\left(I_i, I_j\right) = \frac{\mathbf{v}_i^k \cdot \mathbf{v}_j^k}{\|\mathbf{v}_i^k\|_2 \cdot \|\mathbf{v}_j^k\|_2}. \tag{0.0.6}$$

Once the nearest-neighbor image has been found, then, its class is assigned to the incoming image:

$$L^*(I) = L^*(I_k) = C_k. \tag{0.0.7}$$

The success of the unsupervised classification depends on the implicit relationship existing between the k-attribute, the p-norm that is used to define the distance criterion for each classifier, and the class information. In the case of supervised classification the learning method $L^*(I_j; \mathbf{m})$ depends also on a set of parameters $\mathbf{m}$ that has to be tuned using class information from a subset of images of the database (testing database). The aim of this paper is to analyze the performance of such type of algorithms in the context of ensemble learning methodologies, by exploiting the structure of its uncertainty space.

One of the main challenges in dealing with very large databases of images is establishing a low-dimensional feature representation having enough discriminatory power. Different techniques based on image attributes have been proposed in the literature [17]. Spectral attributes, obtained by projecting the images onto the corresponding spectral basis, are the most common in face recognition. In consequence, the first part of our research is dedicated to this issue and it is the basis of the whole work.

The face image classification is the next key issue for face recognition problem and ensemble learning techniques are widely used in this field. Ensemble learning methodologies try to exploit the intrinsic uncertainty of any unsupervised/supervised classification problem [27]. The main idea in ensemble learning is to be able to weigh and combine several single classifiers in order to obtain a stronger classifier that outperforms every one of them [28].

In the face recognition context, ensemble learning consists in building a set of diverse classifiers based on different single image attributes. Most of the ensemble learning techniques used for face recognition belong to the category of supervised learning using boosting or bagging approaches based on individual classifiers [3, 5, 16, 23, 25, 31].

2. Objectives and Methodology

The main objective of this research is to design a robust methodology for automatic image classification that can be applied to different fields, and particulary in biometry authentication and also in the biomedical field, by constructing biomedical robots able to asses physicians in unravelling different kind of illness (diagnosis, prognosis and treatment). Particularly, in the medical field, we address the analysis of Triple Negative Breast Cancer (TNBC) samples to automatically establish their histological grade. One of the main challenges in achieving such methodology is the high dimensionality of the digital information involved, and how to extract the relevant features to achieve this predictions, accounting at the same time for their intrinsic uncertainty.

Image classification problems can be regarded as inverse problems of the kind $\mathbf{F}(\mathbf{m};\mathbf{a}) = \mathbf{d}$, where $\mathbf{d}$ is the set of observed data (image classes in this case), $\mathbf{a}$ is the image attribute vector, $\mathbf{m}$ are other numerical parameters involved in the classification, and $\mathbf{F}$ is the classifier. Besides, image classification problems have the additional difficulty that the correct classifier $\mathbf{F}$ is a priori unknown. This fact makes the identification problem involved, different from standard inverse problems, since $\mathbf{F}$, and eventually $\mathbf{m}$ have to be found. Learning algorithms have to be used instead. Our approach consisted in building $\mathbf{F}$ from a family of simple nearest-neighbor classifiers, and selecting the attribute vector $\mathbf{a}$ according to a prior analysis of the main types of attributes: spectral, statistical and regional descriptors. Finally, parameter vector $\mathbf{m}$ has to be identified in a supervised classification setup.

The methodology consists of the following steps:

1. Image features (attributes) extraction, that is, a set of procedures used to extract meaningful information from images, having enough discriminative power, and, at the same time reducing the data dimension by discarding the non-relevant information that might input noise into the classification. For that purpose we performed and exhaustive numerical analysis of various image attributes: covariance-based and covariance-free spectral methods, statistical attributes concerning the color (or gray levels) distribution and its spatial arrangement, and finally, attributes deduced from regional descriptors and image segmentation.

2. Construction of different unsupervised nearest-neighbor classifiers that are based on single attributes. Particularly, we have compared their predictive accuracy in a meaningful statistical way, by averaging the results obtained from different independent simulations of the image database. The term unsupervised refers to the fact that no class information is used to improve the accuracy of the different classifiers. As a result of this analysis, we have shown that although the accuracies obviously depend on the database that was used, no individual method can provide systematically 100% accuracy by itself. Therefore other kind of attributes and supervised learning algorithms are needed.

3. The use of the ensemble learning methodologies to improve the predictive accuracy by combining different independent classifiers. Ensemble learning methodologies try to exploit the intrinsic uncertainty of any unsupervised/supervised

classification problem. Ensemble classifiers aggregate several weighted weak classifiers in some specific way in order to obtain a strong classifier. In this context, a weak classifier is defined to be slightly more accurate than a random guess, meanwhile, a strong classifier is well-correlated with the true classification. Schapire (1990) used this fact to prove that ensemble voting using weak classifiers produces a strong classifier that outperforms every one of them. The diversity of the ensemble of classifiers that are used is a crucial feature in ensemble learning [21, 27]. The *diversity* concept requires in this context that the individual classifiers misclassify different instances. We performed a detailed analysis of the diversity of the individual classifiers that were analyzed, and numerically showed that the ensemble accuracy can drastically improve by reducing the ensemble to the most diverse attributes.

4. The ensemble learning methodology was applied to the face recognition problem and to the TNBC medical images classification, both in an unsupervised and supervised learning setups. Different image databases were used to perform the numerical analysis of the methodologies described before. In the context of face recognition problem, three diverse databases were selected: ORL, FERET and PUT face databases. For the medical images classification, a database of TNBC samples was provided by the Anatomical Pathology Service from the Hospital Universitario Central de Asturias (Oviedo, Spain).

Summarizing, the main objectives of this research are:

1. Analyze and numerical compare the main methods for dimensionality reduction and feature extraction.

2. Design of a robust unsupervised ensemble classifier by combining different kinds of image attributes through ensemble learning methodologies.

3. Extend the previous methodology to the supervised case. The final classifier will take into account the uncertainty space of the ensemble classifier.

4. The application of this methodology to the face recognition problem and its generalization in biomedicine. Future research will be devoted to analyze other biometric techniques (iris recognition, fingerprint identification, etc), and also,

other important biomedical applications, such as the design of biomedical robots able to asses medical doctors in their decisions.

3. Manuscript Organization

The structure of this thesis is consistent the publications that are included in the Appendix and that constitute the main steps of our research:

- Chapter **I** describes the numerical analysis and comparison of the main spectral decomposition methods, and their applications to the face recognition problem.
- Chapter **II** is devoted to the design of unsupervised ensemble classification methods for biometric applications, and particularly applied to the face recognition problem.
- Chapter **III** presents design of a robust supervised classifier based on exploring and exploiting the uncertainty space of the ensemble learning classifiers.
- Chapter **IV** comprises the first application of the methodology designed in this PhD thesis to the biomedical field, particularly, to the automatic classification of histological grade of different subgroups of triple negative breast cancer (TNBC), in order to optimize their diagnosis/treatment and prognosis.
- Finally, chapters **V** and **VI** present the conclusions that were achieved in this PhD thesis, and outline the main lines of the future research work (in Spanish and English, respectively).
- The appendix section is structured as follows:
  - Appendix A includes the three main publications in peer to peer journals of this PhD thesis, corresponding to the chapters **I** to **III**:
    1. Numerical Analysis and Comparison of Spectral Decomposition Methods in Biometric Applications. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 28, No. 1, pp. 14560. DOI: 10.1142/S0218001414560011. Published on 13 January 2014.
    2. Unsupervised Ensemble Classification for Biometric Applications. International Journal of Pattern Recognition and Artificial Intelligence,

Vol. 28, No. 4 (2014) 1456007 (32 pages). DOI: 10.1142/S0218001414560072. Published on 11 June 2014.

3. Exploiting the Uncertainty Space of Ensemble Classifiers in Face Recognition. Submitted to the International Journal of Pattern Recognition and Artificial Intelligence. Under Review.

- Appendix B includes two contributions to International Confrences:

  1. Automatic Classification of Cell Patterns for Triple Negative Breast Cancer Identification.
  2. Aligned PSO for Optimization of Image Processing Methods Applied to the Face Recognition Problem. The 5^{th} Joint International Conference on Swarm, Evolutionary, and Memetic Computing - SEMCCO 2013, Chennai, India, December 19-21, 2013.

  These contributions are the base of future publications that will be presented in peer-reviewed journals.

- Appendix C contains a book chapter entitled "Image Processing Methods for Biometric Applications", published in the Studies in Computational Intelligence series by Springer.

Chapter 1

Numerical Analysis of Spectral Decomposition Methods in Face Recognition

Biometry is a vast field of research including different applications in authentication, artificial intelligence, pattern recognition and computer vision. One of the most important problems, and also widely studied, is the face recognition problem, that was used as benchmark to design the original methodology that is presented in this PhD thesis. Although the face recognition problem (and most of the biometric techniques) has the particularity that the structure of the image samples (the spatial disposal of the components) follow the same pattern, the methodology presented in this PhD has been thought to be applicable to other important biometric applications and image recognition problems.

Once said that, this first step of our research presents a numerical analysis and comparison between various spectral image attributes, applied to the face recognition problem. One of the main challenges in this field consists in establishing a low-dimensional feature representation of the images having enough discriminatory power to perform high accuracy classification. Spectral decomposition methods are very suitable for this purpose. The aim of this analysis was to understand, and numerically compare, the main spectral methods proposed in the literature, since some authors praise the use

of some techniques, due to their presumed higher predictive accuracies. This was initially also our thought, since following the literature it seemed that some methods performed better than others, and the art in face recognition consisted in selecting the most performing one to build the final classifier. Nevertheless, we numerically proved (and theoretically explained) that all these methods provide similar results when the energy threshold used to achieve the model reduction is properly set. As a result of this research we concluded that most of the numerical results presented in the literature were biased due to the manner in which the numerical experiments were performed. To solve out this problem, we made up a methodology to perform systematically the same number of random independent simulations (100) with the same learning and testing databases, for all the different attributes. Finally, and the most important, none of these methods was able to provide a perfect classification, leading us to the use of other different image attributes. This subject was explored in the following part of our research, naturally leading us to the use of ensemble learning methodologies.

The analysis of this research concerns: 1) model reduction methods; 2) the design of the different nearest-neighbour classifiers; 3) conclusions and a brief presentation of the results.

1.1. Model Reduction Techniques

Spectral methods applied to a set of images consists in performing the orthogonal projection of each image onto different spectral basis. The spectral methods presented in this research can be classified as follows, according to the way of obtaining the spectral basis:

1. Spectral methods that use the whole database of images

The following methods belong to this category: Principal Component Analysis (PCA), Non-Centered PCA (NCPA), 2DPCA, Fisher's Linear Discriminant Analysis, and Independent Component Analysis (ICA). All these methods are covariance-based and imply diagonalization. ICA is a special case since although it is based on the maximisation of the non-Gaussianity, in practice, always involves whitening as a pre-processing step.

The projection in the general case is as follows:

$$C = VDV^T \in M(n_{pixels}, n_{pixels}), \tag{1.1.1}$$

$$I = \mu + V^T \mathbf{a}, \tag{1.1.2}$$

$$\mathbf{a} = V(I - \mu), \tag{1.1.3}$$

where C is a real symmetric covariance matrix, μ is the image trend, V the matrix whose columns $\mathbf{v}_k$ are the spectral basis vectors, and $\mathbf{a}$ is the projection of image I onto spectral basis V. The model reduction consists in finding the number of spectral components q such as

$$\widetilde{I} = \mu + \sum_{k=1}^{q} a_k \mathbf{v}_k, \tag{1.1.4}$$

$$\|I - \widetilde{I}\|_2 \leq \theta_E, \tag{1.1.5}$$

with $I, \widetilde{I} \in \mathbb{R}^{n_{pixels}}$, θ_E is the prescribed energy threshold, and n_{pixels} is the total number of pixels of any image in the database.

The Non-Centered PCA was introduced for the first time in this research, for the face recognition problem. It is important to remark that the first vector $\mathbf{v_1}$ of the Non-Centered PCA base plays a similar role to μ. Therefore, the model reduction is performed for $\mathbf{v_k}$ terms with $k \geq 2$, since most of the energy would be spanned by $\mathbf{v_1}$.

In practice, the covariance matrix C cannot be (and should not be) diagonalized, since n_{pixels} is usually very large, and besides, the rank of C is at most the number of images N used to compute C ($N \ll n_{pixels} \times n_{pixels}$). The way to deal with the numerical problem is properly explained in the publication in Appendix A, section A.1.

2. Spectral methods that act on single images

The following methods belong to this category: Singular Value Decomposition (SVD), Discrete Wavelet Transform (DWT), Discrete Cosine/Sine Transform (DCT/DST), Discrete Chebyshev Transform (DCHT), Discrete Walsh-Hadamard Transform (DWHT), and Discrete Hilbert Transform (DHT). All these methods, but the SVD, are covariance-free.

The projection in this case is performed as follows:

$$I = USV^T \in M(m, n) \tag{1.1.6}$$

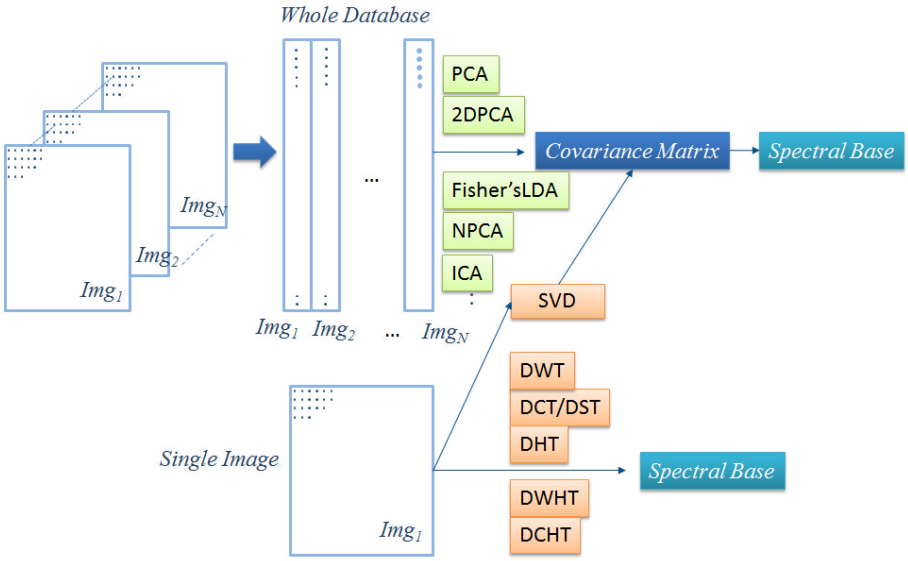


Figure 1.1: Model reduction.

where U, V are orthonormal basis of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, that depend on each of these methods.

The energy compression consists in finding the numbers of transformed pixels p and q such as:

$$\|I - S(1:p, 1:q)\|_F < \theta_E$$

where $S(1:p, 1:q)$ represents the $p \times q$ upper block of S where the energy has been concentrated, and F stands for the Frobenius norm. The dimensionality reduction is achieved from $m \times n$ pixels to $p \times q$ frequency components.

In the case of the SVD, only the q first singular values are needed, because S is blocky-diagonal. Nevertheless, SVD needs the diagonalization of the row and column correlation matrices, II^T and I^TI. For the other methods, the orthogonal matrices U, V are precomputed, and S has not block-diagonal form, that is, the SVD provides the optimum dimensionality reduction in the L_2 sense.

The Discrete Chebyshev Transform (DCHT) was introduced for the first time in this research, for the face recognition problem. Applying the DCHT for model reduction it was a challenging problem since Chebyshev polynomials are unstable for degrees higher than 50. This problem was solved by dividing the image into smaller blocks (less than 50×50) to perform this analysis.

Method	Covariance	Projection Basis
PCA/NPCA	$S = \sum_{k=1}^{N}(X_k - \mu)(X_k - \mu)^T$	Eigenfaces
2DPCA	$S_M = \frac{1}{N}\sum_{i=1}^{N}(I_i - \overline{I})^T(I_i - \overline{I})$	2DPCA basis
Fisher's LDA	$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T$ $S_W = \sum_{i=1}^{C}\sum_{X_k \in C_i}(X_k - \mu_i)(X_k - \mu_i)^T$	Fisherfaces
ICA	covariance-free maximization of non-Gaussianity	ICA basis
SVD	covariance-based	U, V basis
DCT/DST	covariance-free	Fourier basis
DCHT	covariance-free	Chebyshev Polynomials
DWHT	covariance-free	Walsh-Hadamard basis
DHT	covariance-free	Hilbert basis
DWT	covariance-free	Wavelet basis

Table 1.1: Spectral methods and their corresponding projection basis

Figure 1.1 shows how the different spectral methods perform the model reduction according to this classification. This classification is important since in all biometry authentication applications the speed of the classification algorithms is crucial. Therefore, it is crucial that all the image attributes are precomputed before applying the classification algorithm, in order for this to be fast. This is hard to achieve in the case of the covariance-based spectral methods, and also for those methods that act over the whole database, since the spectral basis should be updated depending on new subjects that are added to the database. Also, diagonalization is unfeasible for high dimensions and computationally very expensive (or even impossible).

The following table 1.1 summarises all the spectral methods used in this research, according to the above classification and their respective projection basis.

In conclusion, orthogonal transformations applied to images provide different important outcomes:

- Pixel decorrelation is achieved since orthogonal transformations induce rotations to the principal axes of the image in $\mathbb{R}^n$ and $\mathbb{R}^m$. In the case of PCA, this rotation is induced by the orthonormal basis calculated through the experimental covariance matrix.

- Dimensionality reduction is obtained with very little loss of energy since the transformed image concentrates most of its energy onto the first components (upper block of matrix S).

1.2. Nearest-neighbor Classifiers

The classification of a new incoming image I, corresponding to a new pose of the C individuals (classes) in the Bd database, is based on nearest-neighbour classifiers. In this context, the nearest-neighbour classifier is composed by of three items: 1) the model reduction method; 2) the type of analysis, global or local ; 3) the similarity criterion, that serve us to define a normed or Euclidean vector spaces of images. Items 1) and 2) provide the reduced-size attribute vector in each case.

Attribute Vector: each spectral method provides an individual image attribute, and for that purpose images are represented by their spectral features, that is, by their projection vectors $\mathbf{v}_I$ onto the different spectral basis, corresponding to the selected spectral method of analysis.

Type of Analysis: Two different kind of analysis are performed in this research: global and local analysis. In the case of the global analysis the attributes are calculated over the whole size of the image. For the local analysis, the images are divided into blocks, and for each block the attributes are computed. The final feature vector is constructed by merging into an unique vector all the local attribute vectors of the image, always computed on the same order. It is important to remark that, although the use of local features increases the dimension of the attribute space, its discriminative power is also increased in most of the cases. Moreover, the local analysis facilitates the application of covariance-based methods (PCA, NPCA, Fisher' LDA, 2DPCA) to very large images.

Similarity Criterion: a distance criterion is defined on the space of each spectral attribute, for finding the most similar images for the classification. Two different criteria were compared in this research for all the spectral methods:

1) Minimise the distance between the new incoming image and the database images in a certain norm p, defined over the space of attributes:

$$d(I, I_k) = \|\mathbf{v}_I - \mathbf{v}_k\|_p,\ I_k \in Bd.$$

Different norms $p = 2, 3, 5, \text{inf})$ were compared.

2) Maximise the cosine of the angle between the new incoming image and the database images, represented by their attributes vectors:

$$\cos(I, I_k) = \frac{\mathbf{v}_I \cdot \mathbf{v}_k}{\|\mathbf{v}_I\|_2 \|\mathbf{v}_k\|_2}.$$

The numerical experiments (shown in the next section) consisted in analyzing and comparing the predictive accuracy of the different classifiers for both types of analysis (global and local), and also for both criteria including different p-norms. The criterion that provided the best classification accuracy is shown in table 1.2. These classifiers will be used for ensemble learning in this PhD thesis.

The classification method used in this research belongs to the class of unsupervised learning, since the distance and the cosine are defined ad-hoc, that is, no information from the learning database classes is used in their definition to optimize the classification of the testing database.

1.3. Numerical Results and Conclusions

The numerical analysis were performed over two face images databases: the ORL database of faces provided by ATT Laboratories Cambridge and the Face Recognition Technology database (FERET). ORL provides 400 grey scale images corresponding to 40 subjects, with 10 poses per subject. Meanwhile, the FERET database contains 150 individuals, with 3 corresponding RGB color poses for each subject. In both cases we assume the images were already registered, thus, no alignment preprocessing such as translation, rotation and scaling based on the eyes coordinates has been adopted. Also, large pose variations and considerable lightening conditions changes were not analyzed, since these issues were not part of the aims of this research.

Table 1.2 summarises all the numerical results presented in this research, and it shows the best classification accuracy for each spectral method performed over 100 independent simulations of the ORL database.

Several important conclusions arise from this analysis:

1. All the spectral methods provide approximately the same accuracy when they are used with the same energy cut-off. The reason why this happens is that

Method	Type of Analysis	Criterion	Median	IQR
PCA	Local	p=2	95.00	2.00
NPCA	Local	p=2	94.50	2.50
2DPCA	Local	p=2	95.00	2.50
Fisher's LDA	Local	p=2	95.00	2.00
ICA	Global	cos	91.50	2.50
SVD	Local	p=3	93.00	2.00
DCT	Global	p=3	94.75	2.50
DST	Local	p=2	94.50	2.00
DCHT	Global	p=2	95.00	2.00
DWHT	Global	p=3	95.00	1.50
DHT	Global	p=2	94.00	2.25
DWT	Global	p=2	95.50	2.00

Table 1.2: Spectral methods and their corresponding projection basis

orthogonal transformations are isometries, and all the transformed images (matrices) are related by similarity relationships of the kind $I = UDV^T$, where U and V are orthogonal matrices. Thus, all the similar matrices (I and D) contain the same information if no model reduction is performed. In all the numerical experiments we used an energy threshold of 99.5%, obtaining a very good dimensionality reduction, and a median accuracy of 94.75% considering all the spectral methods. In our opinion the different accuracies shown in the literature for all these methods are due to the use of different energy cut-offs for dimensionality reduction, and also to the different parameters used in the face recognition experiments (databases), that is, no unique computing platform was previously used to perform the numerical comparison of all these different methods. This is not the case of the present research where all the spectral methods have been compared using the same numerical conditions. By tuning this parameter correctly all these methods should perform similarly.

2. The previous conclusion reinforces the fact that all the covariance-free spectral decomposition techniques based on single images (DCT, DST, DCHT, DWT, DWHT, DHT) are very interesting since they provide high accuracies and are not computationally expensive compared to covariance-based techniques.

3. The use of local spectral features generally provide higher accuracies than global features for the spectral methods that use the whole training database (PCA, NPCA, 2DPCA, Fisher LDA, ICA). For the methods based on orthogonal transformations of single images, global features perform better.

4. The distance criterion generally provides higher accuracies than the cosine. The use of other p norms ($p > 2$) provides similar results to the Euclidean norm, nevertheless some methods perform better. Norm $p = 1$ always provides the worst results. This fact might be related to how model reduction and energy compression are performed in the case of spectral methods.

5. No spectral method can provide systematically 100% accuracy. Therefore other kind of attributes and supervised learning algorithms are needed.

The complete description of this research can be found in Appendix A, section A.1.

Chapter 2

Face Recognition via Unsupervised Ensemble Classification Methods

One of the main conclusions of the previous chapter is that no spectral method of analysis was able provide systematically 100% accuracy. Therefore, in this research we explore other kinds of image attributes which could furnish complementary information. We also construct other type of classifiers based on these attributes, which are combined using ensemble learning techniques. Consequently, the research presented in this chapter shows three different ensemble learning algorithms for automatic image recognition in biometry (face recognition problem). Diversity among the different individual classifiers that are used is also an important feature in ensemble learning, and we show that the classification accuracy is improved by reducing the set of combined classifiers to the most diverse ones. As a result of this analysis we proposed three different versions of ensemble classifiers: *CAV1*, *CAV2*, and *CBAG*. The main difference among them is the way that image candidates are selected to perform the consensus by majority voting.

The main achievement shown in this chapter was the design of a set of simple and diverse nearest-neighbour classifiers that were combined through ensemble learning approaches that provide systematically (in a "stable" way) very high accuracies for the face recognition problem. This methodology can be adapted to solve other biometric and pattern recognition problems, with applications in different fields, such as

biomedicine, security, semantic web, etc. Finally, the ensemble learning methodology presented here, can be dramatically improved by optimizing the parameters of these classifiers in a supervised approach, and by exploring its corresponding uncertainty space, that will be detailed in the next chapter.

The analysis of this research concerns: 1) the presentation and numerical analysis of the different image attributes which define the different individual nearest-neighbour classifiers; 2) the ensemble learning algorithms and their diversity analysis; 3) conclusions and a brief presentation of the results.

2.1. Image Attributes

In order to complement the information provided by the spectral methods, we extended the set of image attributes. Therefore, we have analyzed the performance of three different types of image attributes that were used for image representation:

Statistical attributes: color histograms, percentiles arrays and omnidirectional variogram accounting for the spatial color distribution.

Segmentation features and regional descriptors: different measures of texture analysis and edge detection.

Spectral attributes: PCA, 2DPCA, Fisher's LDA, DWT and DWHT.

As in the previous chapter, each image attribute together with its corresponding type of analysis (local or global) and the distance criterion (p-norm or cosine), define an individual classifier. Besides, the classification was unsupervised since no class information was used to improve the design of the different classifiers. Nevertheless, a few numerical experiments are shown to introduce the need of supervision.

We have numerically proved that the most performing classifier is based in the local histogram, followed by the local percentiles, and the global analysis via the discrete wavelet transform. We already know from the previous chapter that all the spectral methods provide similar accuracies, but in this research we will also prove, using the diversity analysis, that these classifiers are not diverse and provide redundant information in the classification. All these spectral nearest-neighbour classifiers will be then

reduced to the DWT, which is the method that provides the highest accuracy with the lowest computational cost. The regional descriptors are the attributes that provide the lowest accuracies. These individual nearest-neighbour classifiers use the L1 norm in the case of all the statistical attributes, texture analysis and regional descriptors, while the L2 norm provides higher accuracies for all the spectral attributes. In most of the cases, the local analysis provides better results, only the edges and the DWT perform better globally.

2.2. Ensemble Learning and Diversity Analysis

For the purpose of improving the predictions of the classifiers described in the previous section, we have chosen the ensemble learning methodology. An ensemble of classifiers is a set of classifiers whose individual predictions are combined in some way (typically by weighted or unweighed voting) with the goal to obtain a classifier that outperforms every one of them. This methodology assumes that assembling together a set of weak and diverse classifiers produces a strong classifier with higher accuracy. The word diverse refers to the fact that the different classifiers misclassify different images, while the word weak refers to the fact that the accuracy is better than random guessing.

There are different ways of combining classifiers, depending on the stage at which they operate: at feature level or at score level. We have chosen the approach consisting in weighting each classifier's output score. The final decision of the ensemble classifier was taken by majority voting.

Diversity among the different classifiers that are used is also an important feature in ensemble learning. The diversity analysis performed over the presented attributes confirmed the conclusion that reducing the number of classifiers to the most diverse, improved the results (accuracies and stability) of the three ensemble learning algorithms described before. As a conclusion of this analysis, the five most diverse attributes were chosen: local histogram, variogram, discrete wavelet transform, texture and edges detection. Table 2.1 shows the median accuracies and the IQR of the five individual classifiers constructed from these uncorrelated attributes.

Method	Type of Analysis	Criterion	Median	IQR
Histogram	Local	L_1	98.80	1.50
Variogram	Local	L_1	93.00	2.25
Texture	Local	L_1	92.50	3.00
Edges	Local	L_2	90.85	2.50
DWT	Global	L_2	95.50	2.00

Table 2.1: The individual classifiers based on the less correlated attributes

2.3. Unsupervised Ensemble Learning Algorithms

Three different unsupervised ensemble learning algorithms were designed and compared in this research:

CAV1 The first algorithm called Consensus Algorithm Version 1 (*CAV1*) retains the N_c nearest neighbour images for every individual classifier (based on a single attribute), and the prediction is performed by majority voting (Borda count) taking into account the individual prior accuracy of each classifier giving more reliability to the classifiers that provided better prior accuracies.

CAV2 In the second version, the N_f best candidates are selected by the most performing prior classifier (the local histogram in this case), and the rest of classifiers will act on this reduced-size database of preselected images.

CBAG Finally, the third version is based on ensemble prediction using different classifiers and bagging. In this case the bags are obtained by partitioning the learning database into different learning data subsets that are randomly drawn. For each of these subsets the classifiers are randomly generated combining several attributes from the above mentioned categories.

In all the three cases the individual classifiers are finally combined by taking into account the majority voting score of their decisions.

Figure 2.1 shows a global sketch for the work-flow of these algorithms.

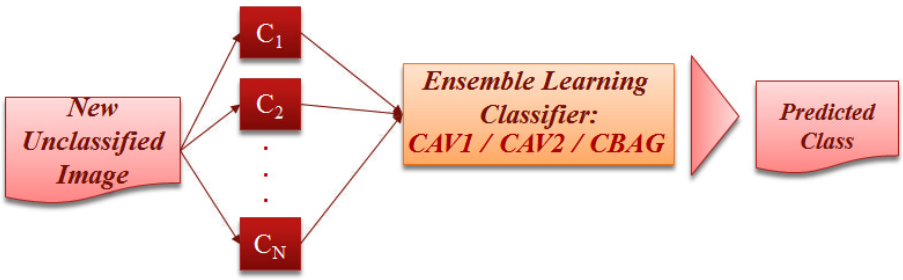


Figure 2.1: Work-flow of the ensemble learning algorithms *CAV1*, *CAV2* and *CBAG*.

2.4. Numerical Results and Conclusions

The numerical results were performed over the ORL database of faces provided by ATT Laboratories at Cambridge, that we have already used in the first chapter.

The main results shown in this research are the following:

1. The statistical attributes (local histogram and percentiles) are the individual classifiers that provided the higher accuracies, followed by the spectral methods (DWT), and the regional features (texture analysis).

2. No single attribute is able to provide systematically 100% accuracy over the ORL database.

3. The accuracy and stability of the classification is increased by consensus classification (ensemble learning techniques).

4. The diversity analysis on the classifiers allowed us to select the five most diverse attributes to be used in the learning algorithms explained before. These attributes are: the local histogram, the local variogram, the local texture, the local edges and the global DWT. The use of this set of five uncorrelated classifiers, improved the accuracies and stability in the prediction obtained by the ensemble learning algorithms: CAV1 and CAV2 performed very close to the local histogram and the stability of the classification increased even more.

5. Optimum results are obtained by optimizing the parameters of these classifiers using a member of the Particle Swarm Optimization (PSO) family. These results are in accord with the conclusions that are presented in the literature using

ensemble learning methodologies, that is, it is possible to build strong classifiers by assembling different weak (or simple) classifiers based on different and diverse image attributes.

These encouraging results brought us to the use of supervised ensemble techniques in face recognition and in other important biometric problems that will be presented in the next chapter.

The complete description of this research can be found in Appendix A, section A.2.

Chapter 3

Uncertainty Analysis of Ensemble Classifiers Applied to Face Recognition

As we have concluded in the previous chapter, a supervised approach was needed to improve the results of the designed ensemble learning classifiers. In this chapter we present a supervised ensemble learning algorithm, named Supervised Consensus Algorithm Version 1 (*SCAV1*), which is based on *CAV1*, the ensemble classifier defined in the previous chapter. This methodology is expanded to the supervised case, by optimizing *CAV1*'s parameters. Classification problems regarded as optimization problems generate an uncertainty space which is exploited in this case for the final decision. The final classifier applies Borda count over the best models found in the uncertainty space of *CAV1*. In this chapter, we also show the application of this methodology to two different image databases (ORL and PUT), obtaining very high and stable accuracies (100% median accuracy and almost a null interquartile range).

3.1. Image Attributes

In order to design and perform the supervised ensemble classification methodology, images were also represented through attributes vectors. We extended the set of image

attributes used in the previous chapter by introducing Zernike moments in order to increase the discriminative power of the ensemble learning classifier. Thus, the final list of image attributes is composed by statistical attributes (histogram and variogram), regional descriptors (texture) and edges detection, spectral attributes (DWT), and moments (Zernike moments).

3.1.1. Zernike Moments

Zernike moments was introduced in this research as a novel image attribute/individual nearest-neighbour classifier for the ensemble learning methodology shown in Appendix A, section A.3. Zernike moments were defined by Zernike in 1934 [32]. The kernel of Zernike moments is a set of orthogonal complex polynomials $V_{nm}(x, y)$ defined over the polar coordinate space inside a unity radius circle ($x^2 + y^2 \leq 1$), as follows:

$$V_{nm}(x, y) = V_{nm}(\rho, \theta) = R_{nm}(\rho)e^{im\theta}, \tag{3.1.1}$$

with $n \in \mathbb{N}, m \in \mathbb{Z}$, and $n - |m|$ even, $|m| > n$, where ρ and θ are the polar coordinates of the the pixel (x, y), and $R_{nm}(\rho)$ is the radial polynomial, defined as:

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!}{s!(\frac{n+|m|}{2} - s)!(\frac{n-|m|}{2} - s)!} \rho^{n-2s}. \tag{3.1.2}$$

Zernike moments of a scalar field $f(x, y)$ are obtained by projecting the image onto the orthogonal basis of polynomials. The Zernike moment of order n with repetition m for a continuous image $f(x, y)$, with values in the interior of the unit circle (vanishing outside the circle) is:

$$Z_{nm} = \frac{n+1}{\pi} \int \int_{x^2+y^2 \leq 1} f(x, y) V^*_{nm}(x, y) dx dy. \tag{3.1.3}$$

In the case of a digital image I of size $M \times N$ pixels, the image is first resized to a square image, the center of the image is considered as the origin, and pixels are mapped to the range of the unit circle. The pixels located outside the circle are ignored. As well, the case of the circumscribed circle could also be considered, as showed in figure 3.2.

Assuming that $N = \min(M, N)$, the discrete Zernike moment of a $N \times N$ image is:

$$Z_{nm} = \frac{n+1}{\pi} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} I(i, j) V^*_{nm}(x_i, y_j) \triangle x_i \triangle y_j, \tag{3.1.4}$$

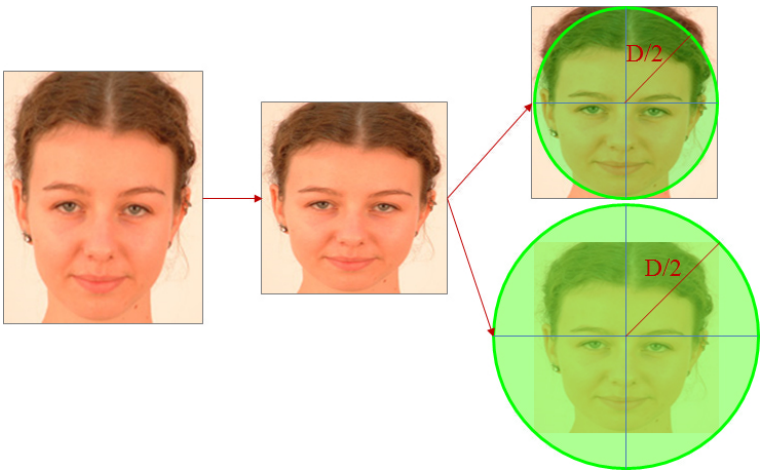


Figure 3.1: Flowchart of Zernike moments calculation for a digital image.

where $x_i = \frac{2i+1-N}{D}$, $y_i = \frac{2j+1-N}{D}$, $\triangle x_i = \triangle y_j = \frac{2}{D}$, with $D = N$ if the image is the image is circumscribed to the circle of analysis, and $D = \sqrt{2}N$ when the image is inscribed (3.2).

Zernike moments are invariant to rotation and scale and very robust in the presence of noise [30]. These properties make them an appropriate feature for automatic face recognition. The Zernike invariants are the magnitudes of the real and imaginary components of the resulting moments. Because of their orthogonality, it is expected that a small set of moments can be used to estimate parameters associated with different models. Moreover, Zernike moments provide shape information underlying the intensity surface of an image, captured at different orders. For example, the zeroth-order moment represents the mean intensity value in an image neighbourhood and first-order moments are related to the centre-of-gravity of the intensity surface, whereas the second-order moment captures the variance of the intensity levels present in the local neighbourhood. Thus, a discontinuity in local intensities results in a high first-order moment, a discontinuity in local gradients results in a high second-order moment, and so on.

Figure 3.2 shows the first five Zernike polynomials of the basis and its respective invariants.

In this chapter the face images were resized to 92×92 pixels. The Zernike nearest-neighbour classifier has been constructed with a basis of polynomials of order 11. This

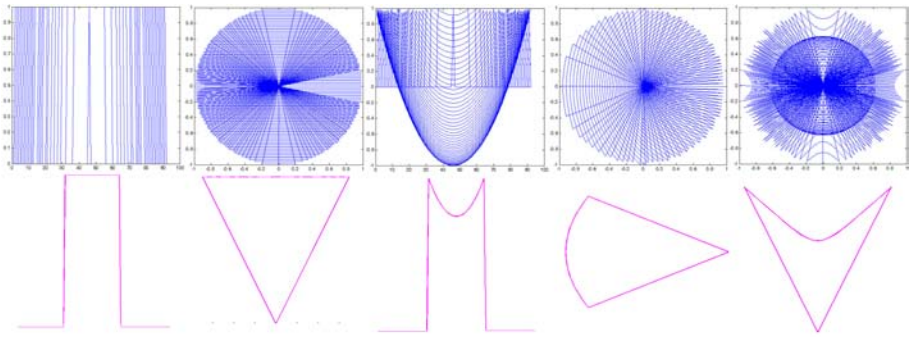


Figure 3.2: The first five Zernike polynomials and their respective invariants.

Correlation	Q-statistic	Disagreement	Double-fault	Entropy	KW	Agreement
[1] 0.28	0.67	0.17	0.04	0.23	0.07	0.17
[2] 0.23	0.56	0.20	0.04	0.28	0.08	0.14

Table 3.1: Diversity of the ensemble of the classifiers.

generated a basis of 42 polynomials, taking into account the necessary condition that the indices (the order n and the repetition m) have to satisfy. The numerical results proved that the local analysis performs better, thus the final Zernike classifier was constructed upon the 32 blocks, obtaining an attribute vector of dimension 32×42. The best classification accuracies were obtained in the case of local analysis with the L2 norm (95.00% of median accuracy).

The diversity analysis performed over the final six classifiers proved that he introduction of the Zernike moments slightly decreased the quality of the diversity statistics, since Zernike moments can be considered as an additional texture descriptor. Moreover, if Zernike moments were considered within the texture descriptors, the diversity statistics will be remain similar to the diversity measures that were originally obtained using the five nearest-neighbour classifiers. Results of the diversity analysis for the ensemble learning methodology are summarized in table 3.1.

[1] For the 6 considered classifiers.
[2] For the selected classifiers, without Zernike moments.

3.2. Supervised Ensemble Learning Methodology as Optimization Problem

In the context of face recognition, supervised learning consists in finding the functional (classifier) L^* that approximates the empirical data. L^* is defined on the input set of face images $\mathcal{M}_{(n,m,c)}$ and takes values in the output set of classes C, as follows:

$$L^* : \mathcal{M}_{(n,m,c)} \rightarrow C, L^*(I_j; \mathbf{m}) = c_j, \tag{3.2.1}$$

where $I_j \in \mathcal{M}_{(n,m,c)}$is an input face image, c_j is its corresponding predicted class, and $\mathbf{m}$ is the set of parameters on which the classifier L^* depends.

The goal of supervised learning is to adjust the parameters $\mathbf{m}$ using class information from the training data set, by solving the optimization problem of minimizing a given cost or error function as follows:

- defining a training set $\{T, \mathbf{c}_{true}\}$, where $T = \{I_1, I_2, ..., I_s\} \subset \mathcal{M}_{(n,m,c)}$ contains s testing images that have been selected from the face images database and $\mathbf{c}_{true} = (c_1, c_2, ..., c_s)$ are the true classes of the images of T.

- minimization of the learning empirical error cost function:

$$\phi(\mathbf{m}) = \|\mathbf{L}^*(T; \mathbf{m}) - \mathbf{c}_{true}\|_q, \tag{3.2.2}$$

  where $\mathbf{L}^*(T; \mathbf{m})$ represents the corresponding class prediction of the images from the training set, for a certain value of the learning parameters $\mathbf{m}$.

- sampling the set of equivalent parameters $\mathbf{m}$ of this classifier:

$$M_{tol} = \{\mathbf{m} : \|\mathbf{L}^*(T; \mathbf{m}) - \mathbf{c}_{true}\|_q \leq tol\}, \tag{3.2.3}$$

Region M_{tol} is related to the topography of the cost function $\phi(\mathbf{m})$ in the region of low misfits, and the intrinsic uncertainty of any inverse/optimization problem [7]. In our case, the methodology consists in analysing the performance of supervised learning in the context of ensemble learning, by reformulating the algorithm *CAV1*, introduced in the previous chapter, in a supervised approach, and exploring its intrinsic uncertainty space to take decisions.

3.2.1. Reformulating Supervised Learning as Ill-Posed Inverse Problems

Supervised learning can be reformulated in the framework of inverse problems [29, 22]. An inverse problem consists in finding the model or models $\mathbf{m}$ whose predictions $\mathbf{L}(\mathbf{m})$ accurately match the observed data $\mathbf{c}_{true}$:

$$\mathbf{L}(T, \mathbf{m}) \simeq \mathbf{c}_{true}, \tag{3.2.4}$$

where $\mathbf{m} = (m_1, ..., m_n) \in M \subset \mathbb{R}^n$ is the estimated model that belongs to a set of admissible models M, defined in terms of some prior knowledge, $\mathbf{c}_{true} \in \mathbb{R}^s$ is the observed data (the true classes to which really belongs the face images in the observation set T, as defined before), and finally, $\mathbf{L}(T, \mathbf{m}) = (L_1(\mathbf{m}), ..., L_s(\mathbf{m}))$ is the forward model (or final classifier in our case).

In practice, inverse problems are ill-posed and their uncertainty analysis of the solution can be used to perform prediction with the corresponding error-bars [11, 12]. In consequence, uncertainty estimation in discrete inverse problems involves sampling the family M_{tol} of equivalent models $\mathbf{m}$ (solutions) that are consistent with our prior knowledge and fit the observed data $\mathbf{c}_{true} \in \mathbb{R}^s$ within the same misfit tolerance tol:

$$\|\mathbf{L}(T, \mathbf{m}) - \mathbf{c}_{true}\| \leq tol. \tag{3.2.5}$$

3.3. Exploring the Uncertainty Space

Exploring the uncertainty space of any learning/inverse problem is usually done by obtaining a representative sample of the set of equivalent models. Sampling-based methods are a natural tool to appraise problems that have a very large number of plausible scenarios. Monte Carlo methods are usually used to accomplish this task, and comprise a broad set of algorithms whose central idea is to learn about a system by simulating it using random sampling strategies [20, 26]. Like other random sampling methods, Monte Carlo methodologies generate sequences of models that converge with time to the posterior distribution of the model parameters in the inverse problems. The main drawback of random sampling methodologies is that they depend too strongly on the dimension of the parameter space and require innumerable forward simulations.

Global optimization methods can provide under exploratory conditions a very good approximation of the model posterior distribution [1, 10]. This strategy is called sampling-while optimizing and is much faster than random sampling. Conversely, the posterior sampling is approximate. Nevertheless, the aim of this sampling procedure is not finding the true posterior distribution of the model parameters for the supervised classifier, but improving the accuracy and the stability of the classification, as it is shown.

Understanding the topography of the cost function in the low misfit regions provide a great benefit in establishing more efficient methods for estimating uncertainty.

The analysis carried out by Fernández-Martínez at al. (2012,2013) [11, 12] showed that the topography of the cost function $\phi(\mathbf{m})$ in the region of low misfits (or high prediction accuracies), correspond to one or several disconnected flat (with null gradients) elongated curvilinear valleys. For linear inverse problems, the region of equivalence is a connected set with a hyperquadric of equivalence, whose axes are proportional to the inverse of the singular values of the forward operator and whose directions are oriented along the singular vectors. Straight infinite valleys are related to the non-injective character of the linear forward operator $\mathbf{L}$, and to the existence of a nontrivial null space. Nonlinear topographies are analogous to the linear equivalence regions, but nonlinearities of the forward operator can cause the error landscape to be deformed or constituted of different disconnected islands of low misfits.

In this chapter, to perform the final decision, we sample the models found by exploring the uncertainty space generated by the optimization problem applied to the supervised approach of *CAV1* ensemble classifier. Uncertainty estimation of the classifier $L^*(I;\mathbf{m})$ consists in sampling the equivalent sets of parameters M_{tol} of this classifier:

$$M_{tol} = \{\mathbf{m} : \|\mathbf{L}^*(T;\mathbf{m}) - \mathbf{c}_{true}\|_q \leq tol.\}. \tag{3.3.1}$$

Figure 3.3 shows the flowchart of the methodology that we propose that is based on the exploration of the uncertainty space of *SCAV1*.

The sampling is performed via an explorative member of the PSO family (RR-PSO), and we obtain a set M_{tol} of parameters that induce a set L_{tol} of equivalent classifiers in the sense that

$$\|\mathbf{L}^*(T;\mathbf{m}_k) - \mathbf{c}_{true}\|_q \leq tol : k = 1, ..., N_s. \tag{3.3.2}$$

Figure 3.4 shows the topography of the cost functional in the PCA space, obtained

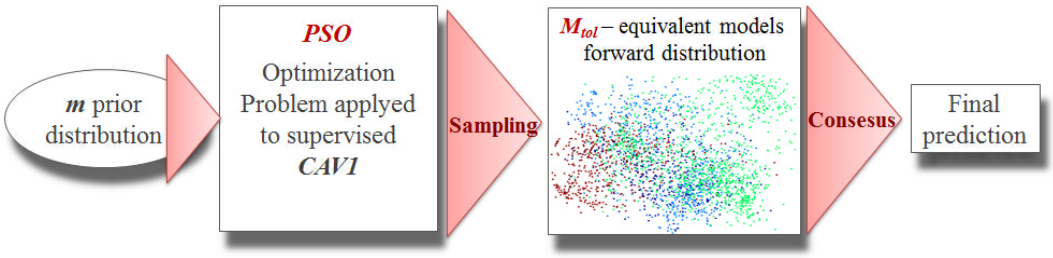


Figure 3.3: The uncertainty space generated by the optimization problem.

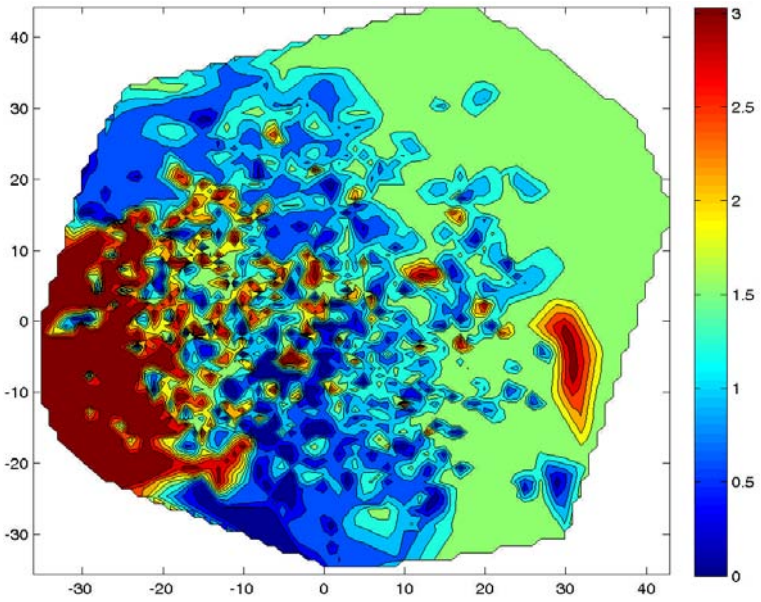


Figure 3.4: The topography of the uncertainty space generated by PSO applied to the supervised approach of *CAV1*.

from a set of equivalent models found by performing PSO optimization to the supervised approach of *CAV1*. In this case, the equivalent models have dimension 11. The first 2 principal components were considered for this graphic representation. It can be observed the complex topography of the cost function in the region of low misfits. The final prediction was accomplished by consensus over the classifiers defined by these models. The reason is as follows: in the presence of data noise, the model with the highest accuracy in the learning database will never provide the right classification (the highest generalization accuracy). The noise in data has two different sources in a classification problem: 1. A wrong class assignment 2. A partially wrong learning algorithm $\mathbf{L}^*$, whose effect in the prediction could be interpreted as an error in the class. These two error sources can be explained as follows: let us name $\mathbf{L}^*_{true}$ the true learning classifier, $\mathbf{m}_{true}$ be the true model parameters that provide the perfect classification $\mathbf{c}_{true}$, that is, $\mathbf{L}^*_{true}(\mathbf{m}_{true}) = \mathbf{c}_{true}$. The first source of noise consists in measuring $\mathbf{c}_{observed} = \mathbf{c}_{true} + \delta\mathbf{c}$. The second source consists in adopting $\mathbf{L}^* = \mathbf{L}^*_{true} + \delta\mathbf{L}^*$. In this case,

$$\mathbf{L}^*(\mathbf{m}_{true}) = (\mathbf{L}^*_{true} + \delta\mathbf{L}^*)(\mathbf{m}_{true}) = \mathbf{c}_{true} + \delta\mathbf{c}_{\delta\mathbf{L}}, \tag{3.3.3}$$

where $\mathbf{c}_{\delta\mathbf{L}}$ is the error in the predicted class generated by an error $\delta\mathbf{L}^*$ in the learning classifier. Obviously, finding $\mathbf{m}_{opt}$ (the optimum model) such as

$$\phi(\mathbf{m}_{opt}) = \min \|\mathbf{L}^*(\mathbf{m}) - \mathbf{c}_{observed}\|, \tag{3.3.4}$$

with $\mathbf{L}^* = \mathbf{L}^*_{true} + \delta\mathbf{L}$ and $\mathbf{c}_{observed} = \mathbf{c}_{true} + \delta\mathbf{c}$, will never provide the true model parameters $\mathbf{m}_{true}$ that gives the perfect classification. Nevertheless, if the noise is low or moderate, it is expected that the model $\mathbf{m}_{true}$ will be located in the region of equivalence M_{tol}, with a lower predictive accuracy than $\mathbf{m}_{opt}$. Fernández-Martínez et al.(2014) presented an interesting analysis on the effect of noise and regularization in linear and nonlinear inverse problems related to this subject [13, 14].

The supervised algorithm that we proposed consists in, giving a new incoming image I_{new}, applying the set of ensemble classifiers in L_{tol} to perform the consensus classification. It is important to remark that the set L_{tol} is formed by ensemble classifiers. This algorithm provides the image with the highest score and a set of other candidate images to be the solution of the face recognition problem. It is expected that when the score of the best candidate image is very high with respect to the other solution candidates, the uncertainty of the identification is very low. Conversely, if

several images have similar scores, the uncertainty of the identification is very high, as is the risk of misclassification.

3.3.1. RR-PSO for Ensemble Learning Parameters Optimization

Particle Swam Optimization (PSO) is a global optimization algorithm that was initially proposed based on the behaviour of different animals in nature, such as bird flocks and fish schools searching for food [19].

A swarm of particles (or models) explores the space of possible solutions in order to optimize a given cost function. The particle swarm algorithm applied to optimization problems is very simple: individuals, or particles, are represented by vectors whose length is the number of degrees of freedom of the optimization problem. To start, a population of particles is initialized with random positions $\mathbf{x}^0$ and velocities $\mathbf{v}^0$. The same objective function is used to compute the objective value of each particle. As time advances, the position and velocity of each particle is updated as a function of its objective function value and of the objective function values of its neighbors. At time-step $k+1$, the algorithm updates positions $\mathbf{x}^{k+1}$ and velocities $\mathbf{v}^{k+1}$ of the individuals as

$$\begin{aligned} v_i^{k+1} &= \omega v_i^k + \phi_1(g^k - x_i^k) + \phi_2(l_i^k - x_i^k), \\ x_i^{k+1} &= x_i^k + v_i^{k+1}, \end{aligned} \tag{3.3.5}$$

where l_i^k is the $i-$th particle's best position, $g^k$the global best position on the whole swarm, $\phi_1 = r_1 a_g$ and $\phi_2 = r_2 a_l$ are the random global and local accelerations, and ω is a real constant called inertia weight. Finally, r_1 and r_2 are random numbers uniformly distributed in $(0, 1)$ to weight the global and local acceleration constants a_g and a_l, also real numbers.

Fenández-Martínez and García Gonzalo (2008) proved that PSO could be physically interpreted as damped mass-spring system with unit mass, damping factor $1 - \omega$, and

total stiffness constant $\phi = \phi_1 + \phi_2$, i. e. the so-called PSO continuous model [9].

$$\begin{cases} x''(t) + (1-\omega)x'(t) + \phi x(t) = \phi_1 g(t) + \phi_2 l(t), \\ x(0) = x_0, \\ x'(0) = v_0, \end{cases} \tag{3.3.6}$$

where $t \in \mathbb{R}$. In this case $x(t)$ stands for the coordinate trajectory of any particle in the swarm. In this model particles interact through the local and global attractors, $l(t)$, and $g(t)$, and mean particle trajectories oscillate around the particle position.

$$o_i(t) = \frac{\phi_1 g(t) + \phi_2 l_i(t)}{\phi_1 + \phi_2}. \tag{3.3.7}$$

Based on this physical model, Fernández-Martínez and García Gonzalo (2009, 2012) deduced a whole family of PSO optimizers, including RR-PSO, by adopting different discretization schemes for the acceleration and velocity [6, 8].

The RR-PSO that we used in this research, was obtained by adopting a regressive discretization scheme in acceleration and velocity of the PSO continuous model [8], as follows:

$$\begin{aligned} v(t+\Delta t) &= \frac{v(t) + \phi_1 \Delta t \left(g(t) - x\left(t\right)\right) + \phi_2 \Delta t \left(l\left(t\right) - x\left(t\right)\right)}{1 + \left(1-\omega\right)\Delta t + \phi \Delta t^2} \\ x(t+\Delta t) &= x(t) + v(t+\Delta t)\Delta t, \ t, \ \Delta t \in \mathbb{R} \\ x\left(0\right) &= x_0, \\ v\left(0\right) &= v_0. \end{aligned} \tag{3.3.8}$$

Fernández Martínez and García Gonzalo (2012) numerically proved that RR-PSO has the greatest convergence rate for a wide range of benchmark functions [8].

The RR-PSO first and second order stochastic stability regions are unbounded, and it has been numerically shown that performing parameters sets (in terms of exploration and convergence) are concentrated around the line $\overline{\phi} = 3\left(\omega - \frac{3}{2}\right)$, mainly for inertia

Algorithm	median	std
SCAV1 / ORL	100	1.2
SCAV1+GS /PUT	99.50	0.52

Table 3.2: *SCAV1* algorithm applied to ORL and PUT face databases.

values $\omega > 2$. This line is independent of the cost function that is optimized, and remains invariant when the number of parameters of the optimization function increases. Furthermore, it is located in a zone of medium attenuation and very high frequency swarm of trajectories. This last property confers to this algorithm a good balance between exploration and exploitation, since this feature allows for a very efficient and explorative search around the oscillation center of each particle in the swarm. Due to this feature, RR-PSO has been chosen in this research to sample the set of parameters $\mathbf{m}$ for *CAV1*, providing the highest predictive accuracies for the face recognition problem on the testing database (M_{tol} set).

3.4. Numerical Results

The methodology described in this chapter was applied to two publicly available face databases ORL and PUT, with very high and stable accuracies of the classification, as shown in table 3.2.

Table 3.2 shows the best results of the supervised methodology proposed in this chapter, when applied to the ORL and PUT databases, using the two versions of the algorithm (with or without group substitution).

In conclusion, we proved that exploiting the uncertainty space of an ensemble classifier (*SCAV1* in this case) is the right strategy to adopt, and provides optimum results. Models in M_{tol} set concern the uncertainty space of the ensemble learning algorithm. These models do not need necessarily to be diverse. In fact, the diversity analysis performed on M_{tol} showed that the diversity measures are not very favorable. Nevertheless, the combination of all the models in M_{tol} confers more stability to the final classifiers. A possibility to increase diversity would be to infer M_{tol} using different bags

of the learning database. Further research will be devoted to improve this methodology, to test it on other databases and to translate it to other biometrical problems, for example with applications in the biomedical field.

The complete description of this research can be found in Appendix A, section A.3.

Chapter 4

Biomedical Application: Histological Classification of Triple Negative Breast Cancers

This research was presented to ***The*** 6^{th} ***International Conference on Advanced Cognitive Technologies and Applications - COGNITIVE 2014***, May 25 – May 30, 2014, Venice, Italy. In this conference we addressed the topic *COGNITION: Artificial intelligence and cognition* which was dedicated to: Expert systems, knowledge representation and reasoning; Reasoning techniques, constraint satisfaction and machine learning; Logic programming, fuzzy logic, neural networks, and uncertainty; State space search, ontologies and data mining; Games, planning and scheduling; Natural languages processing and advanced user interfaces; Cognitive, reactive and proactive systems; Ambient intelligence, perception and vision; Pattern recognition.

In this research we have explored the possibility to design a biomedical robot able to assess physicians to identify the kind of histological grade/survival of different subgroups of triple negative breast cancer (TNBC), in order to optimize their diagnosis/treatment and prognosis. The detailed description of TNBC is presented in the publication from Appendix B, section B.1.

This paper is an important contribution, since it comprises the first application of the methodology designed in this PhD thesis to the biomedical field. Particularly, to the automatic classification of the TNBC, which accounts for $15\% - 25\%$ of all breast

cancers and is associated with a poor outlook. This type of cancer is characterized by the absence of three key receptors (estrogen receptor, progesterone receptor and the human epidermal growth factor receptor 2). Standard treatment, such as hormone therapy, cannot be used for triple negative breast cancer.

Establishing the correct histological grading is crucial important for prognosis and treatment.

Unfortunately, at the present time, histological grading of TNBC is mainly based on visual inspection by medical experts of histological images, and up to our knowledge, no biomedical robot exists to automatize these approaches, to integrate both types of information, and to asses physicians on the diagnostic.

Therefore, we present a methodology for the optimization of the basal cells pattern classification. Different machine learning algorithms are performed on histological images, and on a list of pathological and immunohistochemical variables currently-used in medical practice.

The results obtained applying this methodology on a database of 105 patients showed that pathological and immunohistochemical variables, on one hand, and histological images on the other hand, provide complementary information which improves the classification of TNBC samples.

Consequently, the modelling has to different parts: the first one concerns the classification of samples by their histological grading (2 or 3) following a list of inmunohistochemical variables. This part has been done using the methodology designed in the PhD of Enrique J. de Andrés Galiana, about the design of biomedical robots and its application to genomics and clinics. The second part refers to the use of histological images to solve the same problem. The final aim of this methodology is to use both pieces of information to construct a robust classifier. This last part was not performed in this contribution. Further research will be performed about this subject in the near future.

4.1. Database Description

Hospital Universitario Central de Asturias (Spain) provided us a collection of TNBC samples, which proceeded from a cohort of 105 Caucasians women, between 30 and 94

years, that were enrolled in this study. The acquisition, treatment, and posterior study of the tumor samples were realized in total accordance with the last version of the Helsinki Declaration of 1975 [2].

The final database that we disposed of, consisted of histological images taken at two resolutions (X100 and X400) together with the values of the pathological and immunohistochemical variables, for each patient.

Most of the cancers in this cohort were classified in histological degrees 2 (20 samples) and 3 (89 samples), and only two samples were in degree 1. Also, a few samples have a histological degree, which is unknown.

Figures 4.1 to 4.3 show different basales samples, corresponding to the TNBC histological grades 2 and 3. The nuclear expression of the androgen receptors (AR) was also taken into account, displaying the samples with positive androgen receptors. Hormone receptors status, such as estrogen receptor (ER), estrogen receptor (ER) and AR expressions, are important because they serve to decide whether the cancer is likely to respond to hormonal therapy or other treatments. A positive AR implies a different type of TNBC (apocrine carcinoma)which is a subtype of TNBC that expresses AR, but often lacks ER and PR. Positive values of AR are more expanded through the TNBC samples with histological grade 2: 75% of the samples of this type, meanwhile only 15% of the samples with histological grade 3 present AR positive values.

It is possible to observe the high heterogeneity in the patterns that are present in these images. In the contribution it is explained the criteria used by the pathologist to establish the histological grades of TNBC.

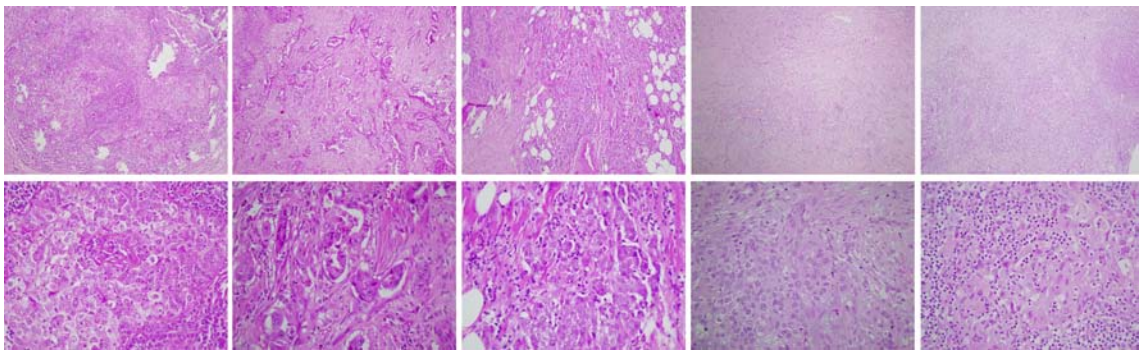


Figure 4.1: Basal images corresponding to histological grade 2, with AR=0, taken at two resolutions X100 and X400.

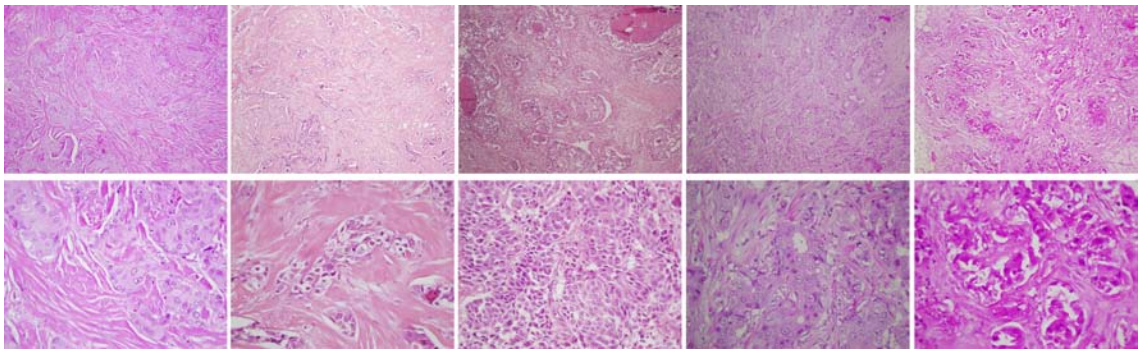


Figure 4.2: Basal images corresponding to histological grade 2, with AR=1, taken at two resolutions X100 and X400.

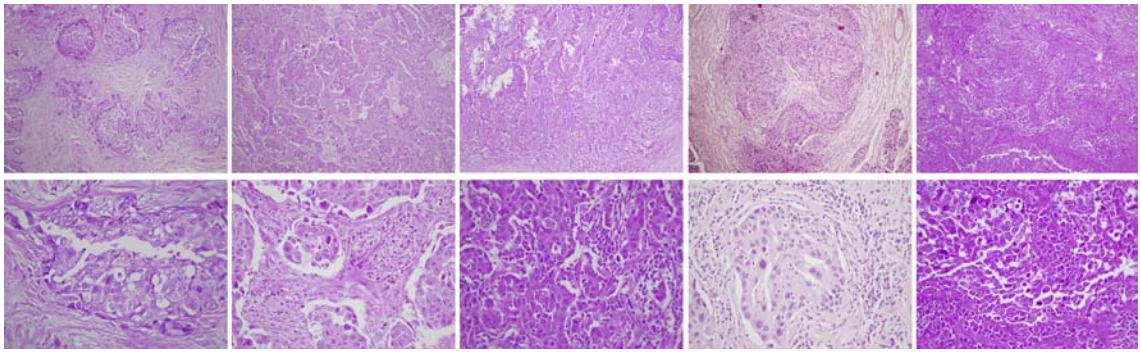


Figure 4.3: Basal images corresponding to histological grade 3, taken at two resolutions X100 and X400.

4.2. Classification Using Pathological and Immunohistochemical Variables

The first aim of our research is to analyze the most discriminatory pathological and immunohistochemical variables of the histological grade. This first part used the methodology designed by E. de Andrés Galiana. A list of currently used pathological and immunohistochemical variables is provided, and each TNBC sample is represented through its feature vector (variables were normalized in the interval $[0, 1]$, according to their respective cumulative distributions. Data preprocessing also includes the imputation of the clinical variables that have not been measured for some patients. Imputation is based in a nearest neighbor interpolation algorithm, and has been widely used in biomedical field to deal with missing data. This technique improves the accuracy in the predictions and do not modify the prognostic variables that are involved.

The first objective is to find the reduced list of such clinical variables that are most discriminative for the histological grade identification. For that purpose we apply the Generalized Fisher's ratio of the j^{th} attribute, for a binary classification problem.

The problem of d feature selection from a list of N_a attributes, in a multiclass problem, is as follows:

- Given a dataset $\{\mathbf{x}_i, y_i\}_{i=1,N}$, where $\mathbf{x}_i \in \mathbb{R}^{Na}$ are the N TNBC samples represented through its N_a features (clinical variables), and $y_i \in 1, ..., N_C$ is the class label (the histological grade of the samples, in this case).
- Finding a features subset of minimum size $d < N_a$ that contains the most discriminative features, that is: the distance between the samples of different classes is maximized, while the distance between the samples in the same class is minimized.

To solve this problem, Fisher's Ratio is calculated for each feature, as follows:

$$GFR_j = \frac{\sum_{k=1}^{N_C} n_k(\mu_{jk} - \mu_j)^2}{\sum_{k=1}^{N_C} n_k(\sigma_{jk})^2}, \tag{4.2.1}$$

where μ_{jk} and σ_{jk} are, respectively, a measure of the center of the distribution (median or mean) and a measure of its dispersion within the k^{th} class, corresponding to the j^{th} feature, μ_j, similarly, is a measure of the center of the distribution of the whole

data set corresponding to the j^{th} feature, and N_C is the number of classes. This is a generalization of Fisher's ratio that is used in Fisher's Linear Discriminant Analysis [15].

In this case we used the generalized Fisher's score for a binary classification problem, since the number of TNBC samples provided for one of the three classes was insufficient (only two) and we neglected it. Features are ranked by their decreasing Fisher's ratios. Most discriminatory attributes corresponds to higher Fisher's ratios. Furthermore, to obtain the minimum-size features list a Recursive Feature Elimination is performed , using the Leave-One-Out method for the prediction.

For the numerical experiments, feature selection was performed, both by considering the median and the mean as measure of the center of the distribution for the Fisher's score calculation, and two different lists of variables were obtained. Table 4.1 shows the ranked lists of features and its corresponding classification accuracies.

Ranked Features	FR	Max. Accuracy (%)
with Median		
Mitotic count (10HPF)	4.56	
Differentiation	4.55	
AR expression	2.60	
Tubular Formation	2.46	
Insitu	2.06	**96.4**
T	2.05	
N	1.99	
Ki67 expression	1.78	
with Mean		
Differentiation	2.41	
Mitotic count (10HPF)	2.00	
Ki67 expression	0.94	
AR expression	0.64	**94.4**
Tubular Formation	0.56	
pro-CollA1 intensity	0.24	
Bcl2 expression	0.22	
pro-Coll1A1 Score	0.21	

Table 4.1: Most discriminatory feature lists, both with median and mean, and their respective classification accuracies.

Interesting, the Fisher ratio with the median produced a higher discrimination than with the mean. In both cases the list of the most discriminatory variables included prognostic variables such as the Mitotic count, differentiation, AR expression or tubular formation, used by the pathologists. The maximum predictive accuracy was 96.4%.

4.3. Classification Using Histological Images

The second aim of this research is to analyze the possibility of performing an automatic histological grade prediction using TNBC images taken at two different resolutions. The main differences reside in the histological variables that have been described in the previous section, that are visually assessed by medical experts. Therefore, in this chapter we explore the possibility of capturing these characteristics using image processing techniques and machine learning.

Consequently, histological grade prediction is treated in this section as an automatic image classification problem, and the most relevant image attributes are analyzed: statistical based (histogram and variogram), spectral (discrete cosine transform), and image segmentation/regional descriptors (edges, texture and Zernike Moments). In this case, all attributes were calculated as global descriptors since TNBC image comparison should not be pixel-based.

TNBC image samples were represented by attribute vectors, and a detailed analysis of their discriminatory power was performed, in each case, at two steps:

- First, we showed the distributions of the histological images with histological grades 2 and 3 using a diagram of a PCA plot in two dimensions, differentiating the samples having a positive androgen receptors (AR). Figures 4.4 and 4.5 show this simple PCA analysis at both resolutions (X10 and X40). It can be observed that TNBC samples with HG = 2 are located in specific zones of the PCA map, mainly on three different clusters, surrounded by samples with HG = 3. Also, it seems that non-apocrine HG2 samples are only located in very restricted areas of the PCA diagram, in both figures. This simple projection methodd serves to help doctors to compare the location of new samples in the PCA map and, this way to confirm the coherence of their analysis with respect to other samples that have already been correctly annotated.

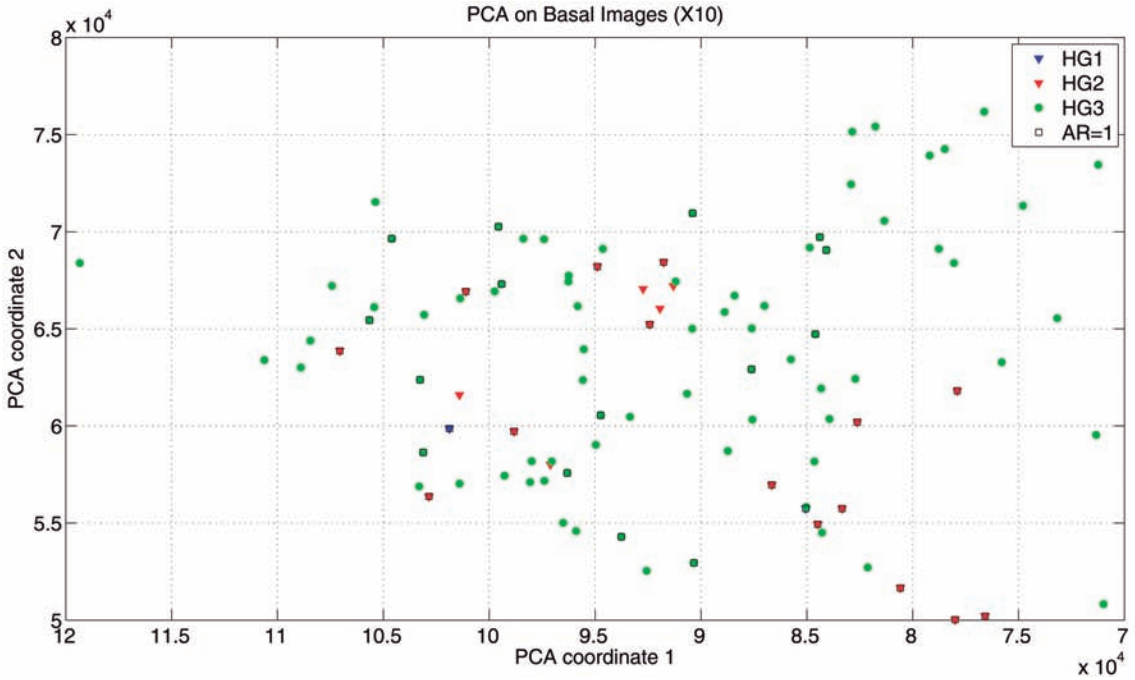


Figure 4.4: PCA 2D plot diagrams of basales at resolution X10.

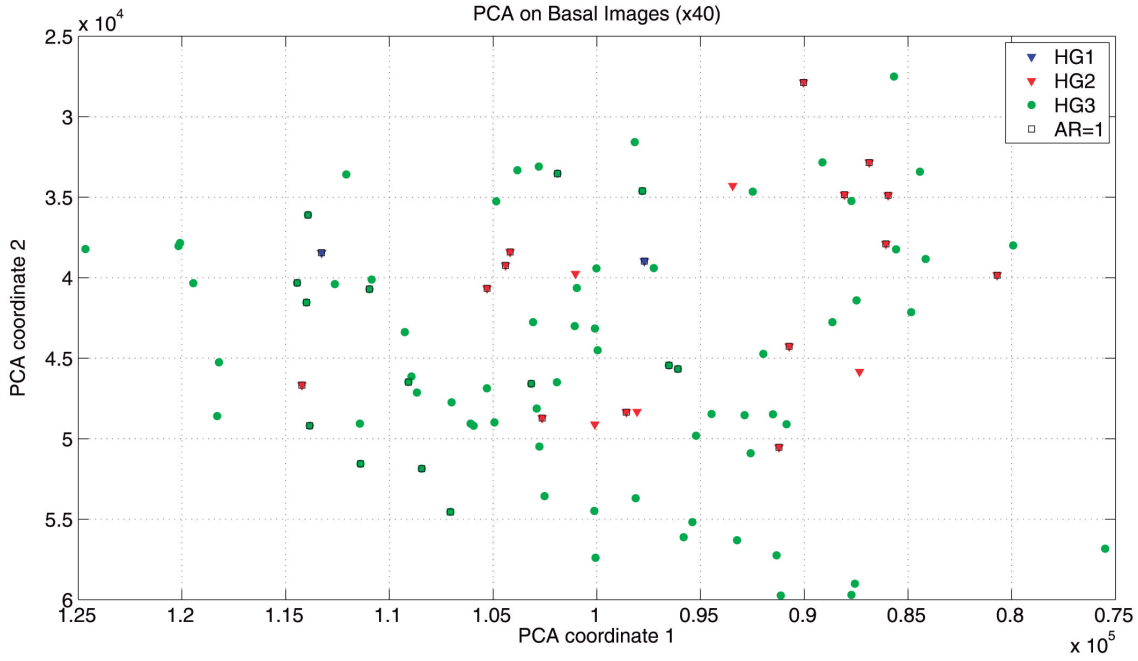


Figure 4.5: PCA 2D plot diagrams of basales at resolution X40.

Figure 4.6 shows the PCA plots in 2D of all the attributes. It can be observed that the HG2 and HG3 samples are located differently in each of these diagrams. These diagrams gave us a global view of how each attribute will differentiate the two histological grades.

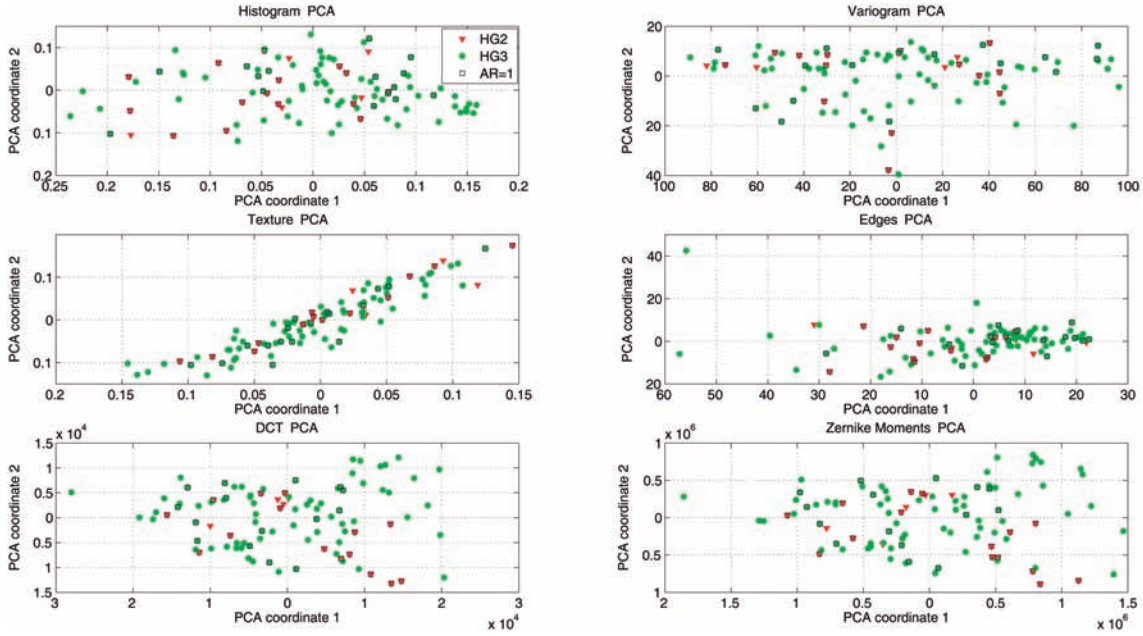


Figure 4.6: PCA 2D plot diagrams for each attribute.

- Second, we showed the discriminatory power of each attribute by representing the means of the samples with HG=2 and HG=3, separately in each color channel.

The conclusions of this analysis are shown in table 4.2, which presents the most discriminative parameters of each attribute. These parameters were used in the design of the individual classifiers based on these attributes. Furthermore, un unsupervised ensemble classification using the listed attributes was performed obtaining a median accuracy of 86.8%.

Posterior researches focuss on the supervised ensemble learning classification and exploration of the uncertainty space generated by the optimization via PSO of the combination of these attributes. Also, future works will consider the use of both pieces of information (pathological and immunohistochemical variables, and histological images) which will provide complementary information. It is expected that it will improve the accuracy in the classification of TNBC samples (histological grade and survival).

Image Attribute	Most discriminative
PCA	4^{th} order harmonics
Color Histogram	Green channel
Variogram	Green, Blue channels
Edges	First coefficients, all channels
DCT	First coefficients, all channels
Zernike Moments	Blue, Green channles

Table 4.2: Histological images attributes and their most discriminative features.

This methodology will be used in the future to asses the histological grade of the TNBC samples, and to analyze the possibility of predicting the patients survival.

The complete description of this research can be found in Appendix B, section B.1.

Chapter 5

Conclusiones y Líneas Futuras de Investigación

En esta tesis doctoral se presenta una metodología robusta para la clasificación, interpretación y reconocimiento automático de imágenes, lo que constituye un reto en el terreno de la inteligencia artificial, el reconocimiento automático de patrones y la visión por ordenador, con múltiples aplicaciones en numerosos campos de la biometría y de la biomedicina.

El diseño de dicha metodología se ha realizado gradualmente en tres etapas que han sido descritas en tres artículos enviados a revistas internacionales con revisión por pares e índice de impacto.

En el primer artículo, titulado ***Numerical Analysis and Comparison of Spectral Decomposition Methods in Biometric Applications***, se ha realizado un análisis y comparación numérica de diferentes técnicas de análisis espectral de imágenes, para la extracción de sus atributos, con especial énfasis en el problema de reconocimiento facial. Dichos métodos se han clasificado en dos categorías diferentes: 1) Métodos basados en el concepto de covarianza/correlación (PCA, NPCA, 2DPCA, Fisher's LDA y SVD) que requieren diagonalización de las matrices de covarianza/correlación correspondientes. Salvo la descomposición en valores singulares, dichos métodos están basados en todo el conjunto de imágenes de la base de datos. 2) Métodos basados en otros tipos de transformaciones ortogonales (DCT/DST, DCHT, DWHT, DWT y DHT) que no implican diagonalización y que se realizan sobre imágenes individuales. La ausencia de diagonalización es una notable diferencia a tener

en cuenta entre ambos tipos de métodos, ya que ésta posee un elevado coste computacional y en ocasiones es incluso imposible. Asimismo, los métodos la transformada de Chebychev Discreta (DCHT) y la PCA no centrada (NPCA) han sido aplicados por primera vez en esta publicación al problema de reconocimiento facial.

La aplicación de estos métodos a las bases de imágenes ORL y FERET ha alcanzado dos objetivos: 1) la reducción de la dimensión, uno de los principales retos en este campo, que consiste en establecer una representación de la imagen mediante vector de atributos de menor dimensión, que posea suficiente poder discriminatorio para realizar una clasificación con alta precisión. Este es un aspecto muy importante, ya que el problema de reconocimiento facial y la clasificación de las imágenes histológicas de cánceres de mama triplemente negativos, requieren una drástica reducción de la dimensión. 2) El diseño de un clasificador por vecino próximo (k-NN) basado en cada uno de los atributos espectrales analizados, un tipo de análisis (local o global) y un criterio de similitud (norma o coseno) definido en el correspondiente espacio de los atributos. Con este propósito, se han considerado dos tipos de análisis (global y local) y diferentes tipos de normas definidas en el espacio de los atributos, que han sido comparadas para cada método espectral, con objeto de poder diseñar de este modo el mejor clasificador en cada caso.

Las siguientes conclusiones se desprenden de este análisis:

1. Los métodos espectrales que actúan sobre todo el conjunto de imágenes poseen un mayor poder discriminador en la modalidad análisis local, mientras que los métodos basados en imágenes individuales proporcionan mejores resultados utilizando el análisis global. Al mismo tiempo, el análisis local permite la aplicación de los métodos basados en covarianza a imágenes de muy alta resolución.

2. Todos los métodos espectrales proporcionan resultados similares cuando el umbral de la energía que se utiliza en la reducción de la dimensión está ajustado adecuadamente. Por esta razón, hemos concluido que la mayoría de los resultados numéricos presentes en la literatura están influidos por la manera en la que los experimentos numéricos se habían realizado. Para resolver este problema, todos los experimentos numéricos fueron diseñados sobre el mismo número de simulaciones independientes, utilizando las mismas bases de datos de aprendizaje y de test para todos los atributos.

3. La transformada de wavelets discreta (DWT) es el método espectral más interesante, dado su mayor poder discriminador y su bajo coste computacional. Futuras investigaciones se realizarán sobre el uso de este método de reducción de la dimensión en diferentes problemas biométricos, y en particular el impacto de la selección del tipo de wavelet y el uso de otras partes de la DWT (diferencias horizontales, verticales y diagonales) en la capacidad predictiva de dicho método.

4. Finalmente, la conclusión más importante ha sido entender que ninguno de estos métodos es capaz de proporcionar una clasificación perfecta, aunque la base de datos de imágenes esté perfectamente registrada. Este hallazgo nos condujo al uso de otros tipos de atributos y a la combinación de diferentes tipos de clasificadores mediante técnicas de aprendizaje/ clasificación por ensambles.

La segunda publicación de esta tesis titulada ***Unsupervised Ensemble Classification for Biometric Applications***, trata sobre el diseño y la combinación de un conjunto diverso de clasificadores por vecino próximo (k-NN), que proporcionan de modo estable una gran precisión en el problema de reconocimiento facial. Esta metodología puede ser fácilmente adaptada a otro tipo de problemas biométricos y de reconocimiento de patrones, con aplicaciones diversas en biomedicina, autentificación, web semántica, etc.

Los objetivos de esta segunda publicación son los siguientes: 1) Extender el conjunto de atributos espectrales analizados en la primera contribución de esta tesis doctoral a otro tipo de atributos tales como los atributos estadísticos (histograma, vector de percentiles, variograma) y atributos de tipo regional (textura y detección de bordes). 2) El diseño de diferentes tipos de clasificadores por ensambles (CAV1, CAV2 y CBAG).

Se obtuvieron las siguientes conclusiones:

1. Los clasificadores k-NN basados en atributos estadísticos, y en particular el histograma local, son los que poseen una mayor precisión en el problema de reconocimiento facial. Además ninguno de estos clasificadores proporciona una precisión del 100% de modo estable o sistemático, sobre pruebas numéricas realizadas sobre la base de imágenes ORL.

2. Los clasificadores por ensambles mejora la estabilidad y la precisión de la clasificación.

3. El análisis de diversidad realizado sobre el conjunto de atributos seleccionados tuvo como objetivo eliminar redundancia en la clasificación y mejorar la estabilidad y la precisión del clasificador. El resultado de dicho análisis permitió seleccionar cinco atributos incorrelados: histograma local, variograma local, textura local, detección de bordes local y DWT global. Dichos clasificadores pueden ser todos pre-computados en el conjunto de imágenes de la base de datos. Esta característica es muy importante en el diseño del clasificador final (velocidad en la clasificación). El uso dichos atributos permitió mejorar la estabilidad y la precisión de los clasificadores por ensambles diseñados (CAV1, CAV2 y CBAG). CAV1 y CAV2 proporcionaron los mejores resultados, próximos en precisión al mejor clasificador individual (histograma local), aumentando la estabilidad del clasificador.

4. Finalmente, se concluyó que la precisión de dichos clasificadores se mejoraba dramáticamente mediante un enfoque supervisado, optimizando los parámetros de los cuales dependen dichos clasificadores, es decir, los pesos individuales de cada clasificador en el ensamble y el número de imágenes seleccionadas para hacer la clasificación final. Dichos parámetros fueron inicializados con la precisión individual de cada clasificador, mientras que el número de imágenes fue determinado mediante experimentación numérica y análisis de sensibilidad de dicho parámetro.

En la tercera publicación de esta trilogía, ***Exploiting the Uncertainty Space of Ensemble Classifiers in Face Recognition***, se aborda la necesidad de supervisión de dichos clasificadores, y en particular el algoritmo CAV1 que proporcionaba los mejores resultados en el enfoque no supervisado. El nuevo clasificador, denominado SCAV1, está basado en 6 clasificadores k-NN diversos, añadiendo a la lista anterior de clasificadores k-NN, el basapo en los momentos de Zernike. Además el clasificador SCAV1 explota la estructura del espacio de incertidumbre inducido por el proceso de supervisión (optimización de los parámetros de SCAV1). El muestro del conjunto de parámetros equivalentes de SCAV1 se realizó vía RR-PSO, dada la gran experiencia que se posee en este tipo de optimizadores y sus características de convergencia/exploración. El clasificador final adopta su decisión por consenso a partir del conjunto de imágenes proporcionadas mediante el proceso de exploración del espacio de incertidumbre de SCAV1.

Como conclusión principal de esta tercera publicación se obtuvo el objetivo final de

esta tesis doctoral, es decir, una cuasi perfecta y estable clasificación (precisión media próxima al 100% y rango intercuartilíco casi nulo), como se mostró numéricamente mediante la aplicación de dicha metodología a las bases de datos faciales ORL y PUT.

El problema de reconocimiento facial ha sido utilizado como problema tipo para el diseño de la metodología de identificación de patrones diseñada en esta tesis doctoral, a través de las bases de imágenes ORL, FERET y PUT. Aunque dicho problema biométrico posee como particularidad, que la estructura de las imágenes de la base de datos (distribución espacial de los diferentes componentes de la imagen) siguen el mismo patrón, dicha metodología has sido diseñada para ser aplicada en un amplio rango de aplicaciones biométricas y problemas de reconocimiento automático de imágenes. Se han obtenido resultados prometedores en la identificación del grado histológico de muestras correspondientes a cánceres triplemente negativos, contribución recientemente poresentada en el congreso internacional COGNITIVE 2014 (Venecia).

Finalmente, las líneas de investigación futuras son las siguientes:

1. El análisis de otras técnicas biométricas (iris, huellas digitales, firmas, etc) y otras aplicaciones biomédicas, como el diseño de diferentes tipos de robots biomédicos que ayuden a los doctores a la toma de decisiones.

2. La revisión de la metodología de aprendizaje por ensambles, intentando reducir/enriquecer el número de atributos necesarios.

3. El análisis de otro tipo de metodologías de aprendizaje, tales como Adaboost o Random Forest.

4. Un análisis más detallado de la DWT como método espectral de reducción de la dimensión. En particular se estudiará el impacto de la elección de la familia de la ondícula y la investigación de cómo elegir la ondícula adecuada, así como el uso de otras partes de la DWT (diferencias horizontales, verticales y diagonales) en la identificación de patrones.

5. Un análisis más completo del espacio de incertidumbre de los problemas de clasificación/aprendizaje supervisados. En particular, se estudiará como promocionar la toma de decisiones en problemas de reconocimiento de patrones con su consiguiente análisis de incertidumbre. Para conseguir tal objetivo será necesario

optimizar el conjunto de entrenamiento (aprendizaje) y la construcción de diferentes tipos de clasificadores (no sólo de tipo k-NN) que aplicados conjuntamente provean la clase estimada y su incertidumbre asociada. El diseño será juzgado satisfactorio cuando la clase predicha para una nueva muestra con una incertidumbre a posteriori baja, nunca sea mal clasificada, y una incertidumbre a posteriori creciente sea capaza de identificar las muestras cuya clasificación es ambigua. Cabe reseñar que dicho enfoque es ante todo pragmático, como lo ha sido el punto de vista adoptado a lo largo y ancho de esta tesis doctoral.

Chapter 6

Conclusions and Future Research

This PhD thesis presents a robust methodology for the automatic image interpretation and recognition, which is a challenging problem in artificial intelligence, pattern recognition and computer vision with applications in many fields of biometry, and biomedicine.

The design of this methodology was realized gradually, at three steps that were presented in three articles published in Peer Reviewed Journals with Impact Index:

1. The first article, titled ***Numerical Analysis and Comparison of Spectral Decomposition Methods in Biometric Applications***, comprises an exhaustive analysis and comparison between various spectral image attributes, with application to the face recognition problem. The methods analyzed in this paper can be classified into two categories: 1) Covariance-based: PCA, NPCA, 2DPCA, Fisher's LDA and SVD, these methods require diagonalization of the covariance/correlation matrices, and they are based on the whole set of images (except SVD), and 2) Covariance-free: ICA, DCT, DST, DCHT, DWHT, DWT and DHT, which do not involve diagonalization, an important difference to take into account, since diagonalization is usually computationally very expensive, and in some cases impossible.

   Also, the spectral methods DCHT and the NPCA were introduced in this paper for biometric applications, and applied to the face recognition problem.

   The application to the the face recognition problem, specifically to the ORL and FERET face databases, achieved two main objectives:

- Model dimensionality reduction, one of the main challenges in this field consisting of establishing a low-dimensional feature representation of the images having enough discriminatory power to perform high accuracy classification. This issue was very important in our case, since the applications to the face recognition problem and to the histological images classifications required a drastic dimensionality reduction due to the extensive databases that we deal with.
- The design of individual nearest neighbor classifiers based on: each spectral attribute, a type of analysis and a distance defined in the correspondent space of attributes. For that purpose, two kinds of analysis were considered (global and local) and different norms were compared in order to deduce the tendency of each spectral attribute and to select the best classifier.

Also, several important conclusions arised from this analysis:

- The spectral methods who act on the whole set of images usually perform better for local analysis, while methods based on orthogonal transformations of single images usually performed better for global analysis. At the same time, local analysis allows the application of covariance-based methods to very large images.
- All the spectral methods performed the same when the energy threshold used to achieve the model reduction was properly set. Therefore, we concluded that most of the numerical results presented in the literature were biased due to the manner in which the numerical experiments were performed. To solve out this problem, the numerical experiments were performed systematically on the same number of random independent simulations (100) with the same learning and testing databases, for all the different attributes.
- The DWT seems to be the most interesting spectral method, due to its high accuracy and its lower computational cost. Therefore, further research will be dedicated to the use of wavelets as a spectral method for model reduction in biometric problems, and particularly, about the impact of the selection of the type of wavelets, and the use of other parts of DWT (horizontal differences, vertical differences and diagonal differences) for the identification.
- Finally, and the most important, none of these methods was able to provide a

perfect classification, leading us to the use of other different image attributes and to the use of ensemble learning methodologies.

2. The second article ***Unsupervised Ensemble Classification for Biometric Applications***, was based on the conclusions of the previous step, and focused on the design of an ensemble learning methodology that combine a set of simple and diverse nearest-neighbor classifiers, with very high and stable accuracies for the face recognition problem. This methodology can be easily adapted to solve other biometric and pattern recognition problems, with applications in different fields, such as biomedicine, security, semantic web, etc.

   The objectives achieved in this research are the following:

   - The set of image attributes was extended and other types of attributes were introduced and analyzed: statistical (histogram, percentiles, variogram), region descriptors (texture and edges), and also, spectral attributes. We concluded that statistical attributes are the individual classifiers that provided the highest accuracies, although no single attribute is able to provide systematically 100% accuracy over the ORL database.
   - The design of three ensemble learning algorithms that combine individual classifiers based on the analyzed image attributes, which improves the stability and the accuracies of each of these classifiers.
   - Diversity analysis was performed over the set of selected attributes in order to eliminate redundancies and improve the ensemble learning methodology. The results of this analysis allowed us to select the five most diverse attributes: the local histogram, the local variogram, the local texture, the local edges and the global DWT. The use of this set of five uncorrelated classifiers, improved the accuracies and stability in the prediction obtained by the ensemble learning algorithms: CAV1 and CAV2 performed very close to the local histogram and the stability of the classification increased even more.

   Finally, we concluded that our methodology could be dramatically improved by the use of a a supervised approach that will set the ensemble learning parameters to optimum ones.

3. The third publication, titled ***Exploiting the Uncertainty Space of Ensemble Classifiers in Face Recognition***, responded to the final conclusion of the previous research about the need of a supervised approach to improve the results of the designed ensemble learning classifiers. The methodology defined in the second article was expanded to the supervised case, by optimizing *CAV1*'s parameters and a new algorithm was defined.

   The new supervised algorithm, called *SCAV1*, is based on a set of 6 diverse nearest-neighbor classifiers (local histogram, local variogram, local texture, global edges, global DWT and global Zernike moments), and takes advantage of the uncertainty space of this ensemble classifier. The optimization/sampling of the parameters (the weights of the nearest neighbor classifiers and the number of image candidates) is done via a powerful particle swarm optimizer (RR-PSO).

   The final classifier uses majority voting on the set of images provided by exploiting the uncertainty space of *SCAV1*. The application of this methodology to two publicly available image databases (ORL and PUT), shows that we are able to obtain almost perfect classification, with very high stable accuracies.

The face recognition problem was used as benchmark to design all the original algorithms presented in this PhD thesis, performing the numerical experiments on the ORL and also on the PUT face databases. Although the face recognition problem (and most of the biometric techniques) has the particularity that the structure of the image samples (the spatial disposal of the components) follow the same pattern, the methodology presented in this PhD has been thought to be applicable to other important biometric applications and image recognition problems. Very promising results were presented in our article for the International Congress Cognitive 2014, for the identification of the histological degree of Triple Negative Breast Cancers (TNBC) samples.

Future research will be devoted to:

1. the analysis of other biometric techniques (iris recognition, fingerprint identification, etc), and also, to other important biomedical applications, such as the design of biomedical robots able to asses medical doctors in their decisions.

2. the review of the ensemble learning methodology, trying to improve it by reducing

the attributes list, and/or by testing other attributes that have not been explored yet.

3. the analysis of other type of learning methodologies, such as, adaboost and random forest.

4. the detailed analysis of the DWT use as spectral method for the model reduction, particularly, the impact of the selection of the type of wavelets, and the use of other parts of DWT (horizontal differences, vertical differences and diagonal differences) for the identification.

5. a complete analysis of the uncertainty of a classification problem will be performed, and the decisions will be made with its corresponding assessment. For that purpose, il will also be necessary to optimize the learning data set, and/or to build different types of classifiers that will be applied together to provide an estimated class. The design will be considered successful when the predicted class of an incoming sample with low posterior uncertainty will never be misclassified, and an increasing uncertainty will point to the samples whose classification is ambiguous.

Appendix A

Publications

A.1. Numerical Analysis and Comparison of Spectral Decomposition Methods in Biometric Applications

Authors:
Juan Luis Fernández Martínez
Ana Cernea

International Journal of Pattern Recognition and Artificial Intelligence
Vol. 28, No. 1, pp. 14560
DOI: 10.1142/S0218001414560011.
Published on 13 January 2014.

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 28, No. 1 (2014) 1456001 (33 pages)

DOI: 10.1142/S0218001414560011

World Scientific
www.worldscientific.com

NUMERICAL ANALYSIS AND COMPARISON OF SPECTRAL DECOMPOSITION METHODS IN BIOMETRIC APPLICATIONS

JUAN LUIS FERNÁNDEZ-MARTÍNEZ* and ANA CERNEA†

Mathematics Department, Oviedo University
C/ Calvo Sotelo s/n 33007 Oviedo, Spain
**jlfm@uniovi.es*
†*cerneadoina@uniovi.es_name*

Received 13 May 2013
Accepted 11 November 2013
Published 13 January 2014

Face recognition is a challenging problem in computer vision and artificial intelligence. One of the main challenges consists in establishing a low-dimensional feature representation of the images having enough discriminatory power to perform high accuracy classification. Different methods of supervised and unsupervised classification can be found in the literature, but few numerical comparisons among them have been performed on the same computing platform. In this paper, we perform this kind of comparison, revisiting the main spectral decomposition methods for face recognition. We also introduce for the first time, the use of the noncentered PCA and the 2D discrete Chebyshev transform for biometric applications. Faces are represented by their spectral features, that is, their projections onto the different spectral basis. Classification is performed using different norms and/or the cosine defined by the Euclidean scalar product in the space of spectral attributes. Although this constitutes a simple algorithm of unsupervised classification, several important conclusions arise from this analysis: (1) All the spectral methods provide approximately the same accuracy when they are used with the same energy cutoff. This is an important conclusion since many publications try to promote one specific spectral method with respect to other methods. Nevertheless, there exist small variations on the highest median accuracy rates: PCA, 2DPCA and DWT perform better in this case. Also all the covariance-free spectral decomposition techniques based on single images (DCT, DST, DCHT, DWT, DWHT, DHT) are very interesting since they provide high accuracies and are not computationally expensive compared to covariance-based techniques. (2) The use of local spectral features generally provide higher accuracies than global features for the spectral methods which use the whole training database (PCA, NPCA, 2DPCA, Fisher's LDA, ICA). For the methods based on orthogonal transformations of single images, global features calculated using the whole size of the images appear to perform better. (3) The distance criterion generally provides a higher accuracy than the cosine criterion. The use of other p-norms ($p > 2$) provides similar results to the Euclidean norm, nevertheless some methods perform better. (4) No spectral method can provide 100% accuracy by itself. Therefore, other kind of attributes and supervised learning algorithms are needed. These results are coherent for the ORL and FERET

*Corresponding author.

databases. Finally, although this comparison has been performed for the face recognition problem, it could be generalized to other biometric authentication problems.

Keywords: Spectral methods; model reduction; face recognition.

1. Introduction

Biometric applications is one of the most important fields of application of image processing techniques and have a great interest in science and technology. Security issues[7] and biomedical applications[23] are fields where biometric applications are widely used.

Many different methods of supervised and unsupervised classification have been already presented in the literature,[27] but no numerical comparison among all of these methods has been performed using the same database and the same training conditions. In this paper, we apply the main spectral decomposition methods for face recognition: Principal Component Analysis (PCA) and its variants, Fisher's Linear Discriminant Analysis, Independent Component Analysis (ICA), Singular Value Decomposition (SVD), Discrete Cosine/Sine Transforms (DCT/DST), Discrete Chebyshev Transform (DCHT), Discrete Walsh-Hadamard Transform (DWHT), Discrete Wavelet Transform (DWT) and Discrete Hilbert Transform (DHT) on the ORL and Face Recognition Technology (FERET) databases of face images. Faces are represented by their spectral features, that is, their projections onto the different spectral basis. Classification is performed *ad hoc* using different norms and the cosine criterion defined in the space of the corresponding spectral attributes.

This analysis sheds light about the use of spectral methods in biometric applications, since it shows that determining the optimum energy cut off for dimensionality reduction is crucial for achieving high accuracy rates, and it is even more important than the spectral method that is used. Also the methods which deal with the whole training database perform better using local spectral features. On the contrary, methods based on single images seem to perform better using global features. The conclusions of this analysis can be generalized to other biometric problems and applications.

2. The Face Recognition Problem and the Algorithm of Nonsupervised Learning

The face recognition problem consists in, given a new incoming face image, identifying the individual corresponding to it, from a database of different face images of known individuals. The learning database (Bd) contains N face images

$$Bd = \{I_{k=1,\ldots,N} \in \mathcal{M}_{m\times n}(\mathbf{N})\},$$

corresponding to C individuals (or classes). We will assume that for each of these individuals the learning database contains n_p poses that are used to perform the classification of the new input images. Therefore the total number of poses in Bd

verifies $N = n_p \cdot C$. In all the experiments performed over the ORL database we dispose of $n_p = 5$ poses for each individual, meanwhile, $n_p = 2$ for the FERET database. In order to perform the classification, images are represented by a feature vector $\mathbf{v}_k \in \mathbf{R}^s$, obtained for each spectral method of analysis. The dimension s_k of $\mathbf{v}_k$ depends on the dimensionality reduction achieved by each spectral technique. For all the spectral methods we have used the same energy cut off of 99.5% to achieve the dimensionality reduction. The different spectral methods used in this paper will be presented in Sec. 3.

The classification algorithm proceeds as follows: given a new image $I \in \mathcal{M}_{m\times n}(\mathbf{N})$, $I \notin Bd$, corresponding to a new pose of any of the C individuals of the database, characterized by its features vector $\mathbf{v}_I$, finding the most similar image in the database Bd, according to a criterion defined over the space of attributes $\mathbf{R}^{s_k}$. Two different criteria are used in this paper:

- Minimize the distance between the new image J and the database images in a certain norm p, defined over $\mathbf{R}^{s_k}$: $d(I, I_k) = \|\mathbf{v}_I - \mathbf{v}_k\|_p$. In this paper, we have used the Euclidean norm ($p = 2$) and also other p-norms such as $p = 3, 5$, and the infinite norm.
- Maximize the cosine of the angle between the above mentioned images represented by their attributes vectors:

$$\cos(I, I_k) = \frac{\mathbf{v}_I \cdot \mathbf{v}_k}{\|\mathbf{v}_I\|_2 \|\mathbf{v}_k\|_2}.$$

  The cosine criterion is related to the existence of a dot product (for instance for $p = 2$).

Although both criteria applied to the same image are equivalent, we will show that they will provide different recognition accuracies when applied over the whole set of images (testing database). In the cases where $\cos(I, I_k)$ provides higher accuracies, that means this criterion performs better with normalized images $\mathbf{n}_I, \mathbf{n}_{I_k}$, since

$$\cos(I, Ik) = \frac{\mathbf{v}_I}{\|\mathbf{v}_I\|} \cdot \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|} = \mathbf{n}_I \cdot \mathbf{n}_{I_k}.$$

The algorithm presented in this paper belongs to the class of nonsupervised classification since the distance $d(I, I_k)$, and the cosine $\cos(I, I_k)$, are defined *ad hoc*, that is, no information from the learning database classes is used in their definition to optimize the classification of the testing database. Finally, local and global features of the images are used to compute the set of attributes and perform their numerical comparison.

3. Spectral Decomposition Methods for Face Recognition

The spectral methods used in this paper can be divided into two categories: (1) those that act in the whole database of images (PCA, NonCentered PCA (NCPA),

2DPCA, Fisher's LDA and ICA); (2) those that act on single images, although they could also be applied to the whole training image database. This category includes the rest of orthogonal transformations that are used in this paper. Another important classification concerns the spectral methods that involve diagonalization of covariance/correlation matrices (PCA, NPCA, 2DPCA, Fisher's LDA and SVD), and those that are covariance-free (they do not involve diagonalization). This difference is very important since diagonalization is usually computationally very expensive, and in some cases impossible.

3.1. *The Frobenius norm of a matrix and the energy content of an image*

A gray color image of size $m \times n$ pixels can be considered as a rectangular matrix in $\mathcal{M}_{m\times n}$. This concept is also valid in the case of a color image, considering independently each color channel. Therefore, the energy content of an image is related to the energy of a matrix considered as a $2-D$ discrete signal.

The energy of a matrix $A \in \mathcal{M}_{m\times n}$ can be defined through its Frobenius norm. This concept is of paramount importance to understand energy compression and dimensionality reduction of images. The entrywise norm of a matrix consists in treating the $m \times n$ matrix as a vector of size mn. The p-entrywise norm of $A \in \mathcal{M}_{m\times n}$ is defined as:

$$\|A\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{\frac{1}{p}}. \tag{1}$$

The special case $p = 2$ is the so-called Frobenius norm of A:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}. \tag{2}$$

It is obvious to prove that:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2 = \text{trace}(A^{\text{T}}A) = \text{trace}(AA^{\text{T}}). \tag{3}$$

Therefore,

$$\|A\|_F = \sqrt{\text{trace}(A^{\text{T}}A)} = \sqrt{\text{trace}(AA^{\text{T}})}. \tag{4}$$

3.2. *Linear operators and orthogonal transformations*

A gray digital image $I(m, n)$ can be regarded as the matrix of a linear operator between two linear spaces $\mathbb{R}^n$ and $\mathbb{R}^m$. In the case of a color image there exist three different linear operators, one for each color channel (R, G and B).

Given an image $I(m,n)$, it is possible to define in several ways, two different orthogonal transformations U, V, such as:

$$I = USV^{\mathrm{T}}, \tag{5}$$

where U and V form orthonormal basis of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

The interests of the orthogonal transformations are the following:

- They conserve the energy of the image, that is the Frobenius norm of matrix I. The demostration is as follows:

$$\|I\|_F = \sqrt{\mathrm{trace}(I^{\mathrm{T}}I)} = \sqrt{\mathrm{trace}(VS^{\mathrm{T}}SV^{\mathrm{T}})} = \sqrt{\mathrm{trace}(S^{\mathrm{T}}S)} = \|S\|_F.$$

- Matrix S is blocky diagonal if U and V contain as columns the left and right singular vectors provided by the SVD of I. In other cases, such as DCT, DST, DCHT and DWT, these orthogonal transformations serve to decorrelate the pixels of I by compressing its energy onto the first harmonics (spectral modes of the image). The pixel decorrelation is based on the fact that orthogonal transformations (in our case U and V) induce rotations to the principal axes of the image in $\mathbb{R}^m$ and $\mathbb{R}^n$. In the case of PCA, this rotation is induced by the orthonormal basis calculated through the experimental covariance matrix. Other covariance-free techniques are presented to induce these rotations.
- Energy compression consists in finding the number of transformed pixels p, q of S (1), such as:

$$\|I - S(1:p, 1:q)\|_F < \theta,$$

  where θ is a prescribed energy threshold, and $S(1:p, 1:q)$ represents the $p \times q$ upper block of S where the energy has been concentrated. The dimensionality reduction is achieved from $m \times n$ pixels to $p \times q$ frequency components. In the case of the *SVD*, only the q first singular values are needed, because S is blocky-diagonal.

Orthonormal transformations based on a set of training images $\{I_1, I_2, \ldots, I_N\}$, follow similar principles.

3.3. *Methods based on the whole set of training images*

3.3.1. *Centered and noncentered PCA (eigenfaces)*

PCA is a popular unsupervised method that finds a linear subspace in which the variance of the projected data is maximized. PCA applied to image analysis and face recognition[19,34,36] is an orthogonal transformation aimed at decorrelating the pixels of a given image, concentrating its energy in a small number of principal components, and reducing the face recognition problem dimensionality. Eigenfaces has been widely investigated and became one of the most popular approaches in face recognition.[21] PCA provides high accuracy recognition rates and it is a fairly robust

method to lighting variations, nevertheless its performance degrades with scale changes.[16]

Although PCA has been classically interpreted as a statistical method of data analysis, in reality it has a close relationship with linear algebra and the SVD. PCA finds an orthonormal basis of the *pixel space*($\mathbb{R}^{N\text{pixels}}$) by diagonalizing the mean centered ($\mu \neq 0$) or noncentered ($\mu = 0$) empirical image covariance matrix, $S \in \mathcal{M}_{N\text{pixels}\times N\text{pixels}}$, constructed with the whole database as follows:

$$S = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^{\mathrm{T}} = X_c X_c^{\mathrm{T}},$$

where $\mu = \frac{1}{N}\sum_{k=1}^{N} X_k$ is the images sample mean, N is the number of sample images contained in the learning database, N_{pixels} is the number of pixels of each image (we suppose that all images in the sample have the same dimensions), $X = [X_1, X_2, \ldots, X_N] \in \mathcal{M}_{N\text{pixels}\times N}$, where $X_i \in \mathbb{R}^{N\text{pixels}}$ are the database images transformed into $1 - D$ column vectors, and $X_c = X - \mu$ is the centered image matrix, once the image experimental mean has been substracted.

In both cases (centered and noncentered PCA), S is a symmetric semidefinite positive matrix, thus, it admits orthogonal diagonalization as follows:

$$S = UD_1U^{\mathrm{T}},$$

where $U \in \mathcal{M}_{N\text{pixels}\times N\text{pixels}}$. Typically, $_{N\text{pixels}} \gg N$ (the number of training images), and the matrix S becomes very difficult (or even impossible) to diagonalize. The rank of S is $N - 1$ (or N in the case of noncentered covariance), that is, only $N - 1$ (or N) eigenvectors of S (contained in U) are not related to the null space of S.[35] For this reason it is compulsory to diagonalize

$$R = \sum_{i=1}^{N} (X_i - \mu)^{\mathrm{T}}(X_i - \mu) = X_c^{\mathrm{T}} X_c \in \mathcal{M}_{N\times N},$$

which is the experimental covariance matrix in the space of images. Furthermore, we have:

$$R = VD_2V^{\mathrm{T}}.$$

Matrices U, V, D_1 and D_2 are related to the *SVD* of X_c as follows:

$$X_c = U\Sigma V^{\mathrm{T}}, D_1 = \Sigma^{\mathrm{T}}\Sigma, D_2 = \Sigma\Sigma^{\mathrm{T}}.$$

The eigenfaces algorithm consists of:

- Diagonalizing R (instead of S) to determine $V \in \mathcal{M}_{N,N}$.
- Finding $P = X_cV = U\Sigma \in \mathcal{M}_{N\text{pixels}\times N}$. The first $N - 1$ (or N in the case of NPCA) principal components are the column vectors of P, divided by their norm

$$\mathbf{u}_k = \frac{P(:,k)}{\|P(:,k)\|_2}.$$

- The total variance of the images in the database is:

$$\sigma^2_{\text{tot}} = \text{trace}(S) = \text{trace}(R) = \text{trace}(D_1) = \text{trace}(D_2).$$

The dimensionality reduction is obtained by retaining the q first principal components $\mathbf{u}_k$, such as:

$$|\sigma^2_{\text{tot}} - \sum_{k=1}^{q} \lambda_k| < \theta,$$

where θ is the energy threshold and λ_k are the non-null eigenvalues of D_1 (or D_2). The eigenvector associated with the largest eigenvalue is the one that reflects the greatest variance in the images, which means that using a few principal components (the first q in this case), we can represent quite accurately the whole image database variability, obtaining an important dimensionality reduction. In the case of the noncentered PCA the first principal component $\mathbf{u}_1$ plays a similar role to the experimental mean μ in the centered PCA. Thus, this term should not be considered in the energy reconstruction to determine the number of the principal components that are needed to achieve the θ energy cutoff, that is, the number of principal components has to be determined by compressing the energy of

$$X - \lambda_1 \mathbf{u}_1.$$

- Finally, defining $W = U(:, 1:q)$, each image X_i is projected onto the q first $\mathbf{u}_k$ vectors, obtaining the feature vector $Y_i = W^{\text{T}} X_i$, and achieving a dimensionality reduction from n pixels to the q first principal coordinates.

Figure 1 shows the spectrum (the singular values of X_c) for a training set composed of 200 images of the ORL database. In this case we only need $q = 93$ principal components to achieve an energy cutoff of 99.5% ($\theta = 0.05\%$), instead of $N_{\text{pixels}} = 10\,304$ pixels. Figure 2 shows the first six eigenfaces for the above mentioned example.

3.3.2. *2DPCA*

2DPCA for face recognition has been introduced by Yang *et al.*[37] As opposed to conventional PCA, 2DPCA is based on 2D matrices, I_i, rather than 1D vectors, X_i. That is, each image I_i does not need to be previously transformed into a column vector and image covariance is constructed directly using the original image matrices. This covariance matrix is, in contrast with the scatter matrix of PCA, much smaller. The 2DPCA diagonalizes the following mean centered covariance matrix to find the orthogonal projection basis:

$$S_M = \frac{1}{N} \sum_{i=1}^{N} (I_i - \overline{I})^{\text{T}} (I_i - \overline{I}),$$

where $\overline{I}$ is the mean image matrices calculated pixel by pixel over the learning database. As S_M is the sum of symmetric and semidefinite matrix, it admits

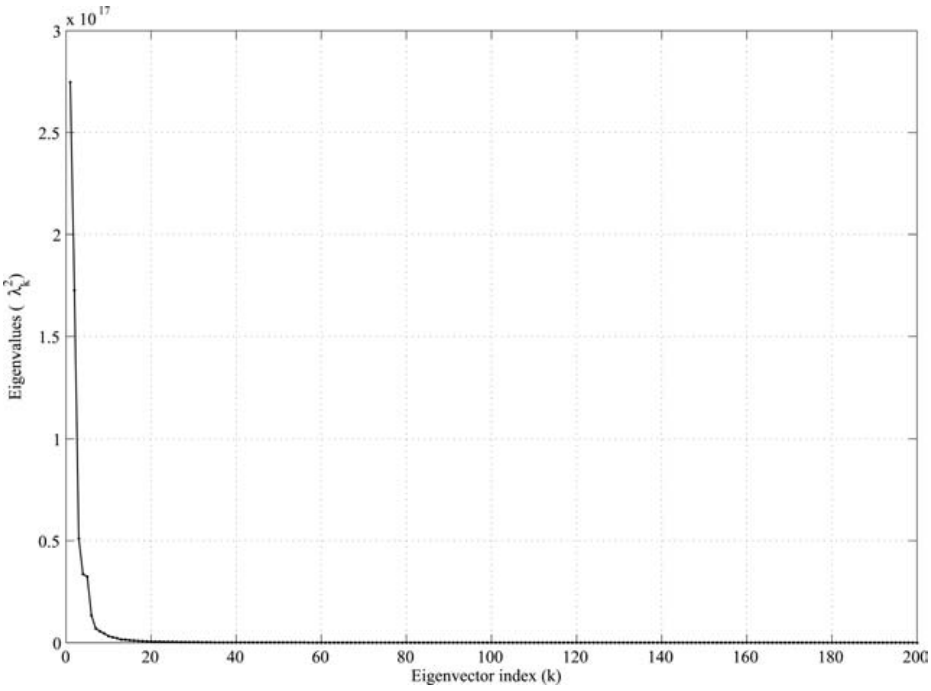


Fig. 1. PCA: energy spectrum for the first 200 eigenfaces of the ORL training database.

orthogonal diagonalization as in the PCA case. If $I_i \in \mathcal{M}(m,n)$, then $S_M \in \mathcal{M}(n,n)$. In the case when $n \gg m$, S_M, which has a maximum rank of $m-1$, it can be calculated by diagonalizing the matrix:

$$R_M = \frac{1}{N}\sum_{i=1}^{N}(I_i - \overline{I})(I_i - \overline{I})^{\mathrm{T}} \in \mathcal{M}(m,m).$$

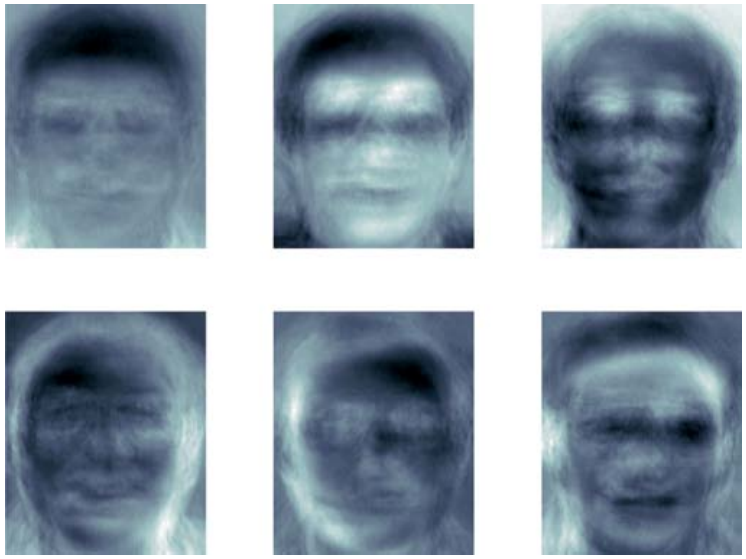


Fig. 2. The first six eigenfaces for the ORL training database.

Eventually, a noncentered 2DPCA, $\overline{I} = 0$, can also be used, as in the PCA case. The dimensionality reduction and the feature projection follows the same logic than for PCA. The recognition rates for 2DPCA seem to be higher than PCA and also to be computationally more efficient.[9,37] Interesting to note that in this case no 2DPCA eigenfaces can be shown, since the orthogonal base lies in the space $\mathbb{R}^m$ or $\mathbb{R}^n$.

3.3.3. *Fisher's linear discriminant (Fisherfaces)*

Fisher Linear Discriminant Analysis (LDA),[10] for short Fisherfaces, has been applied to the face recognition problem by Belhumeur *et al.*[4] This technique is also based on a linear projection W, which attempts to reduce dimensionality while preserving as much of the class discriminatory information as possible. More precisely, the transformation W is selected such as the projections of the sample database X have the maximum degree of separation. Fisher's solution to this problem is the calculation of W which maximizes the differences between classes, normalized by the scatter within-class. For that purpose, between- and within-class scatter matrices are defined:

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}},$$

$$S_W = \sum_{i=1}^{C} \sum_{X_k \in C_i} (X_k - \mu_i)(X_k - \mu_i)^{\mathrm{T}},$$

where C is the total number of classes in the database, N_i is the number of images in each class C_i, and μ_i is the mean of the images in class i.

Fisher's linear discriminant analysis[5] is the linear transformation $W^{\mathrm{T}}X_k$ that maximizes the separation between classes (thus, it maximizes the ratio of the between-scatter S_B to the within-scatter S_W), by optimizing the functional in W:

$$J(W) = \frac{|W^{\mathrm{T}} S_B W|}{|W^{\mathrm{T}} S_W W|}.$$

It can be shown that the column vectors of the optimal matrix

$$W_{\mathrm{opt}} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{C-1}],$$

are the eigenvectors corresponding to the $C-1$ largest eigenvalues of the generalized eigenvalue problem[4]:

$$S_B \mathbf{w}_k = \lambda S_W \mathbf{w}_k, \quad k = 1, \ldots, C-1.$$

For the calculation of W the central idea is to perform the simultaneous diagonalization of matrices S_B and S_W. The solution proposed in Ref. 38 is to find W discarding the null space of S_B which contains no useful information for classification and, on the contrary, preserving the null space of S_W which holds the most

discriminative information. This is achieved by diagonalizing S_B first. Let V be the eigenvectors matrix which orthogonally diagonalizes the symmetric matrix S_B: $\Lambda = V^{\mathrm{T}} S_B V$, with $V^{\mathrm{T}} V = I$. Let Y be the matrix formed by the q first columns of V, formed by discarding the zero eigenvalues and those that are close to 0. Then we have: $Y^{\mathrm{T}} S_B Y = D_B$, where D_B is the $q \times q$ principal submatrix of Λ, and finally:

$$(YD_B^{-\frac{1}{2}})^{\mathrm{T}} S_B Y D_B^{-\frac{1}{2}} = I.$$

Calling $Z = YD_B^{-\frac{1}{2}}$, now we need to determine the matrix which diagonalizes $Z^{\mathrm{T}} S_W Z$. Let U be this matrix: $U^{\mathrm{T}} Z^{\mathrm{T}} S_W Z U = D_W$, with $U^{\mathrm{T}} U = I$.

It can be concluded that matrix $W = U^{\mathrm{T}} Z^{\mathrm{T}}$ diagonalizes S_W and unitizes S_B at the same time and thus it is the one which fulfills Fisher's criterion:

$$\max J(W) = \frac{|W^{\mathrm{T}} S_B W|}{|W^{\mathrm{T}} S_W W|} = \frac{|I|}{|D_W|}.$$

S_B and S_W can be written as follows:

$$S_B = \Phi_B \Phi_B^{\mathrm{T}},$$
$$S_W = \Phi_W \Phi_W^{T},$$

with

$$\Phi_B = [\sqrt{N_1}(\mu_1 - \mu), \sqrt{N_2}(\mu_2 - \mu), \ldots],$$
$$\Phi_W = [x_1 - \mu_1, x_2 - \mu_1, \ldots, x_{N_1} - \mu_1, \ldots].$$

As it happens in the PCA case, it is numerically more convenient to diagonalize $\Phi_B^{\mathrm{T}} \Phi_B \in \mathcal{M}_{C \times C}$, and $\Phi_W^{\mathrm{T}} \Phi_W \in \mathcal{M}_{N_{\mathrm{imag}} \times N_{\mathrm{imag}}}$. This way, an important dimensionality reduction is obtained, since $C \ll N_{\mathrm{pixels}}$ and $N_{\mathrm{imag}} \ll N_{\mathrm{pixels}}$.

The feature vector is obtained by projecting the images using W_{opt}. Figure 3 shows the first six Fisherfaces, i.e. the first six vectors of the projection basis. Extensive experimental results demonstrated that Fisherfaces have error rates lower than eigenfaces in face recognition.[4] The main drawback of this method is that it needs enough images in each class to calculate S_W. Nevertheless some LDA variants exist that are able to deal with the small sample size problem.[6] Also a two-dimensional LDA (similar to 2DPCA) has been proposed.[33] In this paper, we use the standard LDA approach presented in Ref. 38.

3.3.4. *Independent component analysis*

ICA is a signal processing technique whose goal is to express a set of signals (considered as random variables) $X_1, \ldots, X_N$ as linear combination of statistically independent components of a random vector $S = [S_1, \ldots, S_n]^{\mathrm{T}} \in \mathcal{M}_{n \times N\mathrm{pixels}}$, called the source vector, with the restriction $n \leq N$.[15,20] ICA problem consists in finding the linear transform A (called the mixing matrix), and the independent source

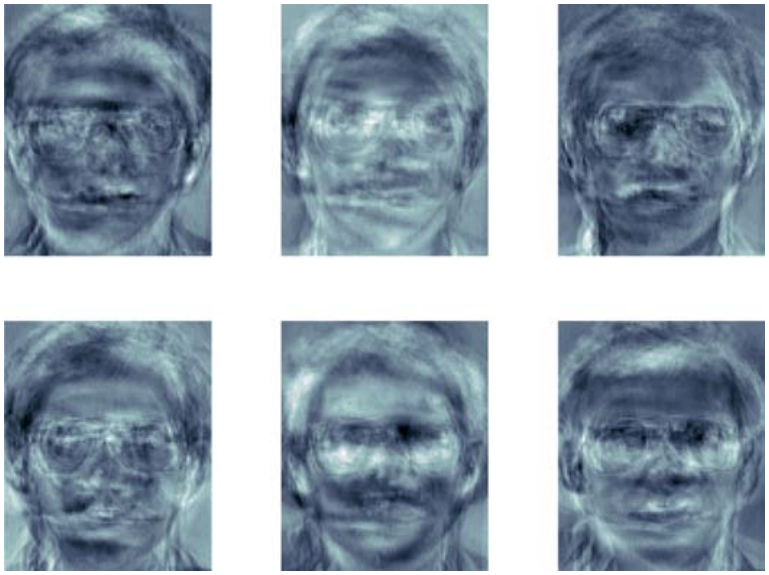


Fig. 3. The first six Fisher's faces for the ORL training database.

components S, such as

$$X^{\mathrm{T}} = AS,$$

where $X \in \mathcal{M}_{N\text{pixels}\times N}$ contains the training images, and $A \in \mathcal{M}_{N\times n}$. Otherwise said, ICA aims at finding the *unmixing matrix* $W \in \mathcal{M}_{n\times N\text{pixels}}$ that gives the best approximation of S: $Y = WX^{\mathrm{T}}$. Preprocessing of X^{T} includes *whitening*, that is, X^{T} is linearly transformed to $\widetilde{X^{\mathrm{T}}} = MX^{\mathrm{T}}$ with uncorrelated components and unit variance, achieved by PCA. At the same time dimensionality is reduced to the number of independent PCA components. This transformation also reduces noise. The problem of finding the matrix A is reduced to finding an orthogonal matrix B such as:

$$\widetilde{X^{\mathrm{T}}} = MX^{\mathrm{T}} = MAS = BS \Rightarrow S = B^{\mathrm{T}}\widetilde{X^{\mathrm{T}}}.$$

ICA problem is nonlinear and it is solved iteratively by *maximization of the non-Gaussianity* of the source components S.

The fast ICA algorithm used in this paper,[15] consists in finding the linear combination of the sphered data, $w^{\mathrm{T}}X^{\mathrm{T}}$, such that has maximum or minimum kurtosis. Feature vectors are determined by projecting the images of the database onto the *nonorthogonal* ICA reduced base. Figure 4 shows the first six ICA faces, i.e. the first six vectors of the projection ICA basis. It can be observed that in this case the ICA faces are more similar to the eigenfaces (Fig. 2) than to the Fisher's faces (Fig. 3) ICA has been used for face recognition by Barlett *et al.*,[2] finding out that ICA provided higher accuracy rates than PCA. Nevertheless ICA is a computationally expensive technique.

Fig. 4. The first six ICA faces for the ORL training database.

3.4. *Methods based on single images*

These methods are applied on any individual image of the database, although they could also be applied to the whole set of training images (as in the PCA case) to find the orthogonal basis to perform the image projection.

3.4.1. *Singular value decomposition (SVD)*

SVD is a very well-known linear algebra technique which is closely related to PCA. It has been applied in face recognition problem by Hsu and Chen,[13] concluding that SVD achieves recognition rates similar to PCA and Fisherfaces. SVD applied to an image I consists in finding orthogonal matrices, U, V, such as

$$I = U\Sigma V^{\mathrm{T}},$$

having Σ diagonal form by blocks. Σ is called the SVD of I, and contains r non-null positive singular values λ_i, where $r = \mathrm{rank}(I)$. U and V orthonormal bases are deduced from the diagonalization of the row and column image correlation matrices II^{T} and $I^{\mathrm{T}}I$. As in the 2DPCA case, SVD works on every individual image of the learning database. Image I can be represented in the SVD feature space by the diagonal matrix Σ_q containing the first q singular values $(\lambda_1, \lambda_2, \ldots, \lambda_q)$ of Σ. Given a new incoming image J, the SVD recognition algorithm projects J onto the orthogonal bases U_k, V_k of the image I_k of the learning database (Fig. 5). This projection will not produce a diagonal matrix for J, but it would be close to it if images J and I are similar. Finally, the algorithm compares both matrices using different p-norms and the cosine criterion induced by the Euclidean scalar product defined over this set of matrices.

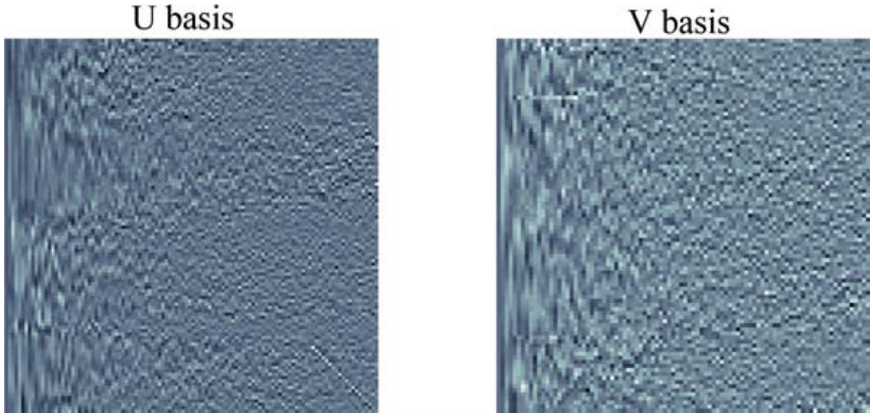


Fig. 5. SVD: U and V basis for a training image of the ORL database.

3.4.2. *Discrete cosine/sine transform (DCT/DST)*

DCT/DST is a covariance-free model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image into cosines/sines of increasing frequency. DCT has been used by Hafed and Levine[11] in face recognition, showing that DCT applied to normalized images is very robust to variations in geometry and lightning. Mathematically, DCT is a discrete Fourier transform operating on real data, by expressing a signal in terms of a sum of sinusoids with different frequencies and amplitudes.

For an image I_k the DCT is defined as:

$$D(u,v) = c(u)c(v) \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I_k(i,j) \cdot \cos \frac{\pi(2i+1)u}{2m} \cos \frac{\pi(2j+1)v}{2n},$$

where $u = 0, \ldots, m-1$ and $v = 0, \ldots, n-1$ and

$$c(\alpha) = \begin{cases} \dfrac{1}{\sqrt{N}}, & \text{if } \alpha = 0, \\ \sqrt{\dfrac{2}{N}}, & \text{if } \alpha \neq 0. \end{cases}$$

N is either the number of rows (m) or columns (n) of the image. DCT can be expressed in matrix form as an orthogonal transformation

$$D = U_{DC} I_k V_{DC}^{\mathrm{T}},$$

where matrices U_{DC} and V_{DC} are orthogonal. This transformation is separable and can be defined in higher dimensions.

The feature vector of an image I_k is constituted by the $q_1 - q_2$ block of D, $D(1 : q_1, 1 : q_2)$, where q_1, q_2 are determined by energy reconstruction considerations

using the Frobenius norm of the image I_k. The energy methodology used to find the q_1, q_2 values is the same in all the spectral methods. Discrete sine transform can also be defined in a similar way using sines instead of cosines.

3.4.3. *Discrete Chebyshev transform (DCHT)*

DCHT is a covariance-free model reduction technique that attempts to decorrelate image by projecting its rows and columns onto Chebyshev polynomials.[24] This transformation is defined as follows:

$$\begin{aligned} \mathrm{dcht}(u,v) &= \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} I(i+1,j+1)t_u(i)t_v(j), \\ u &= 0,1,\ldots,m-1;\ v=0,1,\ldots,n-1, \end{aligned}$$

where $t_u(i)$, $t_v(j)$ are $1-D$ discrete scaled Chebyshev polynomials as defined below. For a given positive integer N (usually the matrix size) and a value x in $[0, N-1]$, the scaled Chebyshev polynomials can be defined recurrently as follows:

$$\begin{aligned} t_0(x) &= \frac{1}{\sqrt{N}}, \\ t_1(x) &= (2x+1-N)\sqrt{\frac{3}{N(N^2-1)}}, \\ t_n(x) &= (\alpha_1 x+\alpha_2)t_{n-1}(x)+\alpha_3 t_{n-2}(x), \\ n &= 2,3,\ldots,N-1,\ \ x=0,1,\ldots N-1, \end{aligned}$$

where

$$\begin{aligned} \alpha_1 &= \frac{2}{n}\sqrt{\frac{4n^2-1}{N^2-n^2}}, \\ \alpha_2 &= \frac{1-N}{n}\sqrt{\frac{4n^2-1}{N^2-n^2}}, \\ \alpha_3 &= \frac{1-n}{n}\sqrt{\frac{2n+1}{2n-3}}\sqrt{\frac{N^2-(n-1)^2}{N^2-n^2}}. \end{aligned}$$

When their degree becomes large, the Chebyshev polynomials tend to exhibit numerical instabilities. For that reason, the following recurrence relationship may be used instead[22]:

$$t_n(0) = -\sqrt{\frac{N-n}{N+n}}\sqrt{\frac{2n+1}{2n-1}}t_{n-1}(0), \quad n=1,\ldots,N-1,$$

$$
\begin{aligned}
t_n(1) &= \left(1 + \frac{n(1+n)}{1-N}\right) t_n(0), \\
t_n(x) &= \gamma_1 t_n(x-1) + \gamma_2 t_n(x-2), \\
n &= 1, 2, \ldots, N-1, \;\; x = 2, 3, \ldots, N-1,
\end{aligned}
$$

with

$$
\begin{aligned}
\gamma_1 &= \frac{-n(n+1) - (2x-1)(x-N-1) - x}{x(N-x)}, \\
\gamma_2 &= \frac{(x-1)(x-N-1)}{x(N-x)}.
\end{aligned}
$$

The model reduction may be performed by constructing two orthogonal matrices, U_m and V_n, such as: $D = U_m I V_n$, where the ith row of the matrices U_m and V_n are the Chebyshev moments of the $i-$degree Chebyshev polynomial. As mentioned previously, the Chebyshev polynomials of degree higher than 50 are very unstable. In such cases, the reduced base may be calculated by transforming the image by blocks of size less than 50×50 pixels. As the DCT, this transformation is separable and can be defined in higher dimensions. The feature vector of an image I_k is calculated the same way as in the DCT case. Although DCHT has already been used for image processing,[22] up to our knowledge it has never been used in face recognition.

3.4.4. *Discrete Walsh-Hadamard transform (DWHT)*

The Hadamard transform is the projection of a signal onto a set of square waves called Walsh functions. The Walsh functions are real and only take the values $+1$ and -1. In the discrete case, the Walsh functions (also called Hadamard matrices) can be constructed recursively as follows:

$$
\begin{aligned}
H_0 &= 1, \\
H_m &= \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix}, \quad m > 0.
\end{aligned}
$$

H_m is called the Hadamard matrix of order m and is a square matrix which has 2^m rows and columns. The Hadamard matrix has the following orthogonality property

$$H_m H_m^T = 2^m I_{2^m}.$$

The Hadamard matrix can also be obtained by defining its element in the ith row and jth $(i, j = 0, 1, \ldots, 2^n - 1)$ column, as follows:

$$H_n(i,j) = (-1)^{\sum_{k=0}^{n-1} i_k j_k} = \prod_{k=0}^{n-1} (-1)^{i_k j_k},$$

where

$$i = \sum_{k=0}^{n-1} i_k 2^k = (i_{n-1} i_{n-2} \ldots i_1 i_0)(i_k = 0, 1),$$

$$j = \sum_{k=0}^{n-1} j_k 2^k = (j_{n-1} j_{n-2} \ldots j_1 j_0)(j_k = 0, 1),$$

are the binary representations of i and j, respectively.[3]

The discrete Hadamard transform of a one-dimensional signal $X = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$ is

$$\mathrm{DWHT}(X) = H_n X,$$

with $n = \log_2 N$. Furthermore,

$$\mathrm{DWHT}(X)(j) = \sum_{i=0}^{N-1} x_{i+1} H_n(i, j) = \sum_{i=0}^{N-1} x_{i+1} (-1)^{\sum_{k=0}^{n-1} i_k j_k},$$

where i and j are represented in binary form.

Similarly, the DWHT of an image I with $M \times N$ pixels is defined as follows:

$$\begin{aligned} \mathrm{DWHT}(u, v) &= \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} H_m(u, i) I(i+1, j+1) H_n(v, j) \\ &= \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} I(i+1, j+1) (-1)^{\sum_{k=0}^{m-1} u_k i_k + \sum_{l=0}^{n-1} v_l j_l}, \end{aligned}$$

where u, v, i and j are represented in binary form.[30] Furthermore, the DWHT can be written in matrix form as follows:

$$\mathrm{DWHT}(I) = H_m I H_n^{\mathrm{T}},$$

with $m = \log_2 M$ and $n = \log_2 N$.

The DWHT performs an orthogonal, symmetric transform, that computationally is very fast and simple. Projecting an image onto the Walsh-Hadamard basis functions only involves multiplication of each pixel by ± 1.[12] The DWHT has been used in data compression.[1] DWHT has been applied to face recognition problems by Sadykhor *et al.*[32]

3.4.5. *Discrete Hilbert transform (DHT)*

The Hilbert transform of a 1D signal $X(t)$ is defined as the convolution of $X(t)$ with the Hilbert Kernel, $h(t) = 1/(\pi t)$:

$$H(t) = X(t) \star h(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{X(\xi)}{t - \xi} d\xi.$$

Because the kernel function $h(t)$ is not integrable, the integral defining the convolution does not converge. Instead, the Hilbert transform is defined using the integral Cauchy principal value:

$$H(t) = \frac{1}{\pi} \lim_{b \to \infty} \int_{-b}^{b} \frac{X(\xi}{t - \xi} d\xi.$$

In any case, the DHT of a discrete signal $X(n)$, where $n \in \mathbb{Z}$ was defined by Kak in Ref. 17 as follows:

$$\mathrm{DHT}(X(k)) = \begin{cases} \dfrac{2}{\pi} \displaystyle\sum_{n\,\mathrm{odd}} \frac{X(n)}{k-n}, & \text{if } k \text{ even}, \\ \dfrac{2}{\pi} \displaystyle\sum_{n\,\mathrm{even}} \frac{X(n)}{k-n}, & \text{if } k \text{ odd}. \end{cases}$$

The discrete version for 2D digital images is described in Ref. 31. DHT has been used in face recognition in Ref. 26.

The Hilbert transform has also been combined with the Huang transform,[14] which is an empirically-based data analysis technique to decompose digital data into oscillating intrinsic components, also called intrinsic mode functions (IMF). The Huang Hilbert transform (HHT) can be used to represent nonlinear and non-stationary processes. HHT has been applied to face recognition by Zang *et al.*[39] This method is not used in this paper because it is computationally very expensive. Instead, we use the Hilbert transform alone.

3.4.6. *Discrete wavelet transform (DWT)*

Wavelets are compact functions (defined over a finite interval) with zero mean and some regularity conditions (vanishing moments). The Wavelet transform converts a function into a linear combination of basic functions, called wavelets, obtained from a prototype wavelet through dilatations, contractions and translations. The DWT of an image $I \in \mathcal{M}(m, n)$ allows finding two orthogonal transformations based on wavelets, named U_w and V_w, such as

$$I = U_w D_{WT} V_w^{\mathrm{T}}.$$

These orthogonal matrices can be constructed as follows:

$$U_W = \left[\frac{H}{G}\right]_m^{\mathrm{T}}, \quad V_W = \left[\frac{H}{G}\right]_n^{\mathrm{T}}. \tag{6}$$

H represents a low pass or averaging portion of the matrices U_W^T, V_W^T, and G is the high pass or differencing portion. In all of cases we have

$$D_{WT} = \begin{pmatrix} HIH^{\mathrm{T}} & HIG^{\mathrm{T}} \\ GIH^{\mathrm{T}} & GIG^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} B & V \\ H & D \end{pmatrix},$$

Table 1. Spectral methods and their corresponding projection basis.

Method	Covariance	Projection basis
PCA/NPCA	$S = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^{\mathrm{T}}$	Eigenfaces
2DPCA	$S_M = \frac{1}{N} \sum_{i=1}^{N} (I_i - \bar{I})^{\mathrm{T}} (I_i - \bar{I})$	2DPCA basis
Fisher's LDA	$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$	
	$S_W = \sum_{i=1}^{C} \sum_{X_k \in C_i} (X_k - \mu_i)(X_k - \mu_i)^{\mathrm{T}}$	Fisherfaces
ICA	Covariance-free	
	maximization of non-Gaussianity	ICA basis
SVD	Covariance-based	U, V basis
DCT/DST	Covariance-free	Fourier basis
DCHT	Covariance-free	Chebyshev Polynomials
DWHT	Covariance-free	Walsh-Hadamard basis
DHT	Covariance-free	Hilbert basis
DWT	Covariance-free	Wavelet basis

where B is the blur, V are the vertical differences, H are the horizontal differences and D are the diagonal differences. DWT can be applied several times to further reduce the dimension of the attribute vector. In the present case, we apply DWT twice and we use the blur B as feature attribute to solve the face recognition problem. Different kind of filters can be used for this purpose.[8] These families of wavelets define a DWT having a maximum number of vanishing moments. We have used the Daubechies-2 family. DWT was applied to face recognition by Kakarwal and Deshmukh.[18]

Table 1 summarizes all the spectral methods used in this paper, and their respective projection basis.

4. Numerical Results

To perform the numerical analysis, we have used the ORL database of faces provided by AT&T Laboratories Cambridge and the FERET database. The ORL database contains 10 different images of 40 distinct individuals. All the images were taken against a dark homogeneous background, varying the lighting, facial expressions and facial details. The database provides upright and frontal poses. The size of each image is 92×112 pixels, with 256 gray levels per pixel.

Although the aim of this paper is not to analyze the performance of the spectral methods for different types of image complexities, such as scale and pose variations, lighting conditions, etc., we also provide some preliminary results for the FERET database. Furthermore, in both cases no alignment preprocessing such as translation, rotation and scaling based on the eyes coordinates has been adopted, since registration and dimensionality reduction are two different problems and should be approached independently.

In all the experiments over ORL, the learning database is composed of five poses of each individual, that are randomly selected. The rest of the poses in the database

are used as probe images for establishing the accuracy of the classification for each spectral technique, using both, global and local features. For each attribute, the classification is performed 100 different times, randomly choosing the learning database and the set of probe images (200 images). Nevertheless, once the database has been generated, it is used to perform classification with all the different spectral methods under the same numerical conditions. For instance, in all the cases the energy cutoff for the reconstruction has been fixed to 99.5% of the Frobenius norm of the transform.

Two different kind of analysis are performed using global and local attributes. In the case of the global analysis, the attributes are calculated over the whole size of the images in the learning database. In the case of the local features, this analysis is performed by dividing all images into blocks. For each block, the attributes are computed, and the final feature vector is defined by merging in an unique vector all the local attribute vectors of the image, always computed on the same order. Figure 6 shows the sketch of this process. Although the use of local features increases the dimension of the attribute space, it is expected to increase its discriminative power. Thus, in the algorithm presented in this paper, no individual classification with posterior fusion of scores is performed. Finally, we provide statistics for the classification calculated over 100 different simulations of each classifier (attribute): minimum, maximum, median and mean accuracy, and also interquartile range and standard deviation, for the distance computed using different norms and the cosine criteria shown in Sec. 2. These analyses were also performed on the FERET database with two frontal poses in the training set, and one pose for testing for each person, as it is usually recommended. This set up only allowed for one simulation on the FERET database.

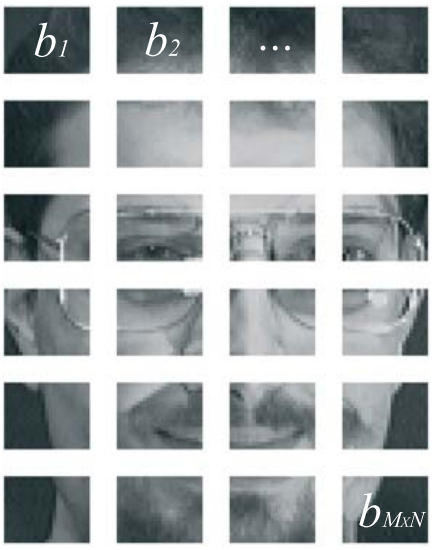


Fig. 6. Image partition in blocks to performing local analysis.

4.1. *Application to the ORL database*

4.1.1. *Global analysis*

The results of the global analysis are summarized in Table 2 for the distance criterion and in Table 3 for the cosine criterion. The conclusions of the global analysis are the following:

(1) For the methods based on the whole set of images:

- 2DPCA provides the higher maximum accuracy rate (99%) and the higher median accuracy (94.50%), using in both cases the distance criterion. 2DPCA also shows the lower accuracy dispersion.
- ICA with cosine provides higher accuracies than ICA with Euclidean distance, but both are lower than PCA/NPCA.
- PCA shows the same accuracy for both criteria.
- Fisherfaces has the lower maximum and median accuracies and the higher dispersions.
- Finally, noncentered PCA (NPCA) performs better than PCA, although the results are very similar.

(2) For the methods based on orthogonal transformations of individual images (SVD, DCT/DST, DCHT, DWHT, DHT and DWT):

- The DCHT, which is introduced in this paper for the face recognition problem, and the DWT provide the higher accuracy rates. The higher median accuracy (95.5%) and the lower dispersion (1.46%) are provided by the DWT.
- DCT, DST, DWHT and DHT are also free covariance spectral methods that are computationally not expensive. Interestingly, the DST performs better

Table 2. Statistics for the global analysis with the Euclidean distance criterion, over 100 independent simulations. Bold faces represent the highest minimum, maximum, median and mean accuracies, and the lowest dispersion.

%	Min	Max	Median	Mean	IQR	Std
PCA	88.50	97.50	93.50	93.30	2.50	1.89
NCPCA	90.50	98.00	94.00	94.11	2.5	1.68
ICA	85.50	94.50	90.50	90.21	3.50	2.20
Fisher	83.50	93.50	88.00	88.13	2.75	2.12
2DPCA	90.50	**99.00**	94.50	94.64	2.50	1.73
SVD1	78.00	89.00	84.00	83.80	2.75	2.38
SVD2	87.00	95.50	91.50	91.54	2.50	1.82
DCT	87.50	96.00	92.00	92.02	**2.00**	1.81
DST	90.00	97.00	93.50	93.58	**2.00**	1.54
DCHT	91.50	98.50	95.00	94.92	**2.00**	1.52
DWHT	90.00	97.00	94.00	93.87	**2.00**	1.65
DHT	90.00	97.50	94.00	93.87	2.25	1.66
DWT	**92.50**	98.50	**95.50**	**95.50**	**2.00**	**1.46**

Table 3. Statistics for the global analysis with the cosine criterion.

%	Min	Max	Median	Mean	IQR	Std
PCA	88.50	97.50	93.50	93.52	**2.00**	1.76
NCPCA	87.50	95.50	92.00	92.00	2.50	1.74
ICA	87.00	96.00	91.50	91.57	2.50	1.95
Fisher	81.00	92.00	86.25	86.23	3.25	2.25
2DPCA	89.00	98.00	93.75	93.77	2.25	**1.52**
SVD1	52.50	77.00	64.25	64.64	6.00	4.53
SVD2	81.50	91.00	86.50	86.33	2.50	2.00
DCT	83.00	93.50	88.00	87.81	2.50	2.04
DST	85.50	95.00	90.50	90.46	2.75	1.89
DCHT	**90.50**	**98.00**	**94.50**	**94.19**	2.50	1.70
DWHT	87.00	95.50	91.50	91.38	2.50	1.73
DHT	88.00	96.00	92.00	91.91	2.50	1.68
DWT	90.00	97.50	94.00	94.03	**2.00**	1.56

than the DCT. Finally, the DHT and the DWHT also provide very good accuracy rates.

- The SVD provides the lower accuracy rate for this energy cutoff (99.5%). If the energy cutoff is increased to 99.99%, then the SVD improves notably, providing similar accuracies to other methods. Nevertheless, SVD is a covariance-based method that needs two diagonalizations. Due to this fact it is computationally infeasible for large images.
- In all these methods the distance criterion is superior to the cosine criterion.

(3) Finally, the use of other p-norms ($p = 3$ and $p = 5$) provides better recognition rates for DCT/DST, SVD, Fisher's LDA and ICA. For the rest of methods the results are slightly worse. The use of the infinite norm always produce lower accuracy rates. Results are shown in Table 4.

Table 4. Statistics for the global analysis using different p-norms. Bold faces represent the highest minimum, maximum, median and mean accuracies, and the lowest dispersion.

%	Norm	Min	Max	Median	Mean	IQR	Std
PCA	3	90.00	94.00	92.00	92.25	2.00	1.25
	5	89.50	95.50	93.50	93.05	2.50	1.81
	∞	84.50	94.00	88.75	88.80	4.00	2.75
NCPCA	3	90.00	93.50	92.00	92.05	1.50	1.23
	5	89.50	95.50	93.50	92.65	3.50	2.06
	∞	87.00	94.00	90.50	90.65	3.50	2.16
ICA	3	86.00	90.50	88.75	88.20	3.00	1.67
	5	83.50	95.50	89.50	89.51	4.00	2.32
	∞	85.00	91.00	88.75	88.45	2.50	1.78
Fisher	3	82.50	90.50	87.50	87.05	5.00	2.94
	5	89.00	95.50	93.50	92.60	3.00	1.98
	∞	80.50	88.00	84.25	84.00	3.50	2.39

Table 4. (*Continued*)

%	Norm	Min	Max	Median	Mean	IQR	Std
2DPCA	3	89.50	94.50	92.50	92.20	1.50	1.43
	5	89.00	95.50	93.50	92.60	3.00	1.98
	∞	77.00	85.00	83.00	82.40	2.50	2.34
SVD	3	87.00	92.00	89.25	89.20	3.00	1.56
	5	86.50	93.00	89.75	89.45	2.50	2.00
	∞	81.00	91.00	87.50	86.65	3.00	3.07
DCT	3	91.50	**96.50**	**94.75**	94.60	2.50	1.55
	5	90.50	96.00	93.75	93.35	**1.00**	1.51
	∞	89.00	94.00	92.50	91.80	3.00	1.75
DST	3	90.50	96.00	92.75	92.85	2.00	1.65
	5	86.50	93.00	90.50	90.45	1.50	2.00
	∞	84.00	91.00	88.25	87.60	2.50	2.01
DCHT	3	**92.50**	96.00	94.50	94.45	2.00	1.21
	5	91.50	96.00	**94.75**	94.45	1.50	1.30
	∞	90.00	96.00	92.75	93.05	3.50	2.02
DWHT	3	**92.50**	**96.50**	95.00	**94.70**	1.50	1.13
	5	89.50	96.50	93.00	92.95	2.00	1.78
	∞	87.50	92.50	90.50	90.15	4.00	1.93
DHT	3	90.50	94.00	92.00	92.30	1.50	**1.11**
	5	83.00	89.00	86.25	86.20	2.50	2.07
	∞	41.50	46.50	44.25	44.15	2.50	1.66
DWT	3	92.00	**96.50**	94.50	94.55	2.50	1.48
	5	88.00	93.50	90.50	90.70	2.50	1.78
	∞	62.00	74.50	70.75	69.60	3.50	3.38

The results of the global analysis are graphically shown for the different statistics (accuracy rates and dispersions) in Figs. 7–10.

4.1.2. *Local analysis*

To perform the local analysis, each image has been divided in this case into 8×4 blocks. The results of the local analysis are summarized in Tables 5 and 6 for both criteria using the Euclidean norm. Table 7 shows the same results using other p-norms. The conclusions of the local analysis are:

(1) For the methods based on the whole set of images:

- PCA, Fisher's LDA and 2DPCA provide the higher maximum (98%) and median (95%) accuracies. Fisher's LDA as a supervised method is expected to provide higher accuracy rates if the number of poses in the training data increases. Nevertheless, this fact is also true for the rest of the methods, since the number of poses in the learning database is a crucial parameter in the classification.
- All of these methods, but ICA and 2DPCA, equal or improve the results shown for the global statistics, meaning that local analysis captures better

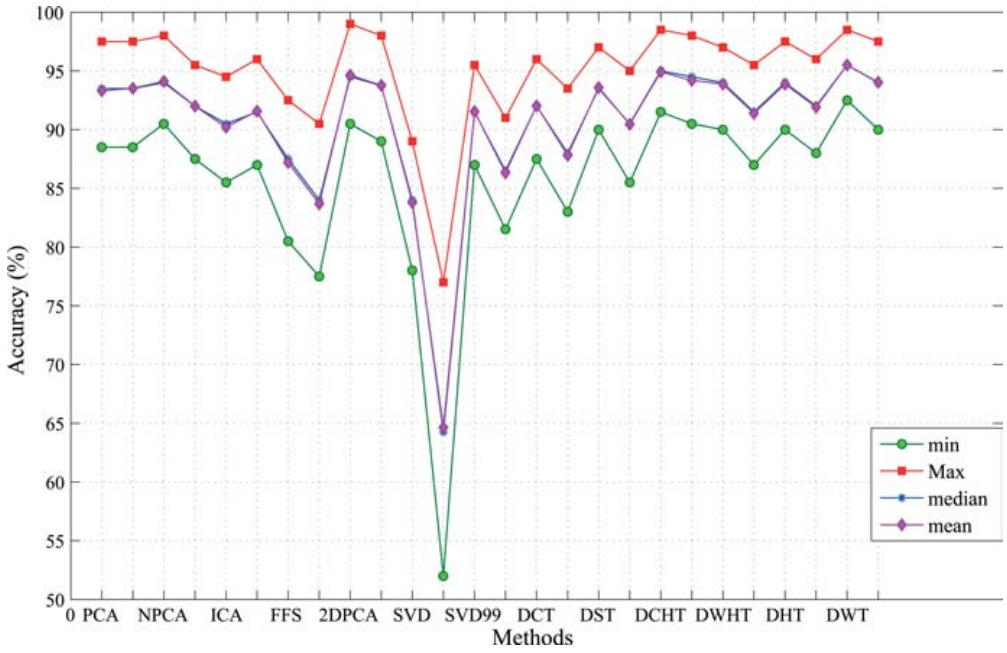


Fig. 7. Global analysis: minimum, maximum, median and mean accuracies for each spectral method, calculated over 100 different simulations.

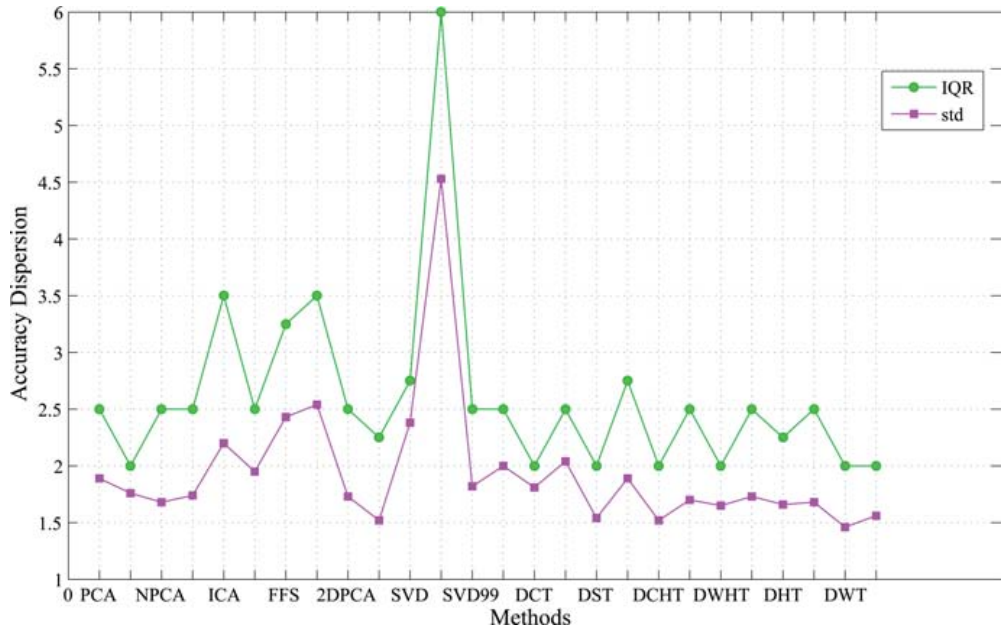


Fig. 8. Global analysis: standard deviation and interquartile range for each spectral method, calculated over 100 different simulations.

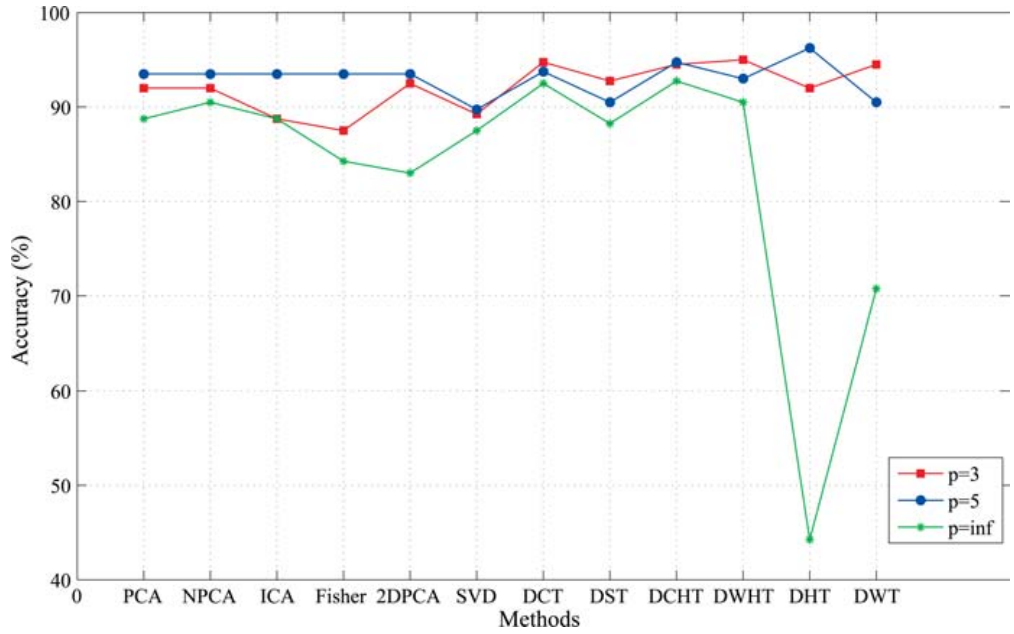


Fig. 9. Global analysis: median accuracies for each spectral method, calculated with different p-norms.

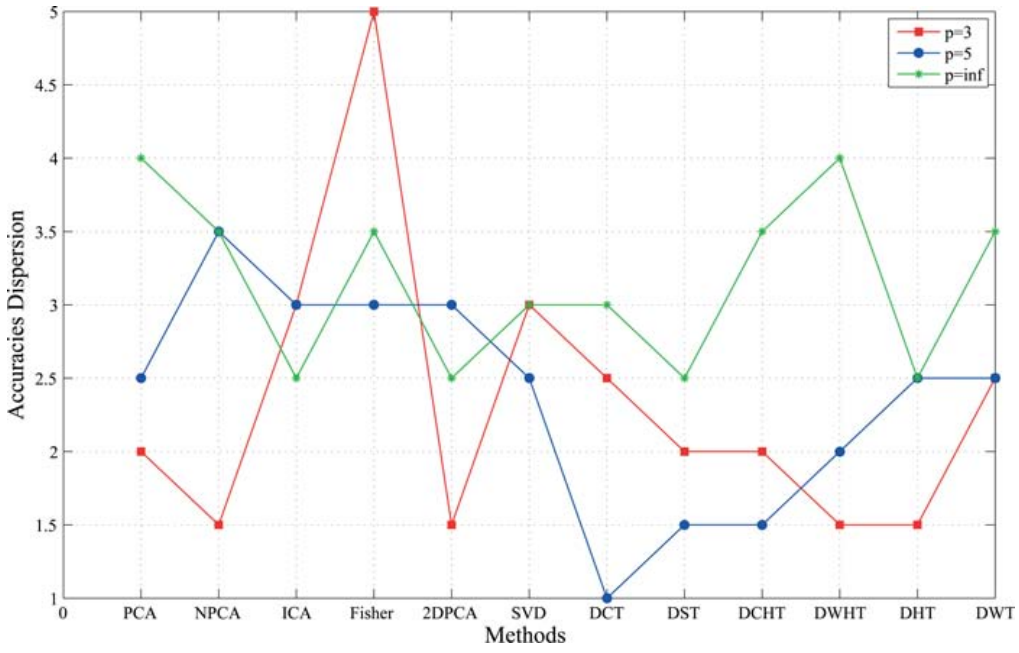


Fig. 10. Global analysis: interquartile range for each spectral method, calculated with different p-norms.

the face variations than the global analysis. The dispersions and the minimum accuracies also improve. This idea has been also presented by Oh *et al.*[25]

- All these methods show similar medium accuracies ($\sim 94.5\%$), but ICA has the lower rate (88%) and is the most expensive method.

Table 5. Statistics for the local analysis with the Euclidean distance criterion, over 100 independent simulations. Bold faces represent the highest minimum, maximum, median and mean accuracies, and the lowest dispersion.

%	Min	Max	Median	Mean	IQR	Std
PCA	**91.00**	**98.00**	**95.00**	**94.83**	**2.00**	**1.60**
NCPCA	90.00	97.50	94.50	94.49	2.50	1.66
ICA	83.50	93.50	88.00	88.19	4.50	2.72
Fisher	90.50	**98.00**	**95.00**	94.96	**2.00**	1.62
2DPCA	90.00	**98.00**	**95.00**	94.73	2.50	1.69
SVD1	87.00	97.00	93.00	92.78	2.50	1.95
SVD2	90.50	97.50	94.50	94.39	2.25	**1.60**
DCT	89.00	97.50	93.25	93.33	2.50	1.75
DST	90.00	97.50	94.50	94.40	**2.00**	1.64
DCHT	89.00	96.50	93.50	93.53	2.50	1.74
DWHT	90.50	97.50	94.50	94.49	2.50	1.62
DHT	87.50	96.00	92.50	92.33	2.50	1.74
DWT	88.50	97.00	93.50	93.47	2.75	1.73

Table 6. Statistics for the local analysis with the cosine criterion.

%	Min	Max	Median	Mean	IQR	Std
PCA	**91.50**	**98.00**	**95.00**	**94.82**	2.50	**1.52**
NCPCA	88.00	96.50	92.50	92.46	**2.00**	1.74
ICA	83.00	93.00	87.50	87.79	4.12	2.56
Fisher	88.00	97.50	93.50	93.58	**2.00**	1.74
2DPCA	89.00	97.50	93.00	93.18	2.50	1.63
SVD1	83.00	94.00	88.50	88.35	3.00	2.28
SVD2	87.50	96.00	92.00	92.03	2.25	1.72
DCT	84.00	95.00	89.50	89.35	2.50	2.10
DST	87.50	95.50	92.00	91.95	2.25	1.74
DCHT	86.50	95.00	90.50	90.68	2.50	1.94
DWHT	87.50	96.00	92.00	92.04	2.50	1.75
DHT	82.50	94.00	89.50	89.15	2.75	2.03
DWT	87.00	95.00	91.50	91.54	3.00	1.80

- Interestingly, PCA performs better using the cosine criterion. In the rest of the cases the distance criterion is superior. Nevertheless, the difference between both criteria are smaller than in the global analysis.

(2) For the methods based on orthogonal transformations of individual images, their performance is more or less the same. The median accuracies are lower than for the global analysis; only SVD performs much better.

(3) Finally, the median accuracy decreases as the index p of the norm increases. Only NPCA, DCT, DCHT, and DWHT improve their accuracy with respect to the Euclidean norm.

The results of the local analysis are graphically shown in Figs. 11–14.

Table 7. Statistics for the local analysis using different p-norms. Bold faces represent the highest minimum, maximum, median and mean accuracies, and the lowest dispersion.

%	Norm	Min	Max	Median	Mean	IQR	Std
PCA	3	91.50	95.00	92.75	93.00	**1.00**	**1.15**
	5	90.00	95.00	93.50	92.95	3.50	1.81
	∞	87.00	91.00	89.75	89.25	2.50	1.53
NCPCA	3	91.50	**97.50**	95.00	94.75	2.00	1.63
	5	90.50	95.00	93.50	93.05	3.50	1.64
	∞	85.00	92.50	89.25	89.15	2.50	2.23
ICA	3	77.00	85.50	82.75	82.30	4.00	2.65
	5	73.00	84.00	78.00	78.35	5.50	3.35
	∞	70.50	78.00	74.00	74.20	4.50	2.65
Fisher	3	90.00	96.00	94.00	93.85	2.50	1.81
	5	90.50	95.00	93.50	93.00	3.50	1.64
	∞	80.50	88.50	84.50	84.60	3.50	2.55
2DPCA	3	90.50	96.50	93.25	93.55	1.50	1.60
	5	90.50	95.00	93.50	93.00	3.50	1.64
	∞	73.50	81.50	79.00	78.20	2.00	2.47
SVD	3	90.50	95.50	93.00	93.30	2.00	1.56
	5	89.50	94.00	92.00	92.15	2.00	1.43
	∞	86.50	92.50	88.50	88.95	3.50	1.97
DCT	3	90.50	**97.50**	**95.25**	94.55	2.50	2.02
	5	90.00	95.50	93.75	93.10	3.00	1.85
	∞	87.00	91.50	88.75	88.85	3.00	1.74
DST	3	**92.00**	96.00	94.00	94.00	2.00	1.43
	5	90.50	96.00	91.75	92.15	1.50	1.63
	∞	84.00	90.00	87.00	86.95	1.50	1.89
DCHT	3	91.50	96.50	94.50	94.35	2.50	1.63
	5	91.50	95.00	93.00	93.10	2.50	1.39
	∞	87.50	91.00	90.00	89.50	2.50	1.37
DWHT	3	91.50	97.00	95.00	**94.70**	1.50	1.56
	5	91.00	95.50	93.00	93.15	2.50	1.47
	∞	86.50	93.00	89.75	89.80	2.50	1.98
DHT	3	87.00	94.00	89.00	89.55	3.00	2.17
	5	76.50	82.50	78.75	79.25	1.50	1.85
	∞	31.50	37.00	34.75	34.35	3.00	1.94
DWT	3	89.00	95.50	93.00	92.70	4.00	2.21
	5	85.00	90.50	89.50	88.70	4.00	2.07
	∞	65.00	75.50	71.25	70.85	6.50	3.63

Finally, Table 8 shows a summary regarding the best accuracy results obtained for each spectral method, together with the type of analysis (local or global), the criterion (p-norm or cosine), and the computational cost in a Pentium(R) 3.4 GHz and 1GB of RAM. The wavelet transform (DWT) is the method that provides the higher accuracy with the lower computational cost. Interesting to note, all the methods provided the best results with the p-norm criterion, but ICA.

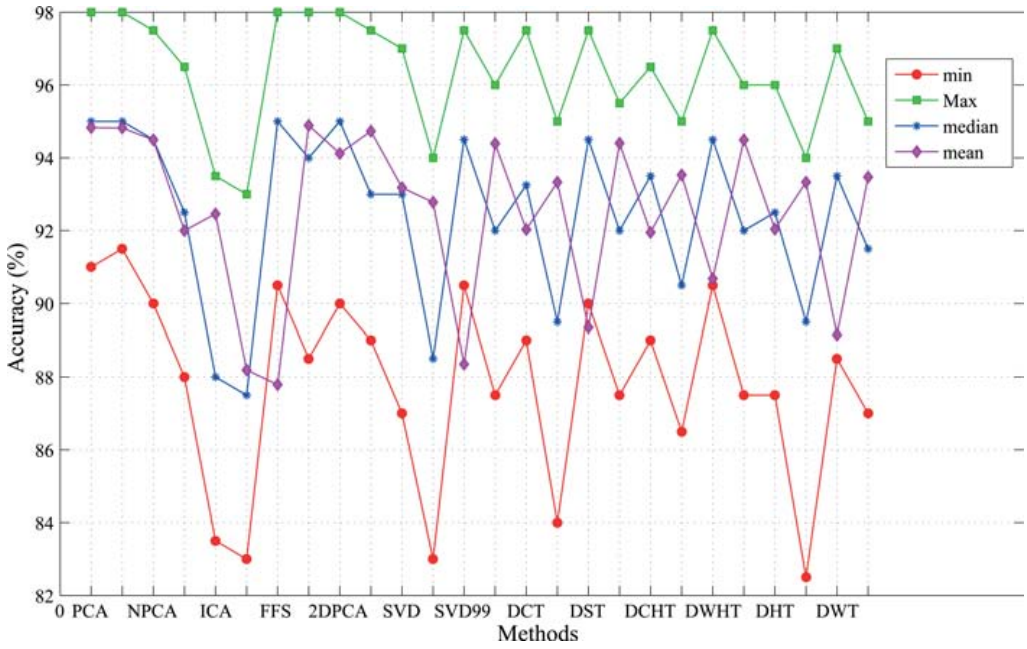


Fig. 11. Local analysis: minimum, maximum, median and mean accuracies for each spectral method, calculated over 100 different simulations.

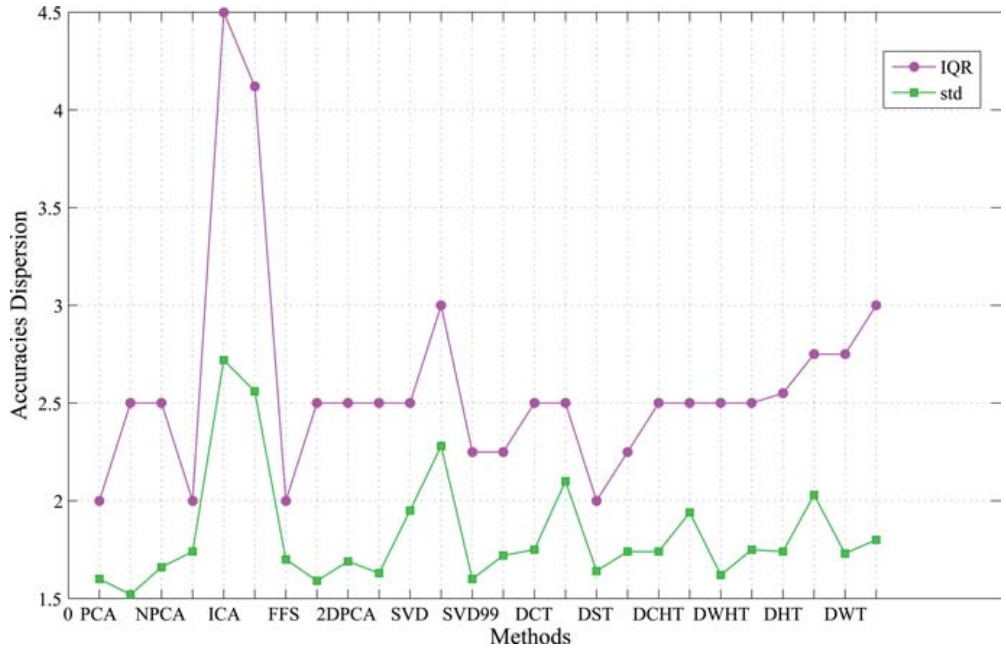


Fig. 12. Local analysis: standard deviation and interquartile range for each spectral method, calculated over 100 different simulations.

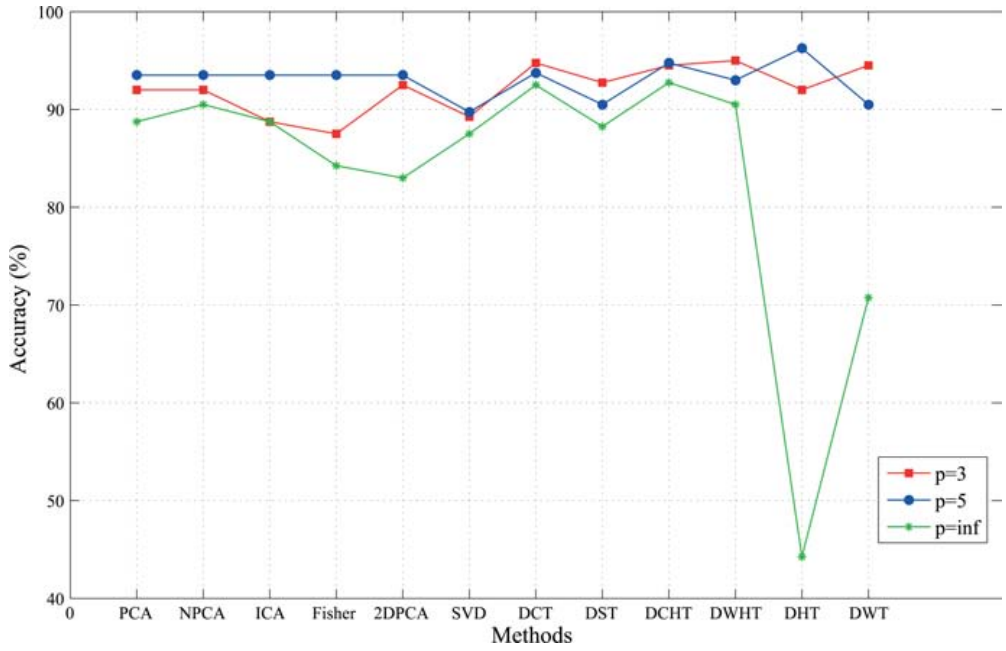


Fig. 13. Local analysis: median accuracies for each spectral method, calculated with different p-norms.

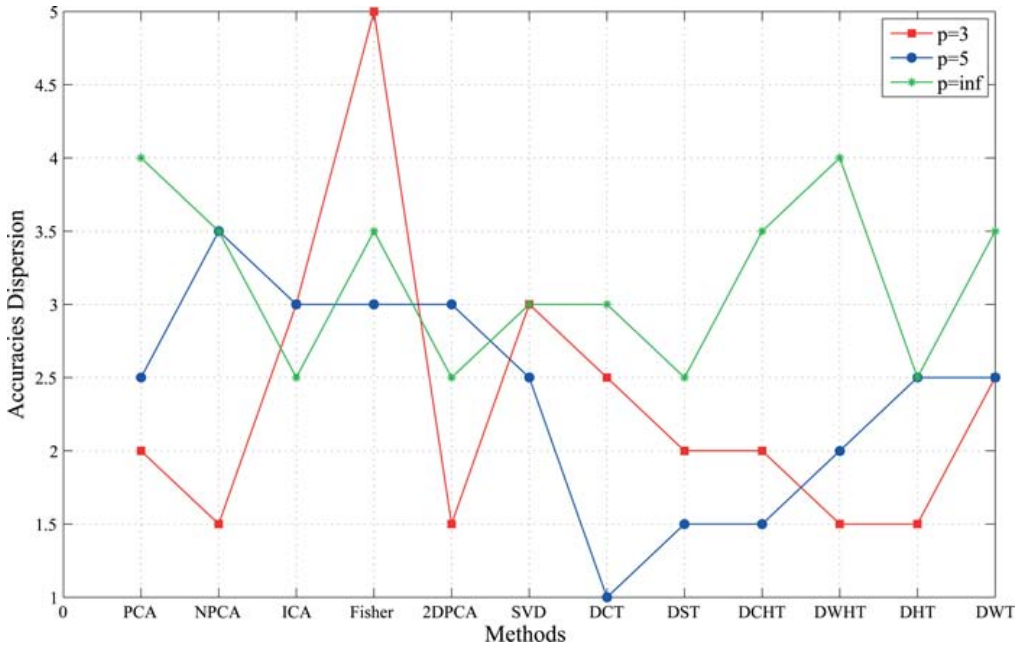


Fig. 14. Local analysis: interquartile range for each spectral method, calculated with different p-norms.

4.2. *Application to the FERET database*

Similar tendencies were also observed for the FERET database, although the recognition accuracies were lower: the maximum accuracy rate obtained for all the methods was 77% (see Table 9).

Table 8. Statistics for the local analysis with the cosine criterion.

Method	Analysis	Criterion	Median	Cost
PCA	Local	$p = 2$	95.00	32.07
NCPCA	Local	$p = 2$	94.50	34.65
ICA	Global	cosine	91.50	62.96
Fisher	Local	$p = 2$	95.00	36.94
2DPCA	Local	$p = 2$	95.00	42.63
SVD	Local	$p = 2$	93.00	112.57
DCT	Local	$p = 3$	95.25	70.73
DST	Local	$p = 2$	94.50	68.90
DCHT	Global	$p = 2$	95.00	20.90
DWHT	Global	$p = 3$	95.00	51.41
DHT	Global	$p = 2$	94.00	37.48
DWT	Global	$p = 2$	95.50	12.15

Table 9. FERET database: recognition rates for the global and local analysis using all the spectral methods. Bold faces represent the highest accuracy rates.

Method	Criterion	Global	Local
PCA	dist	74	74
	cos	75	**77**
NCPCA	dist	74	76
	cos	76	**77**
ICA	dist	65	63
	cos	75	65
2DPCA	dist	77	75
	cos	**78**	76
SVD	dist	60	71
	cos	59	70
DCT	dist	74	74
	cos	73	74
DST	dist	77	76
	cos	77	76
DCHT	dist	70	73
	cos	66	71
DWHT	dist	77	76
	cos	77	76
DHT	dist	76	71
	cos	76	73
DWT	dist	75	75
	cos	76	73

For the methods based on the whole set of images, the cosine criterion seems to provide better recognition rates. The highest accuracy is obtained by 2DPCA with global features. It is not as explicit as for the ORL database, that the use of local features provides better results. In this case Fisher's LDA could not be applied due to the low number of available poses in each class.

For the methods based on single images, global features seem to provide better results than local features, as it happens for the ORL database. In this case, the best methods are the DST and the DWHT. The DCHT does not provide as good results as for the ORL database.

Other authors have achieved higher accuracies (around 93%) using PCA and Fisher's LDA[29] by applying a preprocessing of the FERET images including: translation, rotation, scaling, eyes centering, etc. Nevertheless, as it was already mentioned, this was not the main objective of this research.

5. Conclusion

In this paper, we have compared different spectral methods applied to face recognition problem by nonsupervised classification, using the same database and numerical parameters, such as the number of training and testing images, the energy reconstruction cutoff, the same number of blocks for the local analysis, etc. Global and local analysis were performed over the ORL and FERET databases.

The use of the DCHT and the NPCA is introduced in this paper for biometric applications. Methods based on the whole set of images usually perform better for local analysis, while methods based on orthogonal transformations on single images usually performed better for global analysis. Furthermore, local analysis allows the application of covariance-based methods (PCA, NPCA, Fisher' LDA, 2DPCA) to very large images. PCA, NPCA, 2DPCA and DWT seem to be the best performing spectral methods. All the free covariance techniques DCT, DST, DCHT, DWHT and DHT, also provide very high accuracies and are computationally less expensive.

One of the main conclusions derived from this analysis is that all the spectral methods perform more or less similarly. The reason why this happens is that orthogonal transformations are isometries, and all the transformed images (matrices) are related by similarity relationships of the kind $I = UDV^{\mathrm{T}}$, where U and V are orthogonal matrices. Thus, all the similar matrices (I and D) contain the same information if no model reduction is performed. This is not the case in practice because the dimension of the attribute vector and also the accuracy of the classification depend on the energy threshold that is used. In our opinion the different accuracies shown in the literature for all these methods are due to the use of different energy cutoffs for dimensionality reduction, and also to the different parameters used in the face recognition experiments (databases), that is, no unique computing platform was previously used to perform the numerical comparison of all these different methods. This is not the case of the present paper where all the spectral methods have been compared using the same numerical conditions. In all the cases the crucial

point is choosing the right energy cutoff for model reduction. By tuning this parameter correctly all these methods should perform similarly. This fact reinforces the use of supervised methods to learn this crucial parameter.

Finally, and even more importantly, independent of the existence of different complexities and scale variations in the database that might complicate the face recognition analysis, no spectral method is able to provide systematically 100% accuracy rate operating stand alone. Thus, other types of image attributes are needed to improve the accuracy of the face recognition using for instance ensemble learning techniques.[28]

References

1. A. Akansu and R. Poluri, Walsh-like nonlinear phase orthogonal codes for direct sequence cdma communications, *IEEE Trans. Signal Process.* **55**(7) (2007) 3800–3806.
2. M. Bartlett, J. Movellan and T. Sejnowski, Face recognition by independent component analysis, *IEEE Trans. Neural Netw.* **13**(6) (2002) 1450–1464.
3. K. Beauchamp, *Applications of Walsh and Related Functions, With an Introduction to Sequency Theory*, Microelectronics and signal processing (Academic Press, 1984).
4. P. N. Belhumeur, J. Hespanha and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, in *Computer Vision ECCV'96*, eds. B. Buxton and R. Cipolla, Lecture Notes in Computer Science, Vol. 1064 (Springer, Berlin, Heidelberg, 1996), pp. 43–58.
5. R. Chellappa, C. Wilson and S. Sirohey, Human and machine recognition of faces: A survey, *Proc. IEEE* **83** (1995) 705–740.
6. L. Chen, A new lda-based face recognition system which can solve the small sample size problem, *Pattern Recog.* **33**(10) (2000) 1713–1726.
7. M. Dabbah, W. Woo and S. Dlay, Secure authentication for face recognition, *IEEE Symp. Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007*, April 2007, pp. 121–126.
8. I. Daubechies, *Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics* (Society for Industrial and Applied Mathematics, 1992).
9. W.-L. Fang, Y.-K. Yang and J.-K. Pan, A low-computation approach for human face recognition, *Int. J. Pattern Recogn. Artif. Intell.* **26**(6) (2012) 1256015-1–1256015-23.
10. R. A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* **7** (1936) 179–188.
11. Z. M. Hafed and M. D. Levine, Face recognition using the discrete cosine transform, *Int. J. Comput. Vis.* **43**(3) (2001) 167–188.
12. M. Hassan, I. Osman and M. Yahia, Walsh-hadamard transform for facial feature extraction in face recognition, *Int. J. Comput. Commun. Eng.* **1**(7) (2007) 436–440.
13. C.-H. Hsu and C.-C. Chen, Svd-based projection for face recognition, *IEEE Int. Conf. Electro/Information Technology*, May 2007, pp. 600–603.
14. N. E. Huang and S. S. P. Shen, *Hilbert-Huang Transform and Its Applications*, Interdisciplinary Mathematical Sciences, Vol. 5 (World Scientific Publishing Co. Pte. Ltd., 2005).
15. A. Hyvärinen and E. Oja, A fast fixed-point algorithm for independent component analysis of complex valued signals, *Int. J. Neural Syst.* **10**(1) (2000) 1–8.
16. R. Jafri and H. R. Arabnia, A survey of face recognition techniques, *J. Inform. Process. Syst.* **5**(2) (2009) 41–68.

17. S. Kak, The discrete Hilbert transform, *Proc. IEEE* **58**(4) (1970) 585–586.
18. S. Kakarwal and R. Deshmukh, Wavelet transform based feature extraction for face recognition, *Int. J. Comput. Sci. Appl.* (2010) 100–104.
19. M. Kirby and L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces, *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(1) (1990) 103–108.
20. D. Langlois, S. Chartier and D. Gosselin, An introduction to independent component analysis: Infomax and fastica algorithms, *Tutor. Quant. Meth. Psychol.* **6**(1) (2010) 31–38.
21. H. Moon and P. J. Phillips, Computational and performance aspects of pca-based face-recognition algorithms, *Perception* **30**(3) (2001) 303–321.
22. R. Mukundan, S. Ong and P. Lee, Image analysis by Tchebichef moments, *IEEE Trans. Image Process.* **10**(9) (2001) 1357–1364.
23. A. Nait-Ali, Hidden biometrics: Towards using biosignals and biomedical images for security applications, 2011 *7th Int. Workshop on Systems, Signal Processing and Their Applications (WOSSPA)*, May 2011, pp. 352–356.
24. K. Nakagaki and R. Mukundan, A fast 4×4 forward discrete Tchebichef transform algorithm, *IEEE Signal Process. Lett.* **14**(10) (2007) 684–687.
25. B.-S. Oh, K.-A. Toh, K. Choi, A. Beng Jin Teoh and J. Kim, Extraction and fusion of partial face features for cancelable identity verification, *Pattern Recogn.* **45**(9) (2012) 3288–3303.
26. E. Paquet and M. Rioux, Invariant robust 3-d face recognition based on the hilbert transform in spectral space, *J. Multimedia* **1**(1) (2006) 9–15.
27. M. Parisa Beham and S. Mohamed Mansoor Roomi, A review of face reconnition methods *Int. J. Pattern Recogn. Artif. Intell.* **27** (2013) 35.
28. D. Parikh and R. Polikar, An ensemble-based incremental learning approach to data fusion, *IEEE Trans. Syst. Man, Cybern B, Cybern.* **37**(2) (2007) 437–450.
29. P. Phillips, H. Moon, S. Rizvi and P. Rauss, The FERET evaluation methodology for face-recognition algorithms, *NISTIR 6264 and IEEE Trans. Pattern Anal. Mach. Intell.* **22**(10) (2000) 1090–1104.
30. W. Pratt, J. Kane and H. Andrews, Hadamard transform image coding, *Proc. IEEE* **57**(1) (1969) 58–68.
31. L. Qiao and S. Chen, Image analysis based on two dimensional Hilbert Huang transform, *2012 IEEE Int. Conf. Computer Science and Automation Engineering (CSAE)*, Vol. 2, May 2012, pp. 751–755.
32. R. K. Sadykhov, V. A. Samokhval and L. P. Podenok, Face recognition algorithm on the basis of truncated Walsh-Hadamard transform and synthetic discriminant functions, *FGR* (IEEE Computer Society, 2004), pp. 219–222.
33. P. Sanguansat, W. Asdornwised, S. Jitapunkul and S. Marukatat, Two-dimensional linear discriminant analysis of principle component vectors for face recognition, in *ICASSP 2006 Proc. 2006 IEEE International Conf. Acoustics, Speech and Signal Processing, 2006.* Vol. 2 (2006), p. II.
34. L. Sirovich and M. Kirby, Low-dimensional procedure for the characterization of human faces, *J. Opt. Soc. Am. A* **4**(3) (1987) 519–524.
35. F. Song, J. You, D. Zhang and Y. Xu, Impact of full rank principal component analysis on classification algorithms for face recognition, *Int. J. Pattern Recogn. Artif. Intell.* **26**(3) (2012) 1256005-1–1256005-23.
36. M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cogn. Neurosci.* **3**(1) (1991) 71–86.

37. J. Yang, D. Zhang, A. F. Frangi and J. Yu Yang, Two-dimensional pca: A new approach to appearance-based face representation and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **26** (2004) 131–137.
38. H. Yu and J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, *Pattern Recogn.* **34** (2001) 2067–2070.
39. D. Zhang, J. Pan, Y. Y. Tang and C. Wang, Illumination invariant face recognition based on the new phase features, *SMC* (2010), pp. 3909–3914.

Juan Luis Fernández-Martínez received his Ph.D. in Mining Engineering from the University of Oviedo (Spain) in 1994 when he joined the Mathematics Department of Oviedo University in 1994 where he is Professor of Applied Mathematics. During 2008–2010 he was Visiting and Research Professor at UC Berkeley-Lawrence Berkeley Laboratories and Stanford University. He currently works on inverse problems and uncertainty quantification applied to the oil industry and also on biomedicine (cancer, genomics and proteomics).

Ana Cernea received her B.Sc. degree in Mathematics from the University of Bucharest, Romania, in 1989 and her B.Sc. degree in Computer Science from the Oviedo University, Spain. Actually she is completing her Ph.D. on image processing optimization. Her fields of research also include global optimization techniques and digital signal processing with applications in biometrics.

A.2. Unsupervised Ensemble Classification for Biometric Applications

Authors:
Ana Cernea
Juan Luis Fernández Martínez

International Journal of Pattern Recognition and Artificial Intelligence
Vol. 28, No. 4 (2014) 1456007 (32 pages)
DOI: 10.1142/S0218001414560072
Published on 11 June 2014

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 28, No. 4 (2014) 1456007 (32 pages)

DOI: 10.1142/S0218001414560072

World Scientific
www.worldscientific.com

UNSUPERVISED ENSEMBLE CLASSIFICATION FOR BIOMETRIC APPLICATIONS

ANA CERNEA* and JUAN. LUIS. FERNÁNDEZ-MARTÍNEZ†

Mathematics Department. Oviedo University
C/Calvo Sotelo s/n, 33007 Oviedo, Spain
**cerneadoina@uniovi.es*
†*jlfm@uniovi.es*

Received 13 November 2013
Accepted 25 April 2014
Published 11 June 2014

In this paper, we propose different ensemble learning algorithms and their application to the face recognition problem. Three types of attributes are used for image representation: statistical, spectral, and segmentation features and regional descriptors. Classification is performed by nearest neighbor using different p-norms defined in the corresponding spaces of attributes. In this approach, each attribute together with its corresponding type of the analysis (local or global) and the distance criterion (norm or cosine), define a different classifier. The classification is unsupervised since no class information is used to improve the design of the different classifiers. Three different versions of ensemble classifiers are proposed in this paper: $CAV1$, $CAV2$, and $CBAG$, being the main differences among them the way the image candidates that perform the consensus are selected. The main results shown in this paper are the following: 1. The statistical attributes (local histogram and percentiles) are the individual classifiers that provided the higher accuracies, followed by the spectral methods (DWT), and the regional features (texture analysis). 2. No single attribute is able to provide systematically 100% accuracy over the ORL database. 3. The accuracy and stability of the classification is increased by consensus classification (ensemble learning techniques). 4. Optimum results are obtained by reducing the number of classifiers taking into account their diversity, and by optimizing the parameters of these classifiers using a member of the Particle Swarm Optimization (PSO) family. These results are in accord with the conclusions that are presented in the literature using ensemble learning methodologies, that is, it is possible to build strong classifiers by assembling different weak (or simple) classifiers based on different and diverse image attributes. Due to these encouraging results, future research will be devoted to the use of supervised ensemble techniques in face recognition and in other important biometric problems.

Keywords: Face recognition; statistical attributes; spectral attributes; segmentation features; ensemble learning.

† Corresponding author.

1. Introduction

Automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision with applications in security authentication,[9] biometry, and in biomedicine,[25] to name some of the application fields.

One of the main challenges in dealing with very large databases of images is establishing a low-dimensional feature representation having enough discriminatory power to perform high accuracy classification and/or prediction with the corresponding uncertainty quantification. Different techniques based on image attributes have been proposed in the literature.[18]

In this paper, we present different ensemble learning algorithms for automatic image recognition in biometry, and particularly in face recognition. This paper is organized as follows:

- In Sec. 2, we present the automatic image recognition problem and we introduce the ensemble learning approaches.
- In Sec. 3, we present the different image attributes that are used in the ensemble learning methodologies. Three different types of attributes are used for image representation:
  - Statistical attributes, including local and global color histograms, percentiles arrays and omnidirectional variogram accounting for the spatial color distribution.
  - Segmentation features and regional descriptors, including different measures of texture analysis and edge detection. This type of attributes are very well known in image processing (see for instance Ref. 34).
  - Spectral attributes that are obtained by projecting the images onto different spectral basis. These methods can be divided into two main categories: (1) covariance-based methods, involving the diagonalization of different image covariance and correlation matrices. (2) covariance-free methods involving different orthogonal transformations defined over the set of single images. A numerical comparison between these methods has been recently performed by Fernández-Martínez and Cernea,[11] showing that the most crucial parameter in the accuracy of the classification was the energy cut-off used to perform the image energy compression and model reduction.

  Two different kinds of analysis are performed using global and local attributes. In the case of the global analysis, the attributes are calculated over the whole size of the images in the learning database. In the case of the local features, this analysis is performed by dividing all images into blocks. For each block the attributes are computed, and the final feature vector is defined by merging in an unique vector all the local attribute vectors of the image, always computed on the same order. In our approach, each type of image attribute combined with a kind of analysis (local or

global) and a similarity criterion (p-norm or cosine defined in the space of attributes) induces a different classifier. First we compare the accuracy of all the single classifiers that are presented in this paper, concluding that the statistical (the histogram and the percentiles) and the spectral attributes (Discrete Wavelet Transform, DWT) are the ones that provide the higher identification rates. Second, these attributes are combined to build the ensemble classifiers.

- In Sec. 4, we present three different unsupervised ensemble learning algorithms:
  - In the first algorithm (called $CAV1$), we retain the N_c nearest neighbor image candidates from every individual classifier (based on a single attribute), and the prediction is performed by majority voting (Borda count) taking into account the individual reliability (prior accuracy) of each classifier.
  - In the second version (called $CAV2$), the N_f first candidates are selected by the most performing prior classifier (the local histogram in this case), and the rest of classifiers will act on this reduced-size database of preselected images. Decisions are also taken by majority voting.
  - Finally, the third version (called (called $CBAG$) is based on ensemble prediction using different classifiers and bagging.[6] In this case, the bags are obtained by partitioning the learning database into different learning data subsets that are randomly drawn, and for each of these subsets the classifiers are randomly generated combining several attributes from the above mentioned categories. These classifiers are finally combined by taking the majority score of their decisions.

Diversity among the different classifiers that are used is also an important feature in ensemble learning.[28] We show that we can improve the accuracy by choosing the five most diverse attributes from our list: local histogram, variogram, DWT, texture and edges detection.

To perform the numerical analysis we have used the ORL database of faces provided by ATT Laboratories at Cambridge (http://www.cl.cam.ac.uk/research/dtg/attarchive). The ORL database contains 10 different face images (poses) of 40 distinct subjects. All the images were taken against a dark homogeneous background, varying the lighting, facial expressions and facial details. The database provides upright and frontal poses. The size of each image is 92×112 pixels, with 256 gray levels per pixel. Nevertheless these algorithms could also be applied to other image databases or biometric problems. The final statistics that are provided are based on 100 independent simulations (folds). In each fold, the ORL database is divided into two different groups: 200 images to learn (learning database) and 200 images to establish the accuracy of the different classifiers (testing database). These simulations are kept the same for all the numerical analyses shown in this paper, since the indexes of the images in the different folds are known before performing the analysis for each attribute.

- Finally, in Sec. 5, we outline the conclusions and the future research work. The statistical and spectral attributes provide the higher accuracies. Nevertheless, no single attribute is able to provide systematically 100% accuracy. The ensemble learning methodology serves to improve the accuracy and stability of the classification provided by any single classifier. Therefore, it is possible to build strong classifiers by assembling different simple classifiers based on diverse image attributes. Optimizing the parameters of the ensemble classifiers that has been proposed in this paper by using fast global optimization techniques seems to be the right way to make this methodology evolving in future research.

2. Automatic Image Recognition and Ensemble Learning

The automatic image recognition problem consists of classifying a given probe image I, by providing a database of training images, corresponding to known individuals.

Given a database of training images (the learning database)

$$Bd = \{I_k \in \mathcal{M}_{(n,m,c)}(\mathbb{N}) : k = 1, \ldots, N\}, \tag{1}$$

organized into q classes:

$$C = \{C_k \in \{1, 2, \ldots, q\}, k = 1, \ldots, N\}, \tag{2}$$

and a new incoming image $I \notin Bd$, the problem consists in estimating its class $C_I^* \in C$, according to the different poses available in Bd.

In this definition, $\mathcal{M}_{(n,m,c)}$ is the space of RGB color (if $c = 3$) or gray-scale (if $c = 1$) images of size $m \times n$, and the learning database typically contains N_p poses for each class $C_i \in C(i = 1, \ldots, q)$, that is, $N = q \cdot N_p$.

To perform the classification it is necessary to construct a learning algorithm, since no mathematical model is known to perform the class prediction. A learning method L^* is a functional defined over the set of images $\mathcal{M}_{(n,m,c)}$ into the set of classes C:

$$L^* : \mathcal{M}_{(n,m,c)} \to C, \quad L^*(I_j) = C_j. \tag{3}$$

In the case of unsupervised learning the methodology is as follows:

(1) First finding the image $I_k \in Bd$ such as:

$$d(I, I_k) = \min_{I_j \in Bd} d(I, I_j), \tag{4}$$

where d is a suitable distance (or norm) criterium defined over $\mathcal{M}_{(n,m,c)}$.

(2) Once this image has been found, its class is assigned to the new image I: $L^*(I) = L^*(I_k) = C_k$.

Along this article, feature vectors and attribute vectors are synonymous and represent the images of the database according to the different methods of analysis.

Let $\mathbf{v}_i^k \in \mathbb{R}^{s_k}$ be the feature vector of length s_k corresponding to image I_i according to the attribute k, then the distance between two images I_i and I_j is calculated as follows:

$$d_k(I_i, I_j) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \tag{5}$$

where p is a norm defined over $\mathbb{R}^{s_k}$. Typically, the L_2 norm ($p = 2$) is used and it is related to the Euclidean scalar product. In this case, it is also possible to maximize the cosine of the angle between these two images according to their attributes, as follows:

$$\cos_k(I_i, I_j) = \frac{\mathbf{v}_i^k \cdot \mathbf{v}_j^k}{\|\mathbf{v}_i^k\|_2 \cdot \|\mathbf{v}_j^k\|_2}. \tag{6}$$

The success of the unsupervised classification depends on the implicit relationship between the k-attribute that is used (global or local, as explained in Sec. 3), and the adopted criterion (norm or cosine) to define the cost function of the classifier, and the class information. Not all the attributes (classifiers) used in this paper perform the same.

In the case of supervised classification, the learning method $L^*(I_j; \mathbf{m})$ depends also on a set of parameters $\mathbf{m}$ that has to be tuned using class information coming from the testing database. Supervised ensemble learning using the present setup (each attribute inducing a different classifier) will be explored in future research.

Ensemble learning methodologies try to exploit the intrinsic uncertainty of any unsupervised/supervised classification problem.[26] The main idea in ensemble learning is to be able to weigh and combine several single classifiers in order to obtain a stronger classifier that outperforms every one of them.[28] Ensemble classification is related to majority vote decisions,[5] that are based on the fact that the probability of being correct for a majority of independent voters (classifiers) is higher than the probability of any "fairly good" single voter (classifier).[17] *Fairly good* refers in this case to a probability of being correct higher than 0.5. Also, the majority consensus probability of being correct tends to 1 as the number of independent voters approaches infinity. Schapire[30] used this fact to prove that ensemble voting using weak classifiers, which are slightly more accurate than random classification, produces a strong classifier.

In the face recognition context, ensemble learning consists in building a set of diverse classifiers based on different single image attributes. Most of the ensemble learning techniques used for face recognition belong to the category of supervised learning using boosting or bagging approaches based on individual classifiers.[8,10,17,22–24,39,42] The algorithms introduced in this paper are unsupervised, although they can be used in the modality of ensemble supervised learning if the parameters on which they depend are optimized using the class information.

The general framework used in this paper for the consensus classification is as follows:

- The data consists of a training set $T \subseteq Bd$ that contains $N_L \leq N$ images belonging to Bd and all of them are characterized by a set of N_a feature vectors corresponding to the N_a considered attributes. In this paper, we consider 11 different attributes that are briefly explained in Sec. 3. All these attributes are pre-computed for all the images. This fact accelerates the classification process.
- Given a new incoming image $I \in \mathcal{M}_{(n,m,c)}$, $I \notin T$, represented through its N_a feature vectors, and corresponding to one of the q classes in the database Bd, the face recognition problem consists in assigning its class C_I^*. In this approach, each of the q attributes and the similarity criterion used to compare the images constitutes a single individual classifier that is going to be used for consensus.
- Based on each classifier, the N_c images closest to I (according to the similarity criterion) are selected. The number of image candidates, N_c, and the weights of any individual classifier, are parameters that might be optimized in the case of supervised methods. In the case of unsupervised learning, the weights coincide with the prior accuracy of the individual classifiers, while the effect of the other parameters are numerically explored in Sec. 4.
- Finally, based on the selected candidates, their score information, and the weights that are given to any individual classifier, the ensemble prediction is performed. This methodology also produces a set of other candidates to be the solution of the face recognition problem, having lower score than the optimum solution found. This fact can be interesting to solve the face recognition problem with its corresponding uncertainty assessment.

3. Face Image Attributes

In this section, we briefly describe the attributes that are used in the ensemble learning methodologies. Faces are described by different kind of image attributes: statistical-based features, spectral features, and image segmentation/regional descriptors features (texture-based features). All these attributes can be calculated for gray-scale and color images, both, locally or globally. In the case of global analysis, the attribute features are calculated over the whole size of the image. In the case of local features, this analysis is performed by dividing the images into blocks. For each block, the local attributes are computed and the final feature vector is formed by merging all the local attributes into a unique vector, always computed in the same order. Figure 1 shows the sketch of this process. The use of local features increases the dimension of the attribute space, but it is also expected to increase the discriminative power of the analysis. In this paper, we have used a partition of the images into 8×4 blocks. Finer subdivisions could be also adopted.

As it have been presented in the introduction, to perform the numerical analysis we have used the ORL database of faces, described before. We provide statistics

Fig. 1. Image partition in blocks for the local analysis $\mathbf{v}_I = (\mathbf{v}_1, \ldots, \mathbf{v}_{M\times N})$.

(minimum, maximum, median, mean accuracy, interquartile range, and standard deviation) for the classification calculated over 100 different simulations for each individual attribute. In this context, mean accuracy refers to the number of images (over 200) that are correctly classified over the 100 different folds. The other statistical measures used in this paper follow the same logic. The classification is performed using different p-norms, and the cosine criteria defined in the space of attributes, that were presented in Sec. 2. According to these previous results we select the attributes with their corresponding kind of analysis (global/local) and their similarity criterion (p-norm/cosine) for the ensemble learning methodology that is presented in Sec. 4.

3.1. *Statistical attributes*

3.1.1. *Color histograms*

An image histogram describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image. For example, a narrow histogram means a low contrast, meanwhile a histogram shifted to the right indicates a bright image; a histogram with two major peaks (called bimodal) implies the presence of an object that is in contrast with the background, etc.[37] Due to this fact it is expected that the histogram will have a great discriminative power. Also, this attribute is fast to compute and can be precalculated for all the images in the learning database.

For a gray-scale digital image I, the histogram represents the discrete probability distribution of the gray levels in the image. For this purpose the gray-scale space

([0, 255] for an `8-bit` image) is divided into L bins, and the number of pixels in each class n_i, $(i = 1, L)$ is calculated. In this case, the attribute vector has dimension L (usually $L = 256$):

$$H_I = (n_1, \ldots, n_L). \tag{7}$$

Relative frequencies can also be used by dividing the absolute frequencies n_i by the total number of pixels in the image.

In the case of RGB images, the same analysis is performed considering the color channels I_R, I_G and I_B independently, and merging all the channels histograms, as follows:

$$H_I = (H(I_R), H(I_G), H(I_B)). \tag{8}$$

The histogram can be calculated either globally over the whole image, or locally. Global histograms do not capture the spatial distribution of the color channels. Local histograms provide this kind of information. Their calculation follows the general procedure for computing any local attribute explained above (see Fig. 1).

Table 1 presents a comparison of the accuracies for the global and local histogram analysis performed on the ORL database over 100 different independent simulations. Bold faces represent the optimum results. This analysis shows:

- The local histogram provides higher accuracies and lower dispersions than the global one for all the classification criteria (L1 and L2 distances and cosine).
- In some simulations, we achieved 100% accuracy using local histograms. This result is very important since it shows that using unsophisticated attributes, it is possible to achieve a perfect classification if the learning database is properly built, that is, if it contains enough information about poses variations for each individual.
- The best results are obtained for the local histogram with the L1 norm (98.80% median accuracy). The L1 norm also provides the lowest dispersion (IQR and std).

3.1.2. *Percentiles*

The p-percentile of a color (or gray) channel c_i is defined as the number x_p such as

$$P(c_i \leq x_p) = p \in [0, 1]. \tag{9}$$

Table 1. Gray level histogram: statistics for the global and local analysis with the L1, L2 norms and the cosine criterion. Bold faces highlight the optimum results.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L1	92.00	100.00	96.50	96.09	2.25	1.73
Global L2	88.50	98.50	95.00	94.99	3.00	1.95
Global cosine	88.50	98.50	95.00	94.97	2.00	1.87
Local L1	**96.00**	**100.00**	**98.80**	**98.74**	**1.50**	**0.83**
Local L2	89.50	**100.00**	96.50	96.10	2.00	1.75
Local cos	91.50	**100.00**	96.55	96.63	1.90	1.75

In practice, we compute the percentiles 1%, 99% and 5% to 95% with 5% step of probability. This discretization of the cumulative probability function produces a feature vector of dimension 21 for each image block of local analysis and color channel.

Percentiles and histograms are related because the histogram is the density probability function $f(t)$, and the cumulative probability function $F(x)$ is the integral:

$$F(x) = \int_{-\infty}^{x} f(t)dt \iff f(x) = F'(x). \tag{10}$$

Nevertheless, histograms and percentiles provide different results for classification purposes.

Table 2 shows the statistics for the global and local analysis using the percentiles. As in the previous case the best result is also obtained with the local analysis using the L1 norm (97.00% median accuracy). These results are slightly worse than those provided by the local histogram, both, in accuracy and dispersion.

3.1.3. *Variogram*

The variogram describes the spatial distribution in each color channel. In spatial statistics,[15] the variogram describes the degree of spatial dependence of a spatial random field or stochastic process, the gray-scale in this case. For a given value of vector h, defined by a modulus and direction, the variogram measures dissimilarity between all pairs of values separated by vector h.

The omnidirectional p-variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^p. \tag{11}$$

Usually $p = 2$. To compute the variogram, each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$.

Table 2. Percentiles: statistics for the global and local analysis with the L1, L2 norms and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L1	86.00	97.00	92.00	91.84	3.00	2.31
Global L2	82.50	96.00	89.00	89.32	3.50	2.37
Global cosine	82.50	94.50	88.75	88.82	3.00	2.55
Local L1	**92.50**	**99.50**	**97.00**	**97.05**	**2.00**	**1.40**
Local L2	91.00	99.00	95.50	95.09	**2.00**	1.60
Local cosine	89.50	97.00	94.00	93.80	**2.00**	1.59

Table 3. Variogram: statistics for the global and local analysis with the L1, L2 norms and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L1	67.50	79.00	74.25	73.98	4.00	2.72
Global L2	65.50	81.00	74.00	74.21	3.50	2.71
Global cosine	63.50	78.00	72.50	72.39	3.00	2.50
Local L1	**88.50**	**96.50**	**93.00**	**92.86**	2.50	1.84
Local L2	85.00	94.50	90.00	89.88	**2.25**	1.81
Local cosine	84.00	94.00	89.50	89.30	2.50	**1.73**

Table 3 shows the experimental results for the global and local analysis using the variogram attribute. The best results are again obtained for the local analysis using the L1 norm (93.00% median accuracy). The mean accuracy is lower than those corresponding to the percentiles and the histogram.

The variogram could also be considered as an image texture descriptor. The texture attributes are described in the following section.

3.2. *Segmentation features and image analysis descriptors*

3.2.1. *Texture analysis*

Texture analysis of an image consists in analyzing regular repetitions of a pattern.[35] For texture analysis, we use the spatial gray level co-occurrence matrix (GLCM) which is based on the joint probability distributions of pairs of pixels.

The GLCM, or spatial dependence matrix of an image I is an estimate of the second-order joint probability function $P_{d,\theta}(i,j)$ of the intensity values of two pixels i and j located at a distance d apart (measured in number of pixels) along a given direction θ.[2,4] Typically the GLCM is calculated for different pairs of d and θ. Figure 2 shows the spatial relationships between a pixel and its adjacent pixels, and the corresponding displacement vector (d,θ).

To calculate the GLCM matrix for a given pair (d,θ), the algorithm proceeds as follows:

- First, the matrix $F_{d,\theta}(i,j)$ of absolute frequencies of pairs of pixels with gray levels i and j at a distance d in the direction θ is built. For instance, Fig. 3 shows the $F_{d,\theta}$ matrix for a 4×4 image with five gray levels, for the displacement vector $(1,0)$.

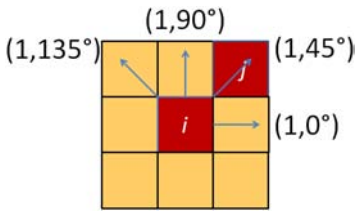


Fig. 2. Spatial relationships of a pixel i with its adjacent pixels.

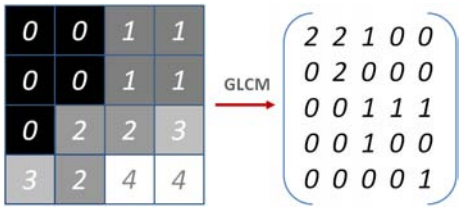


Fig. 3. $F_{1,0}$ for a 4×4 image with five gray levels.

- Second, $F_{d,\theta}$ is normalized as follows:

$$P_{d,\theta}(i,j) = \frac{F_{d,\theta}(i,j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} F_{d,\theta}(i,j)}. \tag{12}$$

The GLCM matrix is square and has a dimension of 256×256 for gray color image.

Different statistical moments can be calculated from the GLCM matrix[4]: contrast, homogeneity, energy, correlation and entropy (refer to Appendix 1 for their definitions). In the present case, we have used a lag $d = 1$ for the directions 0, 45, 90, 135. We have explored the individual accuracy of the different statistical moments of texture, arriving to the conclusion that in the case of local analysis the highest accuracy is provided by the contrast and the homogeneity. The rest of the moments (energy, correlation and entropy) lowered the accuracy provided by the contrast and the homogeneity. For the global analysis we have used all the texture moments. Statistics for global and local texture analysis are shown in Table 4. The median accuracy provided by the global texture analysis is low (around 72.50%); nevertheless, the local analysis improve dramatically the results providing a median accuracy of 92.50% using the L1 norm. This analysis shows that texture is a local attribute for describing and quantifying the spatial arrangement of color or intensities in an image or in a selected region.

3.2.2. *Edges detection*

Edges are determined by sets of pixels where there is an abrupt change in intensity. If a pixel's gray level value is similar to those around it, there is probably not an edge at that point. However, if a pixel has neighbors with widely varying gray levels, it may

Table 4. Texture analysis: statistics for the global and local analysis, with the L1, L2 norms and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L1	65.00	79.50	72.50	72.39	4.00	2.93
Global L2	61.00	76.00	69.25	68.82	3.50	2.95
Global cosine	53.00	68.50	62.50	62.39	3.50	2.74
Local L1	**86.50**	**97.00**	**92.50**	**92.52**	3.00	2.05
Local L2	84.00	94.00	90.50	89.84	3.00	2.15
Local cosine	85.50	94.50	90.75	90.76	**2.75**	**2.04**

represent an edge. Thus, an edge is defined by a discontinuity in the gray-level values.[37] More precisely, we can consider an edge as a property associated to a pixel (x, y) where the image function $f(x, y)$ changes rapidly in the neighborhood of that pixel. In this case, the function $f : \mathbb{R}^2 \to \mathbb{R}$ represents the pixels intensities. Related to f, an edge is a vector variable with two components: magnitude and direction. The edge magnitude is given by the gradient and its direction is perpendicular to the gradient's direction:

$$|\nabla f(x, y)| = \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2}, \tag{13}$$

$$\theta(x, y) = arctg\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right) \pm \frac{\pi}{2}. \tag{14}$$

In the case of digital images, these expressions are approximated by differences schemes. The gradient can be computed using different convolution masks, such as Robert, Sobel, Prewitt, Laplace, Canny, etc. A recent comparison of the mentioned edge detection operators can be found in Ref. 31. In this case, we have used the Canny edge detection operator,[7] which is one of the most commonly used in image processing due to its property of detecting edges in a very robust manner in the case of noisy images. These operators provide an edge map of the same size as the incoming images that are not suitable for recognition purposes due to their size and to the impact of noise in the edge maps. Therefore in order to reduce the dimension and to calculate valuable edge attributes for face recognition, the bidimensional Discrete Cosine Transform (DCT) is applied to the edge map. As a result of this orthogonal transformation, the energy of the edge map is concentrated on the left upper block of the DCT transform. Denoting by E_d the DCT transform of the edge map, the feature vector is formed by the $n_1 \times n_2$ submatrix of E_d, $E_d(1 : n_1, 1 : n_2)$. In our case, it has been determined by numerical experimentation that the optimal parameters were $n_1 = n_2 = 10$ in the case of the global analysis, while $n_1 = n_2 = 3$ in the case of the local analysis. Figure 4 shows sketch of this process for one image of the ORL

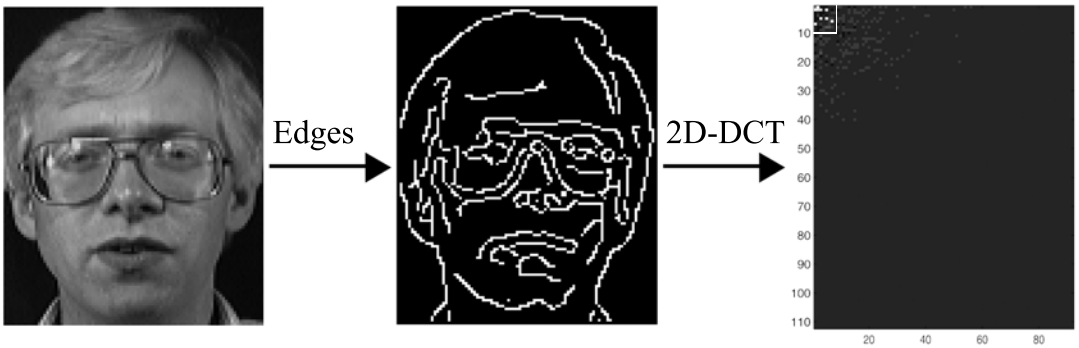


Fig. 4. Method to produce the edges attribute. Once the edge map is produced, the DCT is applied to reduce its dimension.

Table 5. Edge detection: statistics for the global and local analysis with L1, L2 norms and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L1	83.50	**94.00**	89.00	89.18	**2.50**	1.95
Global L2	**85.00**	**94.00**	**90.85**	**90.00**	2.75	**1.89**
Global cosine	83.00	93.50	88.50	88.55	**2.50**	1.92
Local L1	80.00	93.00	86.75	86.68	**2.50**	2.17
Local L2	80.50	90.00	86.00	85.73	3.00	2.18
Local cosine	78.50	89.50	85.50	85.17	3.50	2.38

database. This approach is different from the Line Edge Map (LEM)[22] that uses the Line Segment Hausdorff Distance (LHD) to measure the similarity of face LEMs. This method needs supervision to tune some of the parameters used to define the LHD, and provides accuracies (around 92.00%) similar to the method that we propose.

Table 5 shows the accuracies statistics for the global and local analysis using the edge detection attribute. The median accuracies are similar to the ones provided by texture analysis. The best median accuracy is achieved in the case of the global analysis with the L2 norm (90.85%). It is important to note that this attribute provides lower accuracies for the local analysis and/or with the L1 norm.

3.3. *Spectral attributes*

In this section, we briefly describe the spectral methods that are used in the proposed consensus algorithm. A detailed analysis and comparison among these methods can be found in Ref. 11. Spectral decomposition methods consist in finding orthonormal bases which best separates the image projections, and serve to reduce the dimension of the image attributes space with respect to the pixels space ($\mathbb{R}^{N\text{pixels}}$). These methods can be divided into two categories:

- Covariance-based: PCA, 2DPCA, Fisher's LDA. These methods act on the whole database of training images and involve diagonalization of different kind of experimental covariance/correlation matrices defined over the whole set of training images (Bd).
- Covariance-free: These methods are based on orthogonal transformations U and V defined over $\mathbb{R}^m$ and $\mathbb{R}^n$ and perform on individual images I, such as:

$$I = U\Sigma_I V^T. \tag{15}$$

  Σ_I is called the spectral transform of image I. The following techniques are used in this paper:

  — DWT: matrices U and V are built using different kind of wavelet functions.
  — DCT: this transform uses Fourier basis of cosines.
  — Discrete Walsh-Hadamard Transform (DWHT): the basis are the Walsh functions.

In each case, we give a brief description of the spectral method and its accuracy on the ORL database over 100 simulations. In the numerical experiments, we do not provide the results obtained with the L1 norm, since in previous experiments this norm provided lower accuracies than the L2 norm.[11]

3.3.1. *Principal component analysis (eigenfaces)*

Principal component analysis (PCA) finds the orthogonal basis (eigenfaces) by diagonalizing the experimental covariance matrix of training images:

$$S_{\text{PCA}} = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^T, \tag{16}$$

where $X_k \in \mathbb{R}^{N\text{pixels}}$, $K = 1, \ldots, N$ are the database images transformed into column vectors, $\mu = \frac{1}{N}\sum_{k=1}^{N} X_k$ is the images sample mean, N is the number of sample images contained in the learning database, and N_{pixels} is the number of pixels of each image. The eigenfaces $\mathbf{u}_k$ are the eigenvectors of S_{PCA}, corresponding to the largest eigenvalues. The dimensionality reduction from N_{pixels} to q parameters, is obtained by retaining the q first eigenfaces $\mathbf{u}_k$, spanning most of the database variability. The projection of an image X_j onto this reduced basis is obtained as follows:

$$\mathbf{a}_j = U_q^T (X_j - \boldsymbol{\mu}), \tag{17}$$

with $U_q = [\mathbf{u}_1, \ldots, \mathbf{u}_q]$.

PCA has been applied to image analysis and face recognition.[20,33,36]

Table 6 shows the statistics for the global and local analysis using the PCA attribute. The best accuracy and the lower dispersion are achieved with the cosine criterion for the local analysis (95% of median accuracies).

3.3.2. *2DPCA*

2DPCA diagonalizes the following mean centered covariance matrix to find the orthogonal projection basis:

$$S_{\text{2DPCA}} = \frac{1}{N}\sum_{i=1}^{N} (I_i - \bar{I})^T (I_i - \bar{I}), \tag{18}$$

where I_i belongs to Bd and $\bar{I}$ is the mean image matrix calculated pixel by pixel over the learning database. The covariance matrix S_{2DPCA} is calculated directly using the

Table 6. PCA classifier: statistics for the global and local analysis with the L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	88.50	97.50	93.50	93.30	2.50	1.89
Global cosine	88.50	97.50	93.50	93.52	**2.00**	1.76
Local L2	91.00	**98.00**	**95.00**	**94.83**	**2.00**	1.60
Local cosine	**91.50**	**98.00**	**95.00**	94.82	2.50	**1.52**

Table 7. 2DPCA: statistics for the global and local analysis with the L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	**90.50**	**99.00**	94.50	94.64	2.50	1.73
Global cosine	89.00	98.00	93.75	93.77	**2.25**	**1.52**
Local L2	90.00	98.00	**95.00**	**94.73**	2.50	1.69
Local cosine	89.00	97.50	93.00	93.18	2.50	1.63

original image matrices, and it is much smaller than the PCA covariance3 matrix. 2DPCA has been introduced by Yang *et al.*[40] for face recognition.

Table 7 shows the statistics for the local and the global analysis of the 2DPCA attribute. The best minimum and maximum accuracies are obtained for the global analysis with the L2 norm. Nevertheless, the best median and mean accuracies are obtained for the local analysis with the L2 norm (95.00%).

3.3.3. *Fisher's linear discriminant (Fisherfaces)*

This technique is also based on a linear projection W, which attempts to reduce dimensionality while preserving as much of the class discriminatory information as possible. Fisher's solution to this problem is the matrix W which maximizes the differences between classes, normalized by the scatter within-class.[14] For that purpose, between- (S_B) and within-class (S_W) scatter matrices are defined:

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{19}$$

$$S_W = \sum_{i=1}^{C} \sum_{X_k \in C_i} (X_k - \mu_i)(X_k - \mu_i)^T, \tag{20}$$

where C is the total number of classes in the database, N_i is the number of images in each class C_i, and μ_i is the mean of the images in class i. Numerical implementation of this method can be consulted in Refs. 38 and 41. Fisher's LDA, has been applied to the face recognition problem by Belhumeur,[3] who showed that Fisherfaces have error rates lower than eigenfaces in face recognition. This is due to the fact that Fisher's LDA is a supervised projection method. This method needs to have enough poses at disposal to calculate the Fisher's reduced basis.

Table 8 shows the statistics for the global and local analysis in the case of the Fisher's LDA. The best results are achieved in the case of the local analysis with the L2 norm (95%).

3.3.4. *Discrete wavelet transform (DWT)*

Wavelets are compact functions with zero mean and some regularity conditions (vanishing moments). The wavelet transform converts a function into a linear

Table 8. Fisher's LDA: statistics for the global and local analysis with L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	83.50	93.50	88.00	88.13	2.75	2.12
Global cosine	81.00	92.00	86.25	86.23	3.25	2.25
Local L2	**90.50**	**98.00**	**95.00**	**94.96**	**2.00**	**1.62**
Local cosine	88.00	97.50	93.50	93.58	**2.00**	1.74

combination of basis functions, called wavelets, obtained from a prototype wavelet through dilatations, contractions and translations.

The discrete wavelet transform (D_{WT}) of an image $I \in \mathcal{M}(m,n)$ is defined as follows:

$$D_{\mathrm{WT}} = U_W^T I V_W, \tag{21}$$

where U_W and V_W are two orthogonal matrices constructed as follows:

$$U_W = \left[\frac{H}{G}\right]_m^T, \quad V_W = \left[\frac{H}{G}\right]_n^T, \tag{22}$$

where H represents a low pass or averaging portion of the wavelet filter, and G is the high pass or differencing portion. In all the cases we have

$$D_{\mathrm{WT}} = \begin{pmatrix} HIH^T & HIG^T \\ GIH^T & GIG^T \end{pmatrix} = \begin{pmatrix} B & V \\ H & D \end{pmatrix}. \tag{23}$$

B is the blur, V are the vertical differences, H are the horizontal differences and D are the diagonal differences. DWT can be applied several times to further reduce the dimension of the attribute vector. In the present case, the DWT has been applied recursively twice, and the blur was used as the attribute for face recognition. DWT was applied to face recognition by Kakarwal and Deshmukh.[19]

Table 9 shows the statistics obtained for the DWT attribute. The best accuracies are obtained in the case of the global analysis with the L2 norm (95.50% of median accuracy).

Table 9. DWT: statistics for the global and local analysis with the L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	**92.50**	**98.50**	**95.50**	**95.50**	**2.00**	**1.46**
Global cosine	90.00	97.50	94.00	94.03	**2.00**	**1.56**
Local L2	88.50	97.00	93.50	93.47	2.75	1.73
Local cosine	87.00	95.00	91.50	91.54	3.00	1.80

3.3.5. *Discrete cosine transform (DCT)*

DCT is a free-covariance model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image onto a set of cosines of increasing frequency (Fourier basis).

The cosine discrete transform D_{CT} of an image $I \in \mathcal{M}(m,n)$ is defined as follows:

$$D_{\mathrm{CT}}(u,v) = c(u)c(v)\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} I(i,j)\cdot\cos\frac{\pi(2i+1)u}{2m}\cos\frac{\pi(2j+1)v}{2n} \tag{24}$$

with $u = 0,\ldots,m-1$, and $v = 0,\ldots,n-1$ and

$$c(\alpha) = \begin{cases} \dfrac{1}{\sqrt{N}} & \text{if } \alpha = 0, \\ \sqrt{\dfrac{2}{N}} & \text{if } \alpha \neq 0. \end{cases} \tag{25}$$

N is either the number of rows (m) or columns $(n)+$ of the image. The DCT can be expressed in matrix form as an orthogonal transformation

$$D_{\mathrm{CT}} = U_{\mathrm{DC}} I_k V_{\mathrm{DC}}^T, \tag{26}$$

where matrices U_{DC} and V_{DC} are orthogonal. The feature vector of an image I_k is constituted by the $q_1 \times q_2$ block of D_{CT}, $D_{\mathrm{CT}}(1:q_1, 1:q_2)$, where q_1, q_2 are determined by energy reconstruction considerations using the Frobenius norm of the image I_k.[11] DCT has been used by Hafed and Levine[16] in face recognition, showing that DCT applied to normalized images is very robust to variations in geometry and lightning.

Table 10 shows the statistics obtained for the DCT attribute. The best results are obtained in the case of the local analysis with the L2 norm (93.25% of median accuracy).

3.3.6. *Discrete Walsh-Hadamard transform (DWHT)*

The Hadamard transform is the projection of a signal onto a set of square waves called Walsh functions. The discrete Walsh-Hadamard transform D_{WHT} of an image

Table 10. DCT: statistics for the global and local analysis with the L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	87.50	96.00	92.00	92.02	**2.00**	1.81
Global cosine	83.00	93.50	88.00	87.81	2.50	2.04
Local L2	**89.00**	**97.50**	**93.25**	**93.33**	2.50	**1.75**
Local cosine	84.00	95.00	89.50	89.35	2.50	2.10

Table 11. DWHT: statistics for the global and local analysis, with the L2 norm and the cosine criterion.

Type of Analysis	Min	Max	Median	Mean	IQR	Std
Global L2	90.00	97.00	94.00	93.87	**2.00**	1.65
Global cosine	87.00	95.50	91.50	91.38	2.50	1.73
Local L2	**90.50**	**97.50**	**94.50**	**94.49**	2.50	**1.62**
Local cosine	87.50	96.00	92.00	92.04	2.50	1.75

$I \in \mathcal{M}(m,n)$ is defined as follows:

$$D_{\text{WHT}}(u,v) = \sum_{j=1}^{m} \sum_{i=1}^{n} H_M(u,i) I(i,j) H_N(v,j), \tag{27}$$

where H_M and H_N are the Hadamard matrices of order $M = \log_2 m$ and $N = \log_2 n$.[27]

The DWHT has been used in data compression by Akansu and Poluri[1] and it has also been applied to face recognition problems by Rauf.[29]

Table 11 shows the statistics obtained for the DWHT attribute. The best results are obtained in the case of the local analysis with the L2 norm (94.50% of median accuracy).

As a main conclusion, all the spectral methods (covariance-based and covariance-free) provide similar predictive accuracies. The DWT is the method that provides the highest accuracy and is computationally the fastest.

4. Ensemble Learning and Diversity Analysis

Ensemble learning methods assume that, due to the effect of noise in data, the solution with the minimum misfit on the learning dataset might not be the classifier that provides the right answer. Otherwise said, the solution of any classification/learning problem (considered as an optimization problem) is uncertain. Schapire[30] also proved that assembling together a set of weak and diverse classifiers produces a strong classifier with higher accuracy.

Table 12 shows the list of attributes that have been selected in this paper for ensemble learning applied to the face recognition problem. For each attribute, we specify their corresponding type of analysis (global or local), the similarity criterion that has been used (p-norm or cosine) and its individual median accuracy that has been obtained in Sec. 3. In the case of the DCT and DWHT slightly better accuracy results (95.25% and 95.00%) were obtained using the L3 norm.[11] The median accuracies shown in this table will be used as an initial set of weights to establish the prior reliability of each classifier. In a first attempt, we will use all the attributes in Table 12 to build the final ensemble classifier. The combination of the different classifiers aims to improve the accuracy of a single classifier. Finally, the diversity analysis allowed us to reduce the set to five main diverse attributes, and improving the predictive accuracy of the ensemble by avoiding redundancy. This fact is related to a correct sampling of the posterior distribution of the solutions, in this case for the face recognition problem.

Table 12. List of attributes and their parameters used for consensus.

Attribute	Type of Analysis	Criterion	Median
Histogram	Local	L1	98.80
Percentiles	Local	L1	97.00
Variogram	Local	L1	93.00
Texture	Local	L1	92.50
Edges	Global	L2	90.85
PCA	Local	L2	95.00
2DPCA	Local	L2	95.00
Fisher	Local	L2	95.00
DWT	Global	L2	95.50
DCT	Local	L2	93.25
DWHT	Local	L2	94.50

4.1. *Consensus algorithm version 1 (CAV1)*

The classification of a test image I using $CAV1$ is performed as follows (Fig. 5):

(1) From every classifier shown in Table 12, we retain the first N_c images that are closer to I according to the specified similarity criterion criterium for each individual classifier (p-norm or cosine).
(2) Based on this analysis a matrix $M \in M_{Nc \times Na}$ is built, containing the N_c image candidates for each of the N_a attributes ($N_a = 11$ is the total number of attributes), that is, $M(i, j)$ is the jth classified image by the ith attribute (classifier).
(3) Once M is computed, the array $\mathbf{c} \in \mathbb{R}^{Nd}$ containing the N_d *different* candidates (without repetition) is formed. For any image candidate $I_c \in \mathbf{c}$, its score according to the set of positions (i, j) in matrix M is calculated as follows:

$$\text{score}(I_c) = \sum_{k=1}^{p} P(I_k(i, j)), \tag{28}$$

with:

$$P(I_k(i, j)) = \mathbf{w}(i) \cdot S_c(j), \tag{29}$$

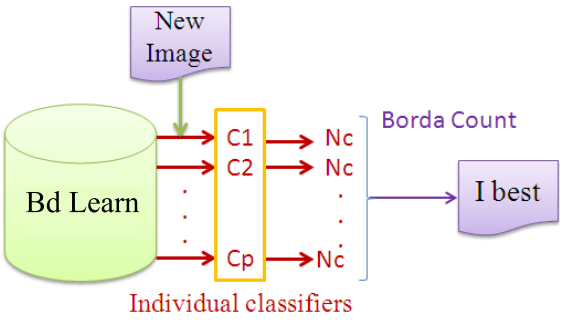


Fig. 5. Graphic flow diagram for algorithm $CAV1$.

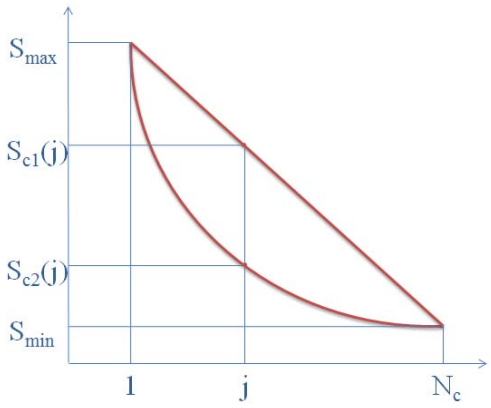


Fig. 6. Score function.

where $P(I_k(i,j))$ is the score of image $I_k(i,j)$, that represents any of the p images instances of I_c in matrix M. The weights vector $\mathbf{w}(i)$ is a trust factor assigned to the ith attribute (classifier), according to its individual accuracy given in Table 12. These weights could be optimized in a supervised learning approach. Some preliminary results are shown. $S_c(j)$ is a monotonic decreasing score function used to rank the different candidates $I_k(i,j)$ according to their position j in matrix M. Two different functions were used (Fig. 6):

$$S_{c1}(j) = S_{\max} - \frac{S_{\max} - S_{\min}}{N_c - 1}(j-1), \tag{30}$$

$$S_{c2}(j) = \frac{S_{\max}}{j^{\log_{N_c}(\frac{S_{\max}}{S_{\min}})}}, \tag{31}$$

where $S_{\max}$ and $S_{\min}$ are the maximum and minimum scores assigned to the image candidates.

(4) After calculating the scores of all the images in matrix M, the final classification of the test image I is performed by selecting the image with the highest score among all the candidate images in M.

(5) This classification method can be applied performing group substitution in matrix M before the score calculation. This means that each of the selected images in matrix M is substituted by its class (or group it belongs to) and the final score is calculated for each of the groups in matrix M.

For a given score function, the performance of the $CAV1$ algorithm obviously depends on the set of weights $\mathbf{w}$ that are used as trust factors for each individual classifier and the number of candidates (N_c) that are used to perform the selection. The first numerical experiments consisted in analyzing the impact of the choice of the score function ($S_{c1}(j)$ and $S_{c2}(j)$) with different $S_{\max}$ and $S_{\min}$ on the predictive accuracies, using the weights provided by the median accuracies shown in Table 12. This analysis has shown that both score functions provided similar results. Thus, we have decided to adopt the linear score function $S_{c1}(j) = N_c - j + 1$ with $S_{\min} = 1$ and $S_{\max} = N_c$.

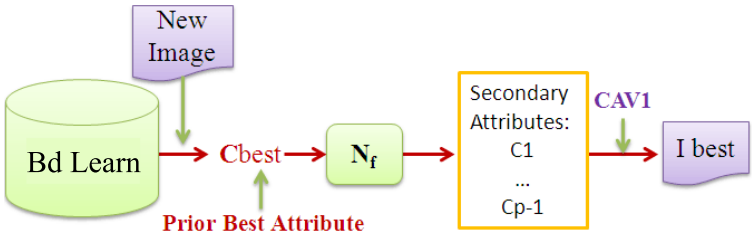


Fig. 7. Graphic flow diagram for algorithm $CAV2$.

The second kind of numerical experiments consisted in studying the effect of the group substitution and the number of image candidates (N_c) used by $CAV1$. The results obtained were almost independent of these parameters. These results depend mostly on the weights (shown in Table 12) that are assigned to the different classifiers. Nevertheless, the best results were obtained with $N_c = 20$ and group substitution (first row in Table 13 to the row corresponding to $CAV1$). The median accuracy (96.75%) is lower than the one provided by the local histogram (98.80%) shown in Table 1. These parameters could be optimally identified using a supervised learning approach. We will show in Sec. 4.4 that the diversity analysis will contribute to improve the predictive accuracy of $CAV1$ by selecting the most diverse classifiers.

4.2. *Consensus algorithm version 2 (CA V2)*

$CAV2$ algorithm is a modification of $CAV1$, as follows (see Fig. 7):

- The classification of a test image I is performed over the learning database by the individual classifier defined by the prior best attribute (local histogram, in this case) and the N_f closest images to I are retained by this classifier. Therefore, the learning database Bd is replaced by the reduced set of the N_f selected images.
- The final classification is performed applying $CAV1$ to this reduced set of images.

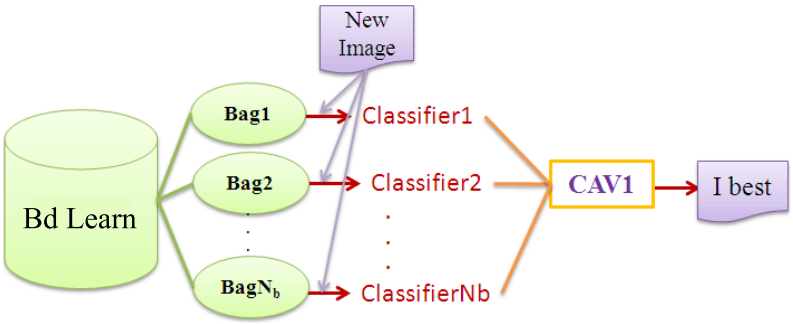


Fig. 8. Bagging algorithm.

The performance of $CAV2$ depends on the set of weights $\mathbf{w}$, on the number of images previously selected by the local histogram attribute (N_f), and finally on the number of image candidates images classified by other attributes (N_c) to perform the selection. Eventually N_f and N_c could be equal. Numerical experimentation using the weights provided by the median accuracies shown in Table 12, and varying N_f and N_c have shown that different combinations of parameters N_f and N_c provided similar results. The best results for CAV1 (97.25% median accuracy) were obtained for $N_f = 20$, $N_c = 4$, and group substitution (2nd row in Table 13). These results were slightly better (both in accuracy and dispersion) than in the case of $CAV1$, but still lower than the local histogram classifier. Again, $CAV2$ parameters could be optimally identified using a supervised learning approach. Also the diversity analysis will contribute to improve the results.

4.3. *Consensus algorithm via bagging (CBAG)*

Bagging learning method is a "bootstrap aggregating" method that aims to increase classification accuracy by training each classifier on a random distribution of the training set.[6] The ensemble's diversity is obtained by partitioning the learning database into different training data subsets (bags), randomly drawn. Each one of these subsets is used to train different classifiers. These classifiers are finally combined by taking the majority score of their decisions.

The flowchart for the $CBAG$ algorithm is as follows (see Fig. 8):

- N_{bags} image subsets are randomly generated from the learning database. Each one contains the same number of poses N_{Pb} for each person (class), already labeled:
$$\{B_k\}_{k=1,\ldots,N_{\text{bags}}} = \{(I_{jk}, C_j)\}_{j=1,N_{Pb}\cdot q,k=1,\ldots,N_{\text{bags}}}, \quad C_j \in \{1,\ldots,q\}. \tag{32}$$
- N_{bags} classifiers are generated, by randomly combining N_a individual attributes (classifiers) from the previously selected attributes according to Table 12:
$$\{L_k\}_{k=1,\ldots,N_{\text{bags}}}, L_k(I_j) = C_j, I_j \in B_k, \quad C_j \in \{1,\ldots,q\}. \tag{33}$$
In this case, the weights are the median accuracies of these attributes, and the number of candidates for each bag is either fixed or randomly selected within the range 3–25.
- A new incoming image is classified over each bag $B_k, k = 1,\ldots,N_{\text{bags}}$ by the corresponding classifier L_k and N_c candidate images are selected for each of the N_{bags}. The final classification is obtained applying $CAV1$ over the set of selected images, as explained in Sec. 4.2.

The best result obtained using $CBAG$ is shown in Table 13. The median accuracy (93.25%) is lower than the other two algorithms ($CAV1$ and $CAV2$). Also the dispersion (3.5%) is much higher. This result is quite surprising according to what it is preconized in the literature about ensemble learning methods,[6] but it could be related to the way that the different bags are generated. The diversity analysis will

Table 13. Statistics for the ensemble learning algorithms using 11 attributes.

Algorithm	Min	Max	Median	Mean	IQR	Std
*CAV*1	**95.00**	**99.00**	96.75	96.75	1.50	1.25
*CAV*2	94.50	98.00	**97.25**	**97.00**	**1.00**	**1.02**
CBAG	87.00	94.00	93.25	91.85	3.50	2.59

contribute to improve these results. Nevertheless, this algorithm is computationally more expensive than *CAV*1 and *CAV*2. *CAV*2 is the fastest.

4.4. *Diversity analysis*

Diversity among the different classifiers that we combine is a key issue in ensemble learning predictions. A simple way to introduce the diversity concept in machine learning is via a linear system of equations. Let us imagine that each equation represents a different classifier. The intersection will be optimally determined if the hyperplanes representing the equations of the linear system are mutually orthogonal. This concept brought to machine learning means that the different classifiers misclassify different images, and the diversity analysis aims to improve the final prediction by avoiding redundancy in classification. The diversity analysis presented in this section brought to the conclusion that reducing the list to only five attributes improves the results of the three ensemble learning algorithms described before. We also show numerical results performed using a list of only five attributes.

Several measures of diversity have been used in this paper (see Appendix A). It is preconized that for good diversity in ensemble classifiers that correlation, Q-statistics and agreement should be closer to zero meanwhile the rest of the diversity measures (disagreement, double-fault, entropy and Kohavi–Wolpert variance) should adopt values close to one[21,32] (see Table 14).

Table 15 shows the diversity measures, both pairwise and nonpairwise, within each of the three groups of classifiers considered. The analysis shows that the degree of correlation between the classifiers of the same group is quite high, specially the attributes belonging to the spectral group. As a consequence it is possible to choose the most relevant classifiers from each category to be used in ensemble learning methodology.

Table 14. Diversity descriptors.

Diversity Measure	Tendency	Pairwise	Reference
Correlation	↘	Yes	21,26
Q-statistic	↘	Yes	21,26
Disagreement	↗	Yes	21,26
Double-fault	↘	Yes	21,26
Entropy	↗	No	21,26
Kohavi–Wolpert	↗	No	21,26
Agreement	↘	No	21,32

Table 15. Diversity within groups of classifiers.

Group	Correlation	Q-Statistic	Disagreement	Double-Fault	Entropy	KW	Agreement
Statistical	0.26	0.80	0.09	0.03	0.12	0.03	0.26
Regional	0.19	0.51	0.35	0.03	0.35	0.09	0.09
Spectral	0.80	0.99	0.03	0.06	0.03	0.01	0.80

Table 16. Diversity between groups of classifiers.

Group	Correlation	Q-Statistic	Disagreement	Double-Fault	Entropy	KW	Agreement
Statistical-Regional	0.15	0.71	0.09	0.01	0.09	0.02	0.12
Statistical-Spectral	0.28	0.89	0.04	0.01	0.14	0.05	0.28
Spectral-Regional	0.33	0.89	0.09	0.02	0.22	0.02	0.29

Table 16 shows the diversity measures between the three groups of classifiers considered in this paper: statistical, regional and spectral. The majority of the diversity measures have improved with respect to the results shown in Table 15 (diversity within-groups). The highest degree of similarity is obtained between the statistical and spectral groups, meanwhile the statistical and regional groups seem to be the most diverse.

Table 17 shows the overall diversity of the ensemble of classifiers, presented in Table 12. Due to the high degree of correlation existing among several of these 11 classifiers, we have selected the five most diverse attributes to be used in the learning algorithms explained before. These attributes are: local histogram, local variogram, local texture, local edges and global DWT. As a result of this selection, a drastic change of the diversity measures values is obtained, as shown in the second row of Table 17.

Table 18 shows the pairwise diversity measures over the five selected classifiers. The most diverse associations are shown in bold faces. Using this set of five uncorrelated classifiers improved the ensemble learning process. Table 19 shows the accuracies obtained for the three ensemble learning algorithms ($CAV1$, $CAV2$ and $CBAG$).

The conclusions of this analysis are the followings:

- In all the cases, the diversity analysis serve to improve all the statistics on accuracy and stability.

Table 17. Diversity of the ensemble of the classifiers.

Correlation	Q-Statistic	Disagreement	Double-Fault	Entropy	KW	Agreement
[a]0.49	0.87	0.11	0.04	0.14	0.05	0.37
[b]0.23	0.56	0.20	0.04	0.28	0.08	0.14

[a]For the 11 considered classifiers.
[b]For the 5 selected classifiers.

Table 18. Pairwise diversity of the selected classifiers. Bold faces indicate the most diverse pairs.

Pair	Correlation	Q-Statistic	Disagreement	Double-Fault
Histogram-Variogram	0.08	0.62	0.12	0.01
Histogram-Texture	0.16	0.71	0.09	0.01
Histogram-DWT	0.29	0.89	0.04	0.01
Histogram-Edges	0.07	0.51	0.39	0.01
Variogram-Texture	0.33	0.82	0.10	0.03
Variogram-DWT	0.39	0.92	0.07	0.02
Variogram-Edges	0.15	0.50	0.33	0.08
Texture-DWT	0.31	0.87	0.08	0.02
Texture-Edges	0.24	0.72	0.35	0.07
DWT-Edges	0.20	0.53	0.33	0.02

Table 19. Ensemble Learning algorithms performed with the final five selected classifiers.

Algorithm	Min	Max	Median	Mean	IQR	Std
*CAV*1	**97.00**	**99.50**	**98.25**	**98.25**	1.50	0.79
*CAV*2	**97.00**	99.00	**98.25**	**98.25**	**1.00**	**0.67**
CBAG	90.50	97.50	94.25	94.00	2.50	1.94

- The best result has been obtained for *CAV*1 and *CAV*2, both with a median accuracy of 98.25% which is very close to the one provided by the local histogram (98.80%). The stability of the results measured by the IQR and the standard deviation have improved the dispersion provided by the local histogram, being *CAV*2 the most stable.

These results can be improved by optimizing the parameters of these ensemble learning classifiers. We have performed some preliminary analysis of the optimization for *CAV*1 and *CAV*2 using a Particle Swarm Optimizer (RR-PSO).[12] The optimum weights that have been found for these two algorithms and the accuracy statistics are shown in Tables 20 and 21. In these tables, it can be observed the following:

- In the case of *CAV*1, the histogram, the edges and the variogram seem to be more important (bigger weights) than the texture and the DWT. The optimum number of candidates for this set of weights is five and no group substitution is needed. With this set of parameters we achieved a median accuracy close to 100%

Table 20. Set of optimized weights for *CAV*1 and *CAV*2 algorithms.

Algorithm	Histogram	Variogram	Texture	Edges	DWT	N_f	N_c	Group Subs
*CAV*1	99.00	72.68	58.29	77.72	50.00	—	5	No
*CAV*2	80.41	82.86	82.28	82.22	84.03	20	15	Yes

Table 21. Statistics for $CAV1$ and $CAV2$ algorithms with optimized weights.

Algorithm	Min	Max	Median	Mean	IQR	Std
$CAV1$	**99.00**	**100.00**	**99.50**	**99.50**	**0.50**	**0.46**
$CAV2$	98.70	**100.00**	99.30	99.20	0.60	0.55

(99.50%). The algorithm is also very stable (low values of IQR and standard deviation). Obviously this set of parameters is not unique.

- In the case of $CAV2$, all the five attributes have similar weights, in the range of 80% to 85%. In this case, the number of first candidates selected by the histogram (N_f) is 20, the number of candidates for the final consensus (N_c) is 15, and group substitution is needed. Also, with this set of parameters we achieved a median accuracy close to 100% (99.30%), being the algorithm very stable as well.

Further research will be devoted to the use of ensemble supervised learning for biometric application and face recognition using these approaches.

5. Conclusion

In this paper we have analyzed, for the face recognition problem, the performance of different types of image attributes: statistical, image analysis descriptors and spectral, both, acting globally over the whole image, or locally, by partitioning the image in different sectors. The classifiers are unsupervised and are based on nearest neighbor estimator using different norms defined on the space of attributes. We have proved numerically that the most performing classifier is based in the local histogram, followed by the local percentiles, and the global analysis via the DWT. All the spectral methods provide similar accuracies. The regional descriptors are the attributes that provide the lower accuracies. These individual classifiers use the L1 norm in the case of all the statistical attributes, texture analysis and regional descriptors, while the L2 norm provides higher accuracies for all the spectral attributes. The local analysis usually provides better results. Only the edges and the DWT perform better globally than locally. In any case, no single attribute is able to achieve systematically 100% accuracy over 100 different independent simulations.

In the second part of this paper, we had investigated the application of ensemble learning methodologies to the face recognition problem, combining the classifiers that have been analyzed in the first part. Ensemble learning and consensus prediction are based on the fact that majority voting using different weak classifiers have a higher accuracy (success probability) than any of the individual voters. We have introduced three different nonsupervised ensemble learning methodologies, named $CAV1$, $CAV2$ and $CBAG$. The main difference among them is the way of selecting the image candidates to perform the consensus. First, we have applied these algorithms using all the individual classifiers (11) based on the image attributes that were analyzed in the first part, showing that $CAV2$ provided the best performance, closely

followed by $CAV1$. $CBAG$ algorithm appeared to be less performing and more computationally expensive than $CAV1$ and $CAV2$. The mean accuracy for $CAV2$ was slightly lower than the one provided by the local histogram, nevertheless the stability was increased by consensus. The diversity analysis on the set of classifiers allowed us to select the five most diverse attributes to be used in the ensemble learning algorithms explained before. These attributes are: the local histogram, the local variogram, the local texture, the local edges and the global DWT. The use of this set of five uncorrelated classifiers, improved the accuracies and stability in the prediction obtained by the ensemble learning algorithms: $CAV1$ and $CAV2$ performed very close to the local histogram and the stability of the classification increased even more. These results were improved by optimizing the parameters on which these two classifiers depend, arriving to median accuracies closer to 100% and very low dispersions. Therefore, in the face recognition problem it is possible to build strong classifiers by assembling different simple classifiers based on diverse image attributes. Future research will be devoted to further improve this methodology using supervised ensemble learning techniques in order to achieve systematically perfect face recognition. Preliminary results have been presented in Ref. 13.

Acknowledgments

We would like to thank the anonymous reviewers whose critics and suggestions served in our opinion to greatly improve the quality of this manuscript.

Appendix A. Texture Analysis

The statistical moments inferred from the GLCM matrix used in this paper are:

- Contrast is a measure of local image variation, typically it means the intensity change between a pixel and its neighbor over the whole image.

$$\text{Cont} = \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j|^2 P_{d,\theta}(i,j). \tag{A.1}$$

  Contrast is zero for a constant image.
- Homogeneity measures the proximity of the distribution of the GLCM matrix to its diagonal.

$$\text{Hom} = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{P_{d,\theta}(i,j)}{1+|i-j|}. \tag{A.2}$$

- Squared energy

$$E_F = \|P_{d,\theta}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{d,\theta}(i,j)^2 \tag{A.3}$$

  which is the square of the Frobenius norm of the GLCM matrix.

- Correlation

$$\rho = \frac{\text{cov}(i,j)}{\sigma_i \sigma_j} = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{d,\theta}(i,j)\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j}, \tag{A.4}$$

where

$$\mu_i = \sum_{i=1}^{n}\sum_{j=1}^{n} iP_{d,\theta}(i,j),$$

$$\mu_j = \sum_{i=1}^{n}\sum_{j=1}^{n} jP_{d,\theta}(i,j),$$

$$\sigma_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} (i-\mu_i)^2 P_{d,\theta}(i,j),$$

$$\sigma_j^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} (j-\mu_j)^2 P_{d,\theta}(i,j).$$

- Entropy

$$\text{Ent} = -\sum_{i=1}^{n}\sum_{j=1}^{n} P_{d,\theta}(i,j)\log P_{d,\theta}(i,j). \tag{A.5}$$

Appendix B. Diversity Analysis

The diversity descriptors have been used in this paper are the following:

- Pairwise measure: these measures are defined between two classifiers, this means that in our case, for the Na considered classifiers, $Na(Na-1)/2$ diversity measures will be computed. A global diversity analyze can be obtained by averaging all the pairwise measures. The pairwise diversity measures used in this paper are correlation, Q-statistic, disagreement, and double-fault. Details about these measures can be consulted in Refs. 21 and 26. For two different classifiers C_i and C_j, let a be the number of images correctly classified by both classifiers, b the number of images correctly classified by C_i and misclassified by C_j, c the number of images correctly classified by C_j and misclassified by C_i, and d the number of images misclassified by both, such as $a+b+c+d=N$.

— Correlation aims to find the correlation between two classifiers and is defined as:

$$\rho_{i,j} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

The highest diversity is achieved when $\rho = 0$ and the lowest for $|\rho_{i,j}| = 1$.

— Q-statistic assesses the similarity of two classifiers outputs and is calculated as follows:

$$Q_{i,j} = \frac{ad - bc}{ad + bc}.$$

Similarly to the correlation measure, the maximum diversity is obtained for $Q_{i,j} = 0$ and the $-1 \leq Q_{i,j} \leq 1$.

— Disagreement is the ratio between the number of images correctly classified by one of the classifiers and misclassified by the other, to the total number of images.

$$D_{i,j} = \frac{b + c}{a + b + c + d} = \frac{b + c}{N}.$$

— Double-fault is the fraction of images misclassified by both classifiers.

$$DF_{i,j} = \frac{d}{N}.$$

- Nonpairwise measures: these refer to the global diversity of the whole set of considered classifiers. The following measures are used in this paper to perform the diversity analysis:

— Entropy: this measure assumes that the highest diversity is achieved if half of the classifiers misclassifies and the other half classifies correctly. Entropy is calculated as follows:

$$E = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N_a - [N_a/2]} \cdot \min(k_i, N_a - k_i),$$

where k_i is the number of classifiers that misclassify the image x_i and N_a is the total number of classifiers.

— Kohavi–Wolpert variance is related to the average disagreement measure over all classifiers, and is calculated as follows:

$$\mathrm{KW} = \frac{1}{NN_a^2} \sum_{i=1}^{N} k_i(N_a - k_i).$$

— Agreement estimates the strength of association among the classifiers.[32] Considering k_{ij} the number of classifiers that assign the image x_i to the class j, the agreement measure can be calculated as follows:

$$k = 1 - \frac{NN_a^2 - \sum_{i=1}^{N} \sum_{j=1}^{C} k_{ij}^2}{NN_a(N_a - 1) \sum_{j=1}^{C} p_j q_j},$$

where $p_j = \sum_{i=1}^{N} \frac{k_{ij}}{NN_a}$ and $q_j = 1 - p_j$, for $j = 1, \ldots, C$. k quantifies the level to which the classifiers agree in their decisions beyond any agreement that could occur by chance.

References

1. A. N. Akansu and R. Poluri, Walsh-like nonlinear phase orthogonal codes for direct sequence cdma communications, *IEEE Trans. Signal Process.* **55**(7) (2007) 3800–3806.
2. S. Aksoy and R. M. Haralick, Content-based image database retrieval using variances of gray level spatial dependencies, in *Proc. IAPR Int. Workshop on Multimedia Information Analysis and Retrieval* (1998), pp. 3–19.
3. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *Computer Vision ECCV '96*, eds. B. Buxton and R. Cipolla, Lecture Notes in Computer Science, Vol. 1064 (Springer Berlin Heidelberg, 1996), pp. 43–58.
4. M. H. Bharati, J. J. Liu and J. F. MacGregor, Image texture analysis: Methods and comparisons, *Chemometr. Intell. Lab. Syst.* **72**(1) (2004) 57–71.
5. P. J. Boland, Majority system and the Condorcet jury theorem, *Statistician* **38**(3) (1989) 181–189.
6. L. Breiman, Bagging predictors, *Mach. Learn.* **24**(2) (1996) 123–140.
7. J. Canny, A computational approach to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell.* **8**(6) (1986) 679–698.
8. K. Choi, K. A. Toh and H. Byun, A random network ensemble for face recognition, *Advances in Biometrics*, Lecture Notes in Computer Science, Vol. 5558 (Springer, 2009), pp. 92–101.
9. M. A. Dabbah, W. L. Woo and S. S. Dlay, Secure authentication for face recognition, in *IEEE Symp. Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007* April 2007, pp. 121–126.
10. H. Ebrahimpour and A. Kouzani, Face recognition using bagging KNN, *Int. Conf. Signal Processing and Communication Systems (ICSPCS)*, Australia, Gold Coast 2007, pp. 17–19.
11. J. L. Fernández-Martínez and A. Cernea, Numerical analysis and comparison of spectral decomposition methods in biometric applications, *Int. J. Pattern Recogn. Artif. Intell.* **28**(1) (2014) 1456001-1–1456001-33.
12. J. L. Fernández-Martínez and E. García-Gonzalo, Stochastic Stability and Numercal Analysis of Two Novel Algorithms of PSO Family: PP-PSO and RR-PSO, *Int. J. Artif. Intell. Tools* **21**(3) (2012) 405–423.
13. J. L. Fernández-Martínez, A. Cernea, E. García-Gonzalo, J. Velasco and B. Ketan Panigrahi, Aligned PSO for optimization of image processing methods applied to the face recognition problem, *Swarm, Evolutionary, and Memetic Computing (SEMCCO)*, Lecture Notes in Computer Science, Vol. 8297 (Springer Berlin Heidelberg, 2013).
14. R. A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* **7** (1936) 179–188.
15. P. Goovaerts, *Geostatistics for Natural Resources Evaluation*, Applied Geostatistics Series (Oxford University Press, Incorporated, 1997).
16. Z. M. Hafed and M. D. Levine, Face recognition using the discrete cosine transform, *Int. J. Comput. Vis.* **43**(3) (2001) 167–188.
17. M. Harandi, J. Taheri and B. C. Lovell, Ensemble learning for object recognition and tracking, in *Pattern Recognition, Machine Intelligence and Biometrics*, 1st edn. (Springer, 2011).
18. R. Jafri and H. R. Arabnia, A survey of face recognition techniques, *J. Inf. Process. Syst.* **5**(2) (2009) 41–68.
19. S. Kakarwal and R. Deshmukh, Wavelet transform based feature extraction for face recognition, *Int. J. Comput. Sci. Appl.* (IJCSA), Issue 1 (2010), 0974-0767.

20. M. Kirby and L. Sirovich, Application of the karhunen-loeve procedure for the characterization of human faces, *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(1) (1990) 103–108.
21. L. I. Kuncheva and C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Mach. Learn.* **51**(2) (2003) 181–207.
22. J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos and S. Z. Li, Ensemble-based discriminant learning with boosting for face recognition, *IEEE Trans. Neural Netw.* **17**(1) (2006) 166–178.
23. R. Mallipeddi and M. Lee, Ensemble based face recognition using discriminant PCA Features, *IEEE Congress on Evolutionary Computation (CEC)* 10–15 June 2012, pp. 1–7.
24. M. Najafi and M. Jamzad, An ensemble based learning for face recognition with similar classifiers, in *Proc. World Congress on Engineering and Computer Science*, San Francisco, USA, 24–26 October 2007.
25. A. Nait-Ali, Hidden biometrics: Towards using biosignals and biomedical images for security applications, *7th Int. Workshop on Systems, Signal Processing and their Applications (WOSSPA), 2011* May 2011, pp. 352–356.
26. R. Polikar, Ensemble based systems in decision making, *IEEE Circuits Syst. Mag.* **6**(3) (2006) 21–45.
27. W. K. Pratt, J. Kane and H. C. Andrews, Hadamard transform image coding. *Proc. IEEE* **57**(1) (1969) 58–68.
28. L. Rokach, Ensemble-based classifiers, *Artif. Intell. Rev.* **33**(1–2) (2010) 1–39.
29. R. Kh. Sadykhov, V. A. Samokhval and L. P. Podenok, Face recognition algorithm on the basis of truncated walsh-hadamard transform and synthetic discriminant functions, *FGR* (IEEE Computer Society, 2004) pp. 219–222.
30. R. E. Schapire, The strength of weak learnability, *Mach. Learn.* **5**(2) (1990) 197–227.
31. G. T. Shrivakshan and C. Chandrasekar, A comparison of various edge detection techniques used in image processing, *Int. J. Comput. Sci. Issues* **9**(5, No 1) (2012) 1694–0814.
32. K. Sirlantzis, S. Hoque and M. C. Fairhurst, Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition, *Appl. Soft Comput.* **8**(1) (2008) 437–445.
33. L. Sirovich and M. Kirby, Low-dimensional procedure for the characterization of human faces, *J. Opt. Soc. Am. A* **4**(3) (1987) 519–524.
34. M. Sonka, V. Hlavac and R. Boyle, *Image Processing, Analysis, and Machine Vision* (Thomson-Engineering 2007).
35. G. N. Srinivasan and G. Shobha, Statistical texture analysis, in *Proc. World Acad. Sci., Eng. Tech.* **36** (2008) 1264–1269.
36. M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cognit. Neurosci.* **3**(1) (1991) 71–86.
37. S. E. Umbaugh, *Computer Vision and Image Processing: A Practical Approach Using CVIPtools* (Prentice Hall Professional Technical Reference, 1998).
38. X. Wang and X. Tang, Random sampling for subspace face recognition, *Int. J. Comput. Vis.* **70**(1) (2006) 91–104.
39. T. Windeatt, Ensemble neural classifier design for face recognition, *2007 (ESANN'2007), European Symp. Artificial Neural Networks*, Bruges Belgium, 25–27 April, pp. 373–378.
40. J. Yang, D. Zhang, A. F. Frangi and J. Y. Yang, Two-dimensional pca: A new approach to appearance-based face representation and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **26** (2004) 131–137.
41. H. Yu and J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, *Pattern Recogn.* **34** (2001) 2067–2070.

42. J. Zhang, L. He and Z. H. Zhou, Ensemble-based discriminant manifold learning for face recognition, in *Advances in Natural Computation*, Lecture Notes in Computer Science, Vol. 4221 (Springer, 2006), pp. 29–38.

Ana Cernea received her B.Sc. degree in Mathematics from the University of Bucharest, Romania, in 1989 and her B.Sc. degree in Computer Science from the University of Oviedo, Spain. Currently, she is completing her Ph.D. on Image Processing Optimization. Her field of research also include global optimization techniques and digital signal processing with applications in biometrics.

Juan Luis Fernández-Martínez received his Ph.D. in Mining Engineering from the University of Oviedo (Spain) in 1994 when he joined the Mathematics Department of University of Oviedo in 1994 where he is Professor in Applied Mathematics. During 2008–2010, he was a Visiting and Research Professor at UC Berkeley-Lawrence Berkeley Laboratories and Stanford University. He currently works in inverse problems and uncertainty quantification applied to the oil industry and also in biomedicine (cancer, genomics and proteomics).

A.3. Exploiting the Uncertainty Space of Ensemble Classifiers in Face Recognition

Authors:
Ana Cernea
Juan Luis Fernández Martínez

International Journal of Pattern Recognition and Artificial Intelligence
In Revision

International Journal of Pattern Recognition and Artificial Intelligence

Exploring the Uncertainty Space of Ensemble Classifiers in Face Recognition

J. L. FERNÁNDEZ-MARTÍNEZ*

Mathematics Department. Oviedo University
C/ Calvo Sotelo s/n, 33007 Oviedo, Spain.
jlfm@uniovi.es

ANA CERNEA

Mathematics Department. Oviedo University,
C/ Calvo Sotelo s/n, 33007 Oviedo, Spain.
cerneadoina@uniovi.es

In this paper we present a supervised ensemble learning algorithm, called *SCAV1*, and its application to face recognition, which takes advantage of the uncertainty space of the ensemble classifiers. Its design includes six different nearest-neighbor classifiers that are based on different and diverse image attributes: histogram, variogram, texture analysis, edges, bidimensional discrete wavelet transform and Zernike moments. In this approach each attribute, together with its corresponding type of the analysis (local or global), and the distance criterion (p-norm) a different individual nearest-neighbor classifier. The ensemble classifier *SCAV1* depends in a set of parameters: the number of candidate images used by each individual method to perform the final classification and the individual weights given to each individual classifier. These parameters are optimized/sampled using a supervised approach via the RR-PSO algorithm. The final classifier exploits the uncertainty space of *SCAV1* and uses majority voting (Borda Count) as a final decision rule. We show the application of this methodology to two different image databases (ORL and PUT), obtaining very high and stable accuracies (100% median accuracy and null interquartile range). In conclusion, exploring the uncertainty space of ensemble classifiers provides optimum results and seems to be the appropriate strategy to adopt for face recognition.

Keywords: Face recognition; statistical attributes; spectral attributes; segmentation features; ensemble supervised learning.

1. Introduction

Automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision. One of the main challenges in dealing with very large databases of images is establishing low-dimensional feature representations having enough discriminatory power to perform high accuracy classification and/or prediction with the corresponding uncertainty quantifi-

*Mathematic Department. Oviedo University, C/ Calvo Sotelo s/n 33007 Oviedo, Spain.

cation. Different techniques based on image attributes have been proposed in the literature[29]. Spectral attributes are the most common in face recognition. These attributes are obtained by projecting the images onto the corresponding spectral basis. These methods can be divided into two main categories: covariance-based and covariance-free methods. Covariance-based spectral methods involve the diagonalization of different image covariance and correlation matrices. PCA[48], 2DPCA[52], ICA[2] and Fisher's LDA[3] belong to this category. The spectral basis provided by these covariance-based methods are computationally expensive since these methods generally use all the poses available in the database. Therefore, these reduced basis should be updated depending on the poses that are available on the database. The covariance-free methods involve different orthogonal transformations defined over the set of single images. To this category belongs the discrete sine and cosine transform (DST/DCT)[24], the discrete wavelet transform (DWT)[30], the discrete Walsh-Hadamard transform (DWHT)[42], etc. Covariance-free methods depend on individual images, their calculation is very fast and can be pre-computed.

A numerical comparison between these methods has been recently performed by Fernández-Martínez and Cernea[13], showing that the most crucial parameter in the accuracy of the classification is the energy cut-off used to perform the image energy compression and the model reduction. The diversity analysis of the individual nearest-neighbor classifiers induced by these spectral attributes, showed that all these methods provide redundant information in the classification[8]. Otherwise said, the corresponding nearest-neighbor classifiers are not diverse. As a result of this analysis, the bidimensional discrete wavelet transform (DWT) was chosen for ensemble learning, since this spectral method shows the highest individual accuracy and the lowest computational cost. Also, no individual classifier based on spectral attributes was able to provide systematically 100% median accuracy over a complete set of independent simulations[13]. Due to this fact Cernea and Fernández-Martínez[8] proposed to use other type of image attributes, such as, statistical and texture and regional image descriptors. As result of this analysis two different ensemble algorithms based on nearest-neighbor classifiers were proposed (*CAV1* and *CAV2*).

CAV1 and *CAV2* were based on 5 main diverse classifiers: histogram, variogram, texture analysis, edges, and bidimensional discrete wavelet transform. In this approach each attribute, together with its corresponding type of the analysis (local or global), and the distance criterion (p-norm) defined a different nearest-neighbor classifier. These ensemble learning algorithms weighted the scores of the different images found by these five different nearest-neighbor classifiers. In a first attempt these algorithms were used under the modality of unsupervised learning by considering the corresponding accuracies of each individual attribute as weights for *CAV1* and *CAV2*. The predictive classification accuracies of *CAV1* and *CAV2* were dramatically improved by performing the optimization of these weights and the number of image candidates used for consensus, using a powerful particle swarm optimizer[8]. This idea is expanded in this manuscript by centering our attention in *CAV1*, and proposing a supervised version, *SCAV1*, that is based on exploring and

exploiting the uncertainty space of *CAV1*.

The structure of this paper is as follows: in Section 2 we introduce the face recognition problem via Supervised Ensemble Learning, analyzing also the state of the art in the field of face recognition. Section 3 briefly describes the attributes for ensemble learning, introducing the Zernike moments nearest-neighbor classifier, and analyzing its diversity with respect to the other five nearest-neighbor classifier. Section 4 explains how to use the uncertainty space in ensemble learning for face recognition. Section 5 briefly describes the RR-PSO algorithm[18] used to optimize the *CAV1* parameters. Section 6 presents *SCAV1* and the numerical results obtained over 10 independent simulations of the ORL and PUT face databases. Finally, we outline the main conclusions and the future research work.

2. Automatic Image Recognition via Supervised Ensemble Learning

The automatic image recognition problem consists in classifying a given probe image I providing a database of training images. Given a database of training images (the learning database)

$$Bd = \left\{ I_k \in \mathcal{M}_{(n,m,c)} \left(\mathbb{N} \right) : k = 1, \ldots, N \right\}, \tag{1}$$

organized into q classes:

$$C = \left\{ C_k \in \{1, 2, ..., q\}, k = 1, \ldots, N \right\}, \tag{2}$$

and a new incoming image $I \notin Bd$, the problem consists in estimating its class $C_I^* \in C$, according to the different poses of I that are at disposal in Bd.

In this definition $\mathcal{M}_{(n,m,c)}$ is the space of RGB color images (if $c = 3$), or gray-scale images (if $c = 1$) of size $m \times n$, and the learning database typically contains N_p poses for each class $C_i \in C$ $(i = 1, ..., q)$.

In previous research[13,8], we have used to perform the class prediction an unsupervised learning algorithm, L^*, which is a functional defined over the set of images $\mathcal{M}_{(n,m,c)}$ into the set of classes C:

$$L^* : \mathcal{M}_{(n,m,c)} \to C, \quad L^*(I_j) = C_j. \tag{3}$$

L^* is based on a nearest-neighbor criterion, that is, finding the image $I_k \in Bd$ such as:

$$d\left(I, I_k\right) = \min_{I_j \in Bd} d\left(I, I_j\right), \tag{4}$$

where d is a suitable distance (or norm) defined over $\mathcal{M}_{(n,m,c)}$. Typically, these distances are defined over the different attribute spaces, as follows:

$$d_k\left(I_i, I_j\right) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \tag{5}$$

where $\mathbf{v}_i^k \in \mathbb{R}^{s_k}$ represents the feature vector of length s_k corresponding to image I_i according to the attribute k, and p is the norm (usually $p = 1$ or $p = 2$ depending on

the method) defined over the k-th attribute vector space ($\mathbb{R}^{s_k}$). Once the nearest-neighbor image has been found, then, its class is assigned to the incoming image:

$$L^*(I) = L^*(I_k) = C_k. \tag{6}$$

The success of the unsupervised classification depends on the implicit relationship existing between the k-attribute, the p-norm that is used to define the distance criterion for each classifier, and the class information. In the case of supervised classification the learning method $L^*(I_j; \mathbf{m})$ depends also on a set of parameters $\mathbf{m}$ that has to be tuned using class information from a subset of images of the database (testing database). The aim of this paper is to analyze the performance of such type of algorithms in the context of ensemble learning methodologies, by exploiting the structure of its uncertainty space.

Ensemble learning methodologies try to take advantage of the intrinsic uncertainty of any classification problem[17,41], that is, considered as a optimization problem there exist different classifiers that provide the same predictive accuracy. The main idea in ensemble learning is to be able to combine and weigh several single classifiers in order to obtain strong classifier that outperforms every one of them[43]. Ensemble classification is also related to majority vote decisions[5], that are based on the fact that the probability of being correct for a majority of independent voters is higher than the probability of any single voter. Schapire (1990)[45] used this fact in machine learning to prove that it is possible to design strong classifier by ensemble learning using weak classifiers. In the present case, in the context of face recognition, ensemble learning consists in building a set of diverse classifiers based on different single image attributes. It has been shown that a necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is to be accurate and diverse[26]. Accurate classifiers are those with error rates lower than random guessing, while diversity refers to the fact that different classifiers of the ensemble misclassify different samples.

The use of ensemble learning methodologies in face recognition is obviously not new, nevertheless most of the research has been focussed in applying well-established ensemble learning methodologies, such as bagging, or bootstrap replicate of the original training set[6], through the use of cross-validated committees[40], or finally through the AdaBoost algorithm[20]. AdaBoost manipulates the training examples to generate multiple classifiers, by changing the weights on training examples that were misclassified. The final classifier is constructed by a weighted vote of these individual classifiers, according to their accuracies on the training set. Also, Hamzeloo et al.(2012)[25] proposed to adopt the ensemble learning formalism and build a weighted nearest-neighbor ensemble classifier for solving general machine learning problems. To the best of our knowledge, this kind of algorithm has not been used in face recognition yet.

The following references are important for ensemble learning in face recognition:

- Guo and Zhang[23] introduced boosting for face recognition. For that purpose

they used Adaboost which is an adaptative algorithm to boost a sequence of classifiers by updating dynamically their weights according to the errors in previous learning. A similar methodology was proposed by Najafi and Jamzad[39], increasing the diversity of the ensemble by training on different subsets of the training set.

- Chawla and Bowyer (2005)[9] explored the use of ensembles in face recognition applying the random subspace method introduced by Ho[28]. This work showed that the ensemble methodology outperforms any single nearest-neighbor classifier learned using the whole database.
- Lu et al (2006)[36] proposed a novel ensemble-based approach to boost performance of traditional Linear Discriminant Analysis (LDA)-based methods used in face recognition. Fan and Lu[12] proposed a fast statistical machine learning ensemble algorithm for face detection.
- Ebrahimpour and Kouzani (2007)[11] used K-nearest-neighbors and bagging.
- Zang et al (2006)[53] used the local linear embedding method to find different low-dimensional spaces for face recognition. The final classification was performed by majority voting.
- Windeatt (2007)[50] proposed an Ensemble Neural Classifier for face recognition, by selecting different multilayer perceptron classifiers based only on their performance on the training set, and used them in ensemble prediction obviating the need for validation.
- Le (2011) proposed a hybrid model combining AdaBoost and neural networks[35]. This algorithm was applied for image classification with features resulted from the faces projection into the different spaces, such as face space, eyes space, face-no-mouth space, and the final result was obtained by ensemble prediction.
- Mallipeddi et al (2012)[37] used an ensemble based face recognition method using discriminant PCA features obtained from different subsets of the training database by bagging. The decision was also done by majority voting. Similar methodologies were also proposed by other authors using boosting or bagging approaches based on individual classifiers[36,10,27,38].

The methodology proposed in this paper consists in generalizing the algorithm *CAV1* introduced by Cernea and Fernández-Martínez[8] to the case of supervised learning, and exploring its uncertainty space to perform majority voting. The new algorithm, named *SCAV1*, is composed of two main steps:

- Step 1: The parameters needed for *SCAV1* are optimized using RR-PSO[18] and different random bags of the training database. For that purpose the database is divided in training, testing and validation. The testing database is used for optimizing the weights of these classifiers, and the validation database to established the generalization of these sets. As a result of this process, only the sets of weights with smaller validation error (higher validation accuracy) are selected, and their posterior histograms are analyzed.

These posterior histograms provide an approximative idea of the uncertainty space of the ensemble classifier *SCAV1*, that is, the set of *SCAV1* parameters with very high accuracies for a representative sample of the training database. Each of these *SCAV1* ensemble classifiers will be composed of six different individual nearest-neighbor classifiers based on the image attributes that are described in section 3. Each classifier selects the N_c images closest to the new incoming image, and the decision is made by consensus. The number of image candidates, N_c, and the weights of any individual classifier, are parameters that will be optimized in this first step.

- Step 2: Once this first step is performed, the final classification of a new incoming image, is made by consensus using the different ensemble learning *SCAV1* classifiers that have been sampled at step 1. Finally the consensus is made taking into consideration the 3 first image candidates provided by each individual *SCAV1* classifier. The number of image candidates (3 in this case) could also be a target for optimization. This methodology also produces a set of other candidates with slightly lower score than the solution that has been adopted. These candidates might be used to perform the uncertainty analysis of the face recognition problem, that is, to identify the subjects in the database that are similar to the new incoming face image.

3. Attributes for supervised ensemble learning

In this section we provide a brief description of the attributes that are used in this paper: color histogram, variogram, texture analysis, edge detection, discrete wavelet transform and Zernike moments. A more detailed description can be found in previous research work concerning unsupervised learning in biometry[8]. Also, we introduce the Zernike moments nearest-neighbor classifier. For each classifier we provide the type of analysis (global or local), the p-norm that is used and its individual accuracy given by a set of 100 independent simulations on the ORL database[8]. In the case of global analysis the attribute features are calculated over the whole size of the image. In the case of local features, this analysis is performed by dividing the images into 8×4 blocks. For each block the local attributes are computed and the final feature vector is formed by merging all the local attributes into an unique vector, always computed in the same order. All these attributes are very fast to be computed and can be precalculated for all the images in the learning database.

3.1. *Color Histogram*

The image histogram describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image: contrast, brightness, biomodality, etc. For a gray-scale digital image I the histogram represents the discrete probability distribution of the gray-levels in the image. In

the case of RGB images the same analysis is performed considering the color channels I_R, I_G and I_B independently, and merging all the channels histograms. The histogram nearest-neighbor classifier is used in the local analysis with the L_1 norm, providing a median accuracy of 98.80%[8].

3.2. *Variogram*

The variogram describes the spatial distribution in each color channel. For a given value of vector h the variogram measures the dissimilarity between all pairs of values separated by vector h. The omnidirectional p-variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^2. \tag{7}$$

To compute the variogram each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$. The variogram nearest-neighbor classifier is used in the local analysis with the L_1 norm, providing a median accuracy of 93.00%[8].

3.3. *Texture Analysis*

Texture analysis consists in analyzing regular repetitions of a pattern on images though the computation of the gray level co-occurrence matrix which is based on the joint probability distributions of pairs of pixels located at a distance d apart (measured in number of pixels) along a given direction θ[4]. Typically the GLCM is calculated for different pairs of d and θ. Texture is a local attribute for describing and quantifying the spatial arrangement of color or intensities in an image or in a selected region. The statistical moments inferred from the GLCM matrix used in this paper are: contrast, homogeneity, energy, correlation and entropy[4]. In the present case to perform the texture analysis we have used a lag $d = 1$ for the directions $0, 45, 90$ and 135. The texture nearest-neighbor classifier is used in the local analysis with the L_1 norm, providing a median accuracy of 92.50%[8].

3.4. *Edges Detection*

Edges describe set of pixels with abrupt changes in intensity[51]. Determination of the edge map is related to gradient calculation of the pixel intensities through different convolution masks to approximate the gradient by differences schemes. In this case the algorithm uses the Canny edge detection operator[7], followed by the bidimensional Discrete Cosine Transform (DCT) to the edge map to reduce further the dimension and decreasing the impact of noise. Denoting by E_d the DCT transform of the edge map, the feature vector is formed by the $n_1 \times n_2$ submatrix

of E_d, $E_d(1:n_1, 1:n_2)$ ($n_1 = n_2 = 10$). The edge nearest-neighbor classifier is used in the global analysis with the L_2 norm, providing a median accuracy of 91.00%[8].

3.5. *Discrete Wavelet Transform (DWT)*

Previous research work has shown that all the spectral attributes provided similar accuracies and do not show a very high diversity in the classification, being the discrete wavelet transform, a free-covariance spectral method that provided the highest accuracy with the lowest computational cost[8].

The discrete wavelet transform (D_{WT}) of an image $I \in \mathcal{M}(m,n)$ is defined as follows:

$$D_{WT} = U_W^T I V_W, \tag{8}$$

where U_w and V_w are two orthogonal matrices constructed as follows:

$$U_W = \left[\frac{H}{G}\right]_m^T, V_W = \left[\frac{H}{G}\right]_n^T, \tag{9}$$

where H represents a low pass or averaging portion of the wavelet filter, and G is the high pass or difference portion. To produce the feature vector for face recognition we have used the blur produced by applying the DWT recursively. The DWT nearest-neighbor classifier is used in the global analysis with the L_2 norm, providing a median accuracy of 95.50%[8].

3.6. *Zernike Moments*

Zernike moments are based on the polynomials of the same name that are orthogonal on the unit ball. These polynomials were proposed by F. Zernike in 1934 and have wide application in optometry to describe aberrations of the cornea from an ideal spherical shape. They are also used in image processing as orthogonal basis to represent properties of an image with no redundancy or overlap of information. Zernike moments have been used in face recognition by Foon et al. (2004)[19].

Zernike moments are widely used as image descriptors due to their properties of orthogonality and rotation invariance. An orthogonal basis of complex functions defined over the set of square integrable functions was introduced by Zernike in 1934. $V_{pq}(x,y)$ is the (p,q) order Zernike function, defined as follows:

$$V_{pq}(x,y) = R_{pq}(\rho)e^{i\theta}, \text{ with } x^2 + y^2 \leq 1, \tag{10}$$

where ρ, θ are the polar coordinates of the pixel (x,y), $R_{pq}(\rho)$ is a radial polynomial in ρ of degree $p \geq 0$ containing no power ρ less than $|q|$, where q is the repetition index for a given order p and has to satisfy $|q| \leq p$, with $p - |q|$ an even number[32].

It can be shown that square integrable functions, $f(x,y)$, defined on the unit disk can be represented through the orthogonal Zernike functions as follows:

$$f(x,y) = \sum_{p=0}^{\infty} \sum_{q=-p}^{p} \rho_p A_{pq} V_{pq}(x,y), \tag{11}$$

where ρ_p is the normalization factor and A_{pq} is the Zernike moment of order p with repetition q, and it is defined as follows:

$$A_{pq} = \int\int_D f(x,y)V_{pq}(x,y)dxdy. \tag{12}$$

Table 1 shows the statistics obtained for the Zernike nearest-neighbor classifier with a basis of polynomials of order 11. The best results are obtained in the case of local analysis with the L_2 norm (95.00% of median accuracy).

Table 1. Zernike Moments: statistics for the global and local analysis, with the L_1, L_2 norm and the cosine criterion.

Type of analysis	min	max	median	mean	IQR	std
Global L_1	83.00	93.50	87.50	87.75	3.00	2.05
Global L_2	82.00	92.50	87.50	87.42	2.50	2.13
Global cosine	79.50	90.50	85.00	84.93	3.00	2.20
Local L_1	89.50	97.50	94.00	94.10	2.00	1.64
Local L_2	**90.00**	**98.00**	**95.00**	**94.61**	**2.00**	**1.59**
Local cosine	87.50	96.50	92.50	92.51	2.00	1.72

Table 2 shows the pair-wise diversity measures computed over the six selected classifiers compared to the five original classifiers used in Cernea and Fernández-Martínez (2014)[8]. The introduction of the Zernike moments slightly decreases the quality of the diversity statistics, since Zernike moments can be considered as an additional texture descriptor. Besides, if Zernike moments are considered within the texture descriptors, the diversity statistics will be remain similar to the diversity measures that were originally obtained using the above mentioned five nearest-neighbor classifiers.

Table 2. Diversity of the ensemble of the classifiers.

Correlation	Q-statistic	Disagreement	Double-fault	Entropy	KW	Agreement
[a]0.28	0.67	0.17	0.04	0.23	0.07	0.17
[b]0.23	0.56	0.20	0.04	0.28	0.08	0.14

[a]For the 6 considered classifiers
[b]For the selected classifiers, without Zernike moments.

3.7. *Final List of Nearest-Neighbor Classifiers*

Table 3 shows the list of attributes that have been selected in this paper for ensemble learning applied to the face recognition problem. For each attribute we specify their corresponding type of analysis (global or local) and the p-norm that is used in building the corresponding individual nearest-neighbor classifiers. We also recall

their individual median accuracies calculated over a set of 100 independent simulations of the ORL database. These accuracies were used in *CAV1* as weights to build the ensemble classifier via majority voting.

Table 3. List of NN classifiers used to build *SCAV1*.

NN Classifier	Type of Analysis	Norm	Median Accuracy
Histogram	Local	L_1	98.80
Variogram	Local	L_1	93.00
Texture	Local	L_1	92.50
Edges	Global	L_2	91.00
DWT	Global	L_2	95.50
Zernike	Global	L_2	95.00

4. Uncertainty space in ensemble learning for face recognition

In the context of face recognition, supervised classification consists in defining a classifier $L^*(I;\mathbf{m})$, which is a functional defined over the set of images $\mathcal{M}_{(n,m,c)}$ into the set of classes C:

$$L^* : \mathcal{M}_{(n,m,c)} \to C, \quad L^*(I_j;\mathbf{m}) = C_j, \tag{13}$$

and also depends on a set of parameters $\mathbf{m}$. These parameters ($\mathbf{m}$) have to be tuned using class information from a subset of images of the database (testing database), by solving the following optimization problem: finding $\mathbf{m}$ such as it is achieved the minimum of the cost function:

$$\phi(\mathbf{m}) = \|\mathbf{L}^*(T;\mathbf{m}) - \mathbf{c}_{true}\|_q. \tag{14}$$

In this definition $T = \{I_1, I_2, ..., I_s\}$ is a set of s testing images that have been selected from the training database; $\mathbf{L}^*(T;\mathbf{m}) = (L^*(I_1;\mathbf{m}), ..., L^*(I_s;\mathbf{m}))$ is their corresponding class prediction according to $\mathbf{m}$; $\mathbf{c}_{true} = (c_1, c_2, ..., c_s)$ the true classes of the images contained in T, and q the norm used to define the misfit (generally $q = 2$).

It has been shown[18] that in any nonlinear inverse or optimization problem, the topography of the cost function $\phi(\mathbf{m})$ in the region of low misfits (or high prediction accuracies), correspond to one or several disconnected flat (with null gradients) elongated curvilinear valleys. Uncertainty estimation of the classifier $L^*(I;\mathbf{m})$ consists in sampling the set of equivalent parameters $\mathbf{m}$ of this classifier:

$$V_{tol} = \{\mathbf{m} : \|\mathbf{L}^*(T;\mathbf{m}) - \mathbf{c}_{true}\|_q \leq tol\}. \tag{15}$$

Once this sampling is performed, typically using Monte Carlo methods or global optimization techniques, we have at disposal a set of parameters vectors:

$$M_{tol} = \{\mathbf{m}_k : k = 1, ..., N_s\}, \tag{16}$$

that induce a set L_{tol} of equivalent classifiers:

$$L_{tol} = \{L^*(I; \mathbf{m}_k) : k = 1, ..., N_s\}, \tag{17}$$

in the sense that

$$\|\mathbf{L}^*(T; \mathbf{m}_k) - \mathbf{c}_{true}\|_q \leq tol : \forall k = 1, ..., N_s. \tag{18}$$

The supervised face recognition problem with uncertainty estimation consists in, giving an incoming image I_{new}, applying the set of ensemble classifiers in L_{tol} to perform the consensus (or majority voting) classification, and obtaining the best solution and the set of other possible solutions given by these classifiers.

In the present manuscript, L_{tol} is composed of a set of *SCAV1* ensemble learning algorithms[8]. This set is diverse by construction, since the different classifiers misclassify different images. Also, although each ensemble classifier in L_{tol} has its own ROC curve[49], that is closer to a perfect ROC than any of the individual classifiers of the ensemble. Then, the consensus classifier based on the ensemble classifiers in L_{tol} will have a perfect ROC curve (step function in the origin or no misclassification). To find the L_{tol} set, we will use in this paper the *RR*-PSO algorithm that is described in the next section.

5. Particle Swarm Optimization and RR-PSO

Particle swam optimization algorithm[33] is a global optimization algorithm inspired in the behavior of bird flocks and fish schools searching for food. A swarm of particles (or models) explores the space of possible solutions in order to optimize a cost function.

The particle swarm algorithm applied to optimization problems is very simple: individuals, or particles, are represented by vectors whose length is the number of degrees of freedom of the optimization problem. To start, a population of particles is initialized with random positions ($\mathbf{x}_i^0$) and velocities ($\mathbf{v}_i^0$). The same objective function is used to compute the objective value of each particle. As time advances, the position and velocity of each particle is updated as a function of its objective function value and of the objective function values of its neighbors. At time-step $k+1$, the algorithm updates positions $\left(\mathbf{x}_i^{k+1}\right)$ and velocities $\left(\mathbf{v}_i^{k+1}\right)$ of the individuals as follows:

$$\begin{aligned} \mathbf{v}_i^{k+1} &= \omega \mathbf{v}_i^k + \phi_1(\mathbf{g}^k - \mathbf{x}_i^k) + \phi_2(\mathbf{l}_i^k - \mathbf{x}_i^k), \\ \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k + \mathbf{v}_i^{k+1}, \end{aligned}$$

with

$$\phi_1 = r_1 a_g,\ \phi_2 = r_2 a_l,\ r_1, r_2 \in U(0,1)\ \omega, a_l, a_g \in \mathbb{R},$$

where $\mathbf{l}_i^k$ is the $i-$th particle's best position, $\mathbf{g}^k$ the global best position on the whole swarm, $\phi_1,\ \phi_2$ are the random global and local accelerations, and ω is a real constant called inertia weight. Finally, r_1 and r_2 are random numbers uniformly

distributed in $(0,1)$, to weight the global and local acceleration constants, a_g and a_l.

PSO is the particular case for $t=k$ and $\Delta t=1$ of the GPSO algorithm[14]:

$$\begin{array}{l} v(t+\Delta t)=(1-(1-\omega)\,\Delta t)\,v\,(t)+\phi_1\Delta t\,(g\,(t)-x\,(t))+\phi_2\Delta t\,(l\,(t)-x\,(t))\,, \\ x(t+\Delta t)=x(t)+v(t+\Delta t)\Delta t. \end{array} \tag{19}$$

This model was derived using a mechanical analogy: a damped mass-spring system with unit mass, damping factor, $1-\omega$, and total stiffness constant, $\phi=\phi_1+\phi_2$, the so-called PSO continuous model:

$$\begin{cases} x''(t)+(1-\omega)\,x'(t)+\phi x(t)=\phi_1 g\,(t)+\phi_2 l\,(t)\,,\ t\in R, \\ x(0)=x_0, \\ x'(0)=v_0. \end{cases} \tag{20}$$

In this case $x(t)$ stands for the coordinate trajectory of any particle in the swarm. In this model particles interact through the local and global attractors, $l\,(t)$, and $g\,(t)$. In this model mean particle trajectories oscillate around the particle position[14]:

$$\mathbf{o}_i(t)=\frac{\phi_1\mathbf{g}(t)+\phi_2\mathbf{l}_i(t)}{\phi_1+\phi_2}. \tag{21}$$

Full stochastic analysis of the PSO continuous and discrete models (GPSO) has performed by Fernández Martínez and García Gonzalo[14,15,16]. This analysis shed light about the relationship between PSO convergence and the PSO parameters tuning: good PSO parameters are located close to the border of the second order stability region for most of the members of the PSO family. Using the above mentioned mechanical analogy Fernández-Martínez and E. García Gonzalo also derived a full family of Particle Swarm optimizers[15,18,21]. The most performing algorithm of the family in terms of the balance of its exploration/exploitation capabilities is the algorithm named RR-PSO, that is obtained by adopting regressive discretizations in acceleration and in velocity of the PSO continuous model (20). The RR-PSO algorithm is:

$$\begin{array}{l} v(t+\Delta t)=\dfrac{v(t)+\phi_1\Delta t\,(g(t)-x\,(t))+\phi_2\Delta t\,(l\,(t)-x\,(t))}{1+(1-\omega)\,\Delta t+\phi\Delta t^2} \\ x(t+\Delta t)=x(t)+v(t+\Delta t)\Delta t,\ t,\ \Delta t\in\mathbb{R} \\ x\,(0)=x_0, \\ v\,(0)=v_0. \end{array} \tag{22}$$

The RR-PSO first and second order stochastic stability regions are unbounded, and it has been numerically shown that performing parameters sets (in terms of exploration and convergence) are concentrated around the line $\overline{\phi}=3\left(\omega-\dfrac{3}{2}\right)$, mainly for inertia values $\omega>2$. This line is independent of the cost function that is

optimized, and remains invariant when the number of parameters of the optimization function increases. Furthermore, it is located in a zone of medium attenuation and very high frequency swarm of trajectories. This last property confers to this algorithm a good balance between exploration and exploitation, since this feature allows for a very efficient and explorative search around the oscillation center of each particle in the swarm.

RR-PSO will be used in this paper to sample the set of parameters $\mathbf{m}$ for *CAV1*, providing the highest predictive accuracies for the face recognition problem on the testing database (M_{tol} set).

6. The consensus algorithm *CAV1* revisited and its supervised version *SCAV1*

6.1. CAV1

In this section we revisit the *CAV1* algorithm, introduced in[8]. Furthermore, the supervised learning approach presented in this paper is based on the ensemble classification involving different supervised *CAV1* classifiers, that are deduced by exploring the uncertainty space of *CAV1*.

The classification of a new incoming image by any individual classifier is performed as in Cernea and Fernández-Martínez (2014)[8]:

(1) From every classifier shown in Table 3, we retain the first N_c images that are closer to I according to the criterium specified for each individual classifier (p-norm or cosine).
(2) Based on this analysis a matrix $M \in M_{Nc \times Na}$ is built, containing the N_c image candidates for each of the N_a attributes ($N_a = 11$ is the total number of attributes), that is, $M(i,j)$ is the j^{th} classified image by the i^{th} attribute (classifier).
(3) Once M is computed, the array $\mathbf{c} \in \mathbb{R}^{Nd}$ containing the N_d *different* candidates (without repetition) is formed. For any image candidate $I_c \in \mathbf{c}$, its score according to the set of positions (i,j) in matrix M is calculated as follows:

$$score(I_c) = \sum_{k=1}^{p} P(I_k(i,j)), \tag{23}$$

with:

$$P(I_k(i,j)) = \mathbf{w}(i) \cdot S_c(j), \tag{24}$$

where $P(I_k(i,j))$ is the score of image $I_k(i,j)$, that represents any of the p images instances of I_c in matrix M and p the number of repetitions. $\mathbf{w}(i)$ is a trust factor assigned to the i^{th} attribute (classifier). $S_c(j)$ is the linear decreasing score function used to rank the different candidates $I_k(i,j)$ according

to their position j in matrix M:

$$S_c(j) = S_{max} - \frac{S_{max} - S_{min}}{N_c - 1}(j-1), \tag{25}$$

$$\tag{26}$$

where S_{max}, S_{min} are the maximum and minimum scores assigned to the image candidates. After calculating the scores for all images in matrix M, the final classification of the test image I is performed by selecting the image with the highest score among all the candidate images in M. This classification method can be applied performing group substitution in matrix M before the final total score calculation. In this case, the score is calculated for each of the groups in matrix M.

6.2. SCAV1

For the supervised classification approach, the parameters to be tuned are the number of image candidates (N_c) and the weights vector ($\mathbf{w}$). Previous research[8] has shown that setting the weights vector $\mathbf{w}$ to the individual accuracies of these nearest-neighbor (NN) classifiers (shown in table 3), the best results were obtained with $N_c = 20$ and group substitution, being the median accuracy 98.25%. We have also shown that optimizing all these parameters the accuracy increased to 99.50%. In this section we explore further this result.

The design of the new supervised version *SCAV1* includes the optimization of the parameters $\mathbf{w}$ and N_c, using the RR-PSO algorithm, presented in the previous section. This optimization provides a set M_{tol} of parameters $\mathbf{m}$ that induce a set of classifiers L_{tol} which produce a misfit lower than a given threshold tol, as defined in Section 4.

The classification of a new incoming image by *SCAV1* is performed as follows:

(1) For each parameters vector $\mathbf{m}_k \in M_{tol}$, a consensus classification by *CAV1* is carried out and the first N_{best} images are retained. We obtain a matrix of candidates M_c which contains N_s rows of N_{best} images, where N_s is the number of optimized models of parameters in M_{tol}. In the present case N_{best} has been set to 3.
(2) A consensus classification by Borda Count is performed over the matrix of candidates M_c built in the previous step, and the best image is selected.
(3) The class of this selected image image is assigned to the incoming image.

Figures 1 and 2 show a sketch for the ensemble algorithm *SCAV1*.

6.3. *Posterior analysis and numerical results for* SCAV1

The numerical experiments were performed over ten independent simulations of the ORL and PUT face databases[44,31]. In the case of the ORL database, in each simulation we have used 200 images for training, 160 images for testing, that is, to

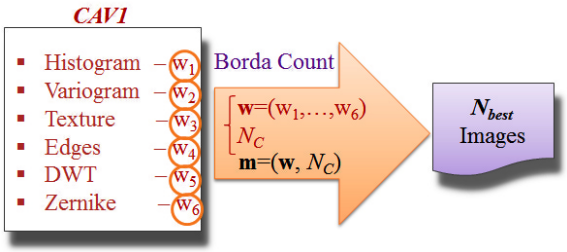


Fig. 1. Tuning parameters for *CAV1* (**m**).

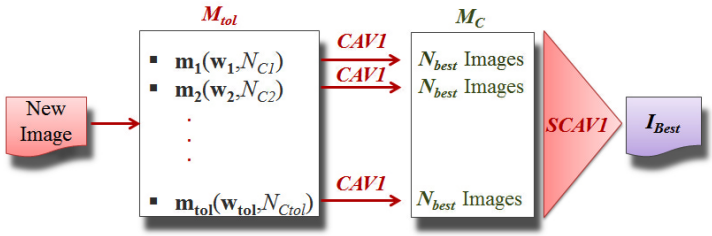


Fig. 2. Consensus algorithm *SCAV1* based on the uncertainty set M_{tol} of *CAV1*.

optimize the parameters of the ensemble learning classifiers, and finally 40 images to perform blind validation and establishing the accuracy rates of these algorithms. The PUT face database is also publicly available and provides 9971 RGB images of 100 different subjects, taken in partially controlled illumination conditions over an uniform background. Images of each person contain neutral expression with changes of poses, such as the head turning from left to right, the head nodding from the raised to the lowered position, the raised head turning from left to right, the lowered head turning from left to right, and also no constraints regarding the face pose or expression. For our numerical experiments we have randomly selected 50 subjects with 22 poses per person, as shown in figure 3. These face images were previously normalized. Therefore, the PUT database that we used, was composed by 1100 faces, and divided as follows: 400 faces to learn, 400 faces to test (for the optimization of *SCAV1* parameters), and 300 images for blind validation and establishing the predictive accuracy of the *SCAV1* ensemble learning methodology.

Figures 4 to 7 show the posterior histograms describing the uncertainty space of the *SCAV1* parameters for the ORL and PUT databases, both, with and without group substitution. In all the cases the search space for the weights was the interval $[50, 100]$ and $[0, 30]$ for the number of selected candidates. It can be observed that the shape of these histograms is highly variable and depends on the database. This fact confirm the need of supervision for a particular image database. In general terms,

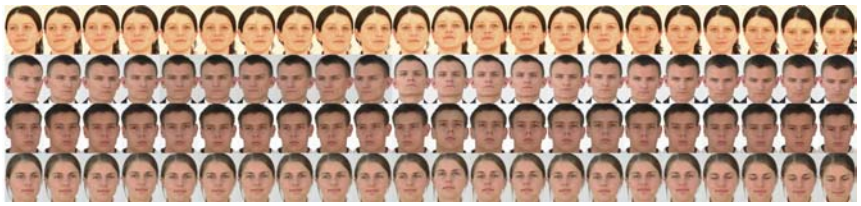


Fig. 3. PUT face database.

the highest modes for the weights correspond to the histogram and DWT, while the variogram, edges, texture and Zernike moments show tendencies towards lower mode weights. The number of candidates shows a mode around 20 images in both cases. These modes are different from the individual accuracies of these nearest-neighbors classifiers shown in table 3, that were used in the case of unsupervised ensemble learning with *CAV1*[8]. This fact serves to explain why the results achieved by the unsupervised *CAV1* were not optimum.

Table 4 summarizes the results obtained in the numerical experiments with both databases. For the ORL database the optimum accuracies were obtained without group substitution, although the results obtained with group substitution were also outstanding. The median accuracy is 100%, and the standard deviation is quite low, since we achieved perfect classification in most of the runs. In the case of the PUT database the results are slightly better without group substitution. In both cases, the stability is very high and seems to be lower with group substitution.

Table 4. Accuracies for *SCAV1* using the optimized parameters. GS stands for group substitution.

Algorithm	min	max	median	mean	IQR	std
SCAV1 / ORL	97.50	100	100	99.25	2.5	1.2
SCAV1+GS / ORL	97.50	100	100	99.00	2.5	1.3
SCAV1 / PUT	98.00	100	99.33	99.13	0.33	0.59
SCAV1+GS / PUT	98.67	100	99.50	99.43	1	0.52

7. Conclusions

In this paper we have presented the design of *SCAV1* for supervised ensemble classification in face recognition. This algorithm is based on a set of 6 diverse nearest-neighbor classifiers (local histogram, local variogram, local texture, global edges, global DWT and global Zernike moments), and takes advantage of the uncertainty space of this ensemble classifier. The optimization/sampling of the *SCAV1* parameters (the weights of the nearest neighbor classifiers and the number of image candidates) is done via a powerful particle swarm optimizer (RR-PSO). The final

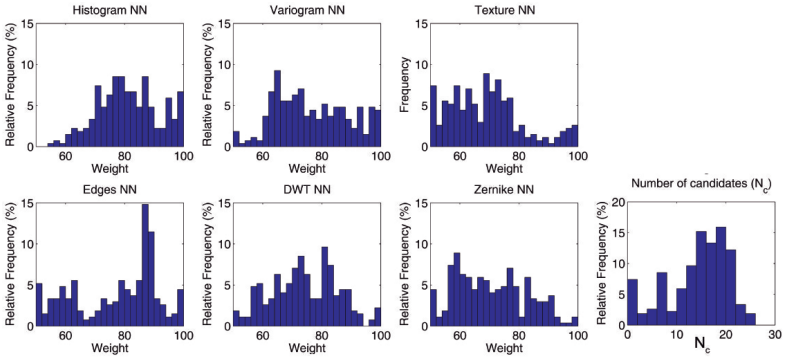


Fig. 4. ORL database: posterior histograms for *SCAV1* parameters without group substitution.

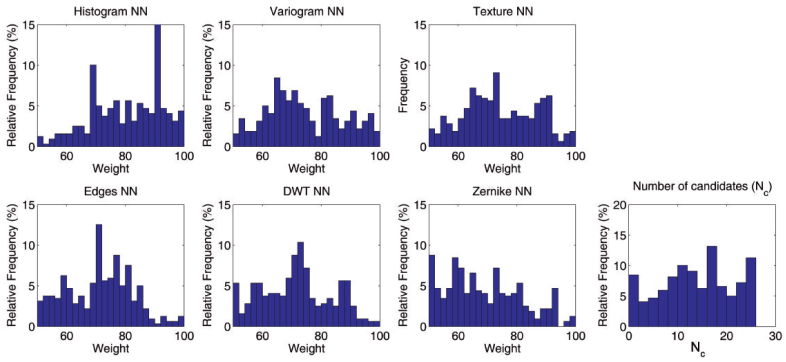


Fig. 5. PUT database: posteriors histograms for *SCAV1* parameters without group substitution.

classifier is made by majority voting on the set of images provided by exploiting the uncertainty space of *SCAV1*. The application of this methodology to two publicly available image databases (ORL and PUT), shows that we are able to obtain almost perfect classification, with very high stable accuracies. In conclusion, exploiting the uncertainty space of an ensemble classifier (*SCAV1* in this case) is the right strategy to adopt, and provides optimum results. This feature is related the Bayesian posterior analysis and the importance sampling of equivalent solutions of inverse and machine learning problems. Nevertheless, further research work should be done to test this methodology in other databases and interesting biometry problems, with application in the biomedical field.

References

1. S. Aksoy and R.M. Haralick, Content-based image database retrieval using variances of gray level spatial dependencies, In *In Proc. of IAPR Intl. Workshop on Multimedia*

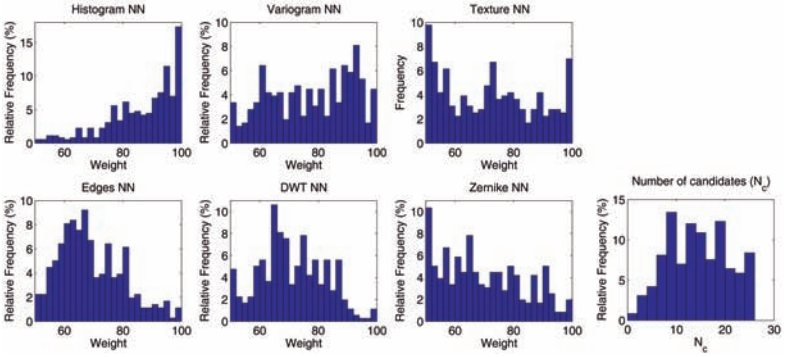


Fig. 6. ORL database: posterior histograms for *SCAV1* parameters with group substitution (GS).

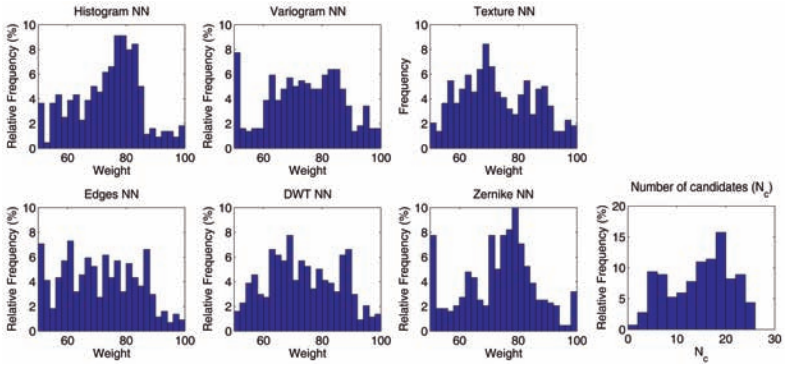


Fig. 7. PUT database: posteriors histograms for *SCAV1* parameters with group substitution (GS).

Information Analysis and Retrieval, pages 3–19, 1998.
2. M. Bartlett, J. Movellan, and T. Sejnowski, "Face recognition by independent component analysis", *IEEE Trans. on Neural Networks*, 13(6):1450 – 1464, nov 2002.
3. P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", In B. Buxton and R. Cipolla, editors, *Computer Vision ECCV '96*, volume 1064 of *Lecture Notes in Computer Science*, pages 43–58. Springer Berlin Heidelberg, 1996.
4. M. H. Bharati, J.J. Liu, and J.F. MacGregor, "Image texture analysis: methods and comparisons", *Chemometrics and Intelligent Laboratory Systems*, 72(1):57-71, 2004.
5. P. J. Boland, "Majority system and the Condorcet jury theorem", *Statistician*, vol. 38, no. 3, pp. 181189, 1989.
6. Leo Breiman, "Bagging Predictors", *Machine Learning*, 24(2):123–140, 1996.
7. John Canny. "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6): pp. 679-698, 1986.
8. A. Cernea and J. L. Fernández-Martínez, "Unsupervised Ensemble Classification for Biometric Applications", *International Journal of Pattern Recognition and Artificial*

Intelligence, In Revision, 2014.
9. N. V. Chawla, K. W. Bowyer, "Ensembles in face recognition: tackling the extremes of high dimensionality, temporality, and variance in data", *IEEE International Conference on Systems, Man and Cybernetics, 2005* , vol.3, pp: 2346-2351, 2005.
10. Choi, Kwontaeg and Toh, Kar-Ann and Byun, Hyeran "A Random Network Ensemble for Face Recognition", In *Lecture Notes in Computer Science*, 5558, pp. 92-101, 2009.
11. H. Ebrahimpour and A. Kouzani, Face Recognition Using Bagging KNN, In *International Conference on Signal Processing and Communication Systems (ICSPCS)* Australia, Gold Coast, pages 17–19, 2007.
12. Z. G. Fan and B. L. Lu, "Fast learning for statistical face detection", *Neural Information Processing*, pp. 187-196, Springer Berlin Heidelberg, 2006.
13. J. L. Fernández-Martínez and A. Cernea, "Numerical analysis and comparison of spectral decomposition methods in biometric applications", *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 28(1), 2014, pp.14560-14593.
14. J. L. Fernández Martínez, and E. García Gonzalo, "The generalized PSO: a new door for PSO evolution", *Journal of Artificial Evolution and Applications*, vol. 2008, Article ID 861275, 15 pages, 2008.
15. J. L. Fernández-Martínez and E. García-Gonzalo, "The PSO family: deduction, stochastic analysis and comparison", *Swarm Intelligence*, vol. 3, pp. 245-273, Dic. 2009.
16. J. L. Fernández-Martínez and E. García-Gonzalo, "Stochastic Stability Analysis of the Linear Continuous and Discrete PSO Models", *Evolutionary Computation, IEEE Transactions on* , vol.15, no.3, pp.405,423, June 2011.
17. J. L. Fernández-Martínez, M. Z. Fernndez Muñiz, and M. J. Tompkins, "On the topography of the cost functional in linear and nonlinear inverse problems", *Geophysics*, vol. 77, issue 1, p. W1, 2012.
18. J. L. Fernández-Martínez and E. García-Gonzalo, "Stochastic Stability and Numercal Analysis of Two Novel Algorithms of PSO Family: PP-PSO and RR-PSO", In *International Journal on Artificial Intelligence Tools*, volume 21, 03, 2012.
19. N. H. Foon, Y.H. Pang, A.T.B. Jin, D.N.C. Ling, "An efficient method for human face recognition using wavelet transform and Zernike moments", *Computer Graphics, Imaging and Visualization, 2004. CGIV 2004. Proceedings. International Conference on*, pp. 65 - 69, 2004.
20. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55(1):119-139, 1997.
21. E. Garcia-Gonzalo, J. L. Fernandez-Martinez, and Ana Cernea "Four-Points Particle Swarm Optimization Algorithms", *Journal of Multiple-Valued Logic and Soft Computing*, Volume 22, Number 3, 2014, p. 239-266.
22. G. D. Guo, H. J. Zhang, and S. Z. Li, "Pairwise face recognition", In *Proceedings of Eighth IEEE International Conference on Computer Vision, ICCV 2001*, Vol. 2, pp. 282-287, 2001.
23. G. D. Guo and H. J. Zhang, "Boosting for fast face recognition", In *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on* pp. 96-100, 2001.
24. Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform", *Int. J. Comput. Vision*, 43(3):167–188, July 2001.
25. Hamzeloo, S.; Shahparast, H.; Jahromi, M.Z., "A novel weighted nearest neighbor ensemble classifier," , *16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp.413-416, 2012.
26. L. K. Hansen and P. Salamon. 1990. "Neural Network Ensembles", *IEEE Trans.*

Pattern Anal. Mach. Intell., Vol. 12(10):993–1001, 1990.
27. M. Harandi, J. Taheri, and B. Lovell, "Ensemble Learning for Object Recognition and Tracking", *Pattern Recognition, Machine Intelligence and Biometrics*, 2(11), pp.261-278, 2011.
28. T. K. Ho, "The random subspace method for constructing decision forests", In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832844, 1998.
29. R. Jafri and H. R. Arabnia, "A survey of face recognition techniques", *Journal of Information Processing Systems*, 5(2):41–68, june 2009.
30. S. Kakarwal and R. Deshmukh, "Wavelet transform based feature extraction for face recognition", *International Journal of Computer Science and Application Issue*, –:–, 2010.
31. A. Kasinski, A. Florek, A. Schmidt "The PUT Face Database", *Image Processing and Communications*, Volume 13, Number 3-4, pp.59-64, 2008.
32. A. Khotanzad and Y. H. Hong, "Invariant Image Recognition by Zernike Moments", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol 12(5), pp. 489-497, 1990.
33. J. Kennedy and R. Eberhart, "Particle swarm optimization", *Proceedings IEEE International Conference on Neural Networks*, vol.4, pp. 1942-1948, 1995.
34. S. Lajevardi and Z. Hussain, "Zernike moments for facial expression recognition", *International Conference on Communication, Computer and Power*, pp: 378-381, 2009.
35. Th. H. Le, "Applying Artificial Neural Networks for Face Recognitionv, Advances in Artificial Neural Systems, vol. 2011, Article ID 673016, 16 pp, 2011.
36. Juwei Lu, K. N. Plataniotis, A. N. Venetsanopoulos and Stan Z. Li, "Ensemble-based discriminant learning with boosting for face recognition", *IEEE Transactions on Neural Networks*, vol.17, no.1, pp.166,178, January 2006.
37. R. Mallipeddi and Minho Lee, "Ensemble based face recognition using discriminant PCA Features", *2012 IEEE Congress on Evolutionary Computation (CEC)* , pp.1–7, 10-15 June 2012.
38. A. Mashhoori and M. Z. Jahromi, "Block-wise two-directional 2DPCA with ensemble learning for face recognition", *Neurocomputing*, Volume 108(2), pp. 111-117, 2013.
39. M. Najafi and M. Jamzad, "An Ensemble Based Learning For Face Recognition With Similar Classifiers", *Proceedings of the World Congress on Engineering and Computer Science, WCECS 2007*, October 24-26, 2007, San Francisco, USA.
40. B. Parmanto, P. W. Munro, and H. R. Doyle, "Improving committee diagnosis with resampling techinques", *Advances in neural information processing systems, MIT Press, Cambridge,*, vol. 8: pp 882888, 1996.
41. R. Polikar, "Ensemble based systems in decision making", *Circuits and Systems Magazine, IEEE*, 6 (3): 21–45, 2006.
42. R. K. Sadykhov, V. A. Samokhval, and L. P. Podenok, "Face recognition algorithm on the basis of truncated Walsh-Hadamard transform and synthetic discriminant functions", In *FGR - IEEE Computer Society*, pp. 219–222, 2004.
43. Lior Rokach. "Ensemble-based classifiers", *Artificial Intelligence Review*, 33, 1(2), pp. 1–39, 2010.
44. F. S. Samaria, A. Harter, "Parameterisation of a Stochastic Model for Human Face Identification", *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp.138,142, 1994.
45. R. E. Schapire, "The Strength of Weak Learnability", *Machine Learning*, 5(2): 197–227, 1990.
46. Y. Shi and R. Eberhart, "Parameter selection in particle swarm optimization", *Evolutionary Programming VII*, pp. 591–600, 1998.

47. K. Sirlantzis, S. Hoque, and M. C. Fairhurst, "Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition", *Appl. Soft Comput.*, 8(1): 437–445, January, 2008.
48. L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces", *Journal of the Optical Society of America A*, 4(3):519–524, 1987.
49. J. A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*, Lawrence Erlbaum Associates, Mahwah, NJ, 1996.
50. T. Windeatt, "Ensemble neural classifier design for face recognition", *Proceedings of European Symposium on Artificial Neural Networks*, pp. 373-378, 2007.
51. S. E Umbaugh, *Computer Vision and Image Processing: A Practical Approach Using CVIPtools*. ISBN 0-13-264599-8. Prentice Hall Professional Technical Reference, 1998.
52. J. Yang, D. Zhang, A. F. Frangi, and J. yu Yang, "Two-dimensional pca: A new approach to appearance-based face representation and recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:131–137, 2004.
53. J. Zhang, L. He and Z. H. Zhou, "Ensemble-Based Discriminant Manifold Learning for Face Recognition" *Advances in Natural Computation - Lecture Notes in Computer Science*, 4221, pp. 29-38, 2006.

Appendix B

Contributions to International Congresses

The following contributions were also made to International Congresses:

The 6^{th} ***International Conference on Advanced Cognitive Technologies and Applications - COGNITIVE 2014***, May 25 – May 30, 2014, Venice, Italy. In this conference we addressed the topic *COGNITION: Artificial intelligence and cognition* which was dedicated to: Expert systems, knowledge representation and reasoning; Reasoning techniques, constraint satisfaction and machine learning; Logic programming, fuzzy logic, neural networks, and uncertainty; State space search, ontologies and data mining; Games, planning and scheduling; Natural languages processing and advanced user interfaces; Cognitive, reactive and proactive systems; Ambient intelligence, perception and vision; Pattern recognition.

The 5^{th} ***Joint International Conference on Swarm, Evolutionary, and Memetic Computing - SEMCCO 2013***, Chennai, India, December 19-21, 2013. This conference aims at bringing together researchers from academia and industry to report and review the latest progresses in the cutting – edge research with Swarm, Evolutionary, Memetic Computing to explore new application areas and to design new bio – inspired algorithms for solving specific hard optimization problems and finally to create awareness on these domains. All accepted papers of SEMCCO 2013 were published in LNCS series of Springer. Indexed in ISI Proceedings, DBLP, ZBIMath/Compuserve, IO-Port, EI, ACM Portal, Scopus, INSPEC, etc.

B.1. Automatic Classification of Cell Patterns for Triple Negative Breast Cancer Identification

Authors:
Juan Luis Fernández Martínez, Ana Cernea, Enrique J. deAndrés-Galiana, Primitiva Menéndez-Rodríguez, José A. Galván and Carmen García-Pravia

Proceedings of COGNITIVE 2014
The Sixth International Conference on Advanced Cognitive Technologies and Applications
May 25 – May 30, 2014 (In Print), Venice, Italy.

Automatic Classification of Cells Patterns for Triple Negative Breast Cancer Identification

Juan Luis Fernández-Martínez*, Ana Cernea*, Enrique J. deAndrés-Galiana*, Primitiva Menéndez-Rodríguez†, José A. Galván‡ and Carmen García-Pravia†
*Mathematics Department. Oviedo University. C/ Calvo Sotelo s/n 33007 Oviedo–Spain
Email: jlfm@uniovi.es
†Servicio de Anatomía Patologica, Hospital Universitario Central de Asturias
‡Translational research unit. Institute of Pathology. University of Bern

***Abstract*—This paper is devoted to present a methodology for the optimization of the basal cells pattern classification. Different unsupervised and supervised learning techniques are applied to the analysis, diagnosis and prognosis of cell patterns classification for triple negative cancers (TNBC), a group of cancers that share with basal like breast cancer very bad prognosis. For that purpose, different machine learning algorithms are performed on histological images, and on a list of pathological and immunohistochemical variables currently-used in medical practice. The main objective is to design a biomedical robot able to assess physicians on the kind of histological grade of different subgroups of TNBC samples in order to optimize the treatment protocol. The proposed methodology is performed over a database of 116 patients. The results show that pathological and immunohistochemical variables and histological images provide complementary information to improve the classification of TNBC samples.**

I. Introduction

Breast cancer (or neoplasia) is a very heterogeneous disease. This term encompasses a variety of entities with distinct morphological features and clinical behaviors. For a long time, breast tumors have been classified according to their morphologic features (histologic type and grade) to ascertain prognostic outcome in patients. Subsequently, molecular markers (the expression of estrogen and progesterone receptor and human epidermal growth factor 2 receptor) were used to provide additional predictive power. Therefore, triple negative breast cancers (TNBC) refers to any breast cancer characterized by the absence of estrogen receptors (ER), progesterone receptors (PR) and human epidermal growth factor 2 receptors (HER2). This classification is important from a clinical and therapeutic point of view, since TNBC are resistant to targeted therapies, because they do not express these receptors [7], [19]. Statistics showed that TNBC accounts for approximately $15\% - 25\%$ of all breast cancer cases [2].

Recently a molecular classification based on gene expression profiles classified tumors into five groups that were not detected using traditional histopathologic methods. This classification includes the basal-like tumors group [22]. These tumors are defined by: (1) the lack of ER, PR, and HER2 expressions; (2) the expression of one or more high-molecular-weight/basal cytokeratins (CK5/6, CK14); (3) the lack of expression of ER and HER2 in conjunction with expression of CK5/6; and (4) the lack of expression of ER, PR, and HER2 in conjunction with expression of CK5/6. Also, from a morphological point of view both basal-like and triple negative breast cancers share a predominance of high histologic grades. The analysis of gene expression profiles showed that 77% of basal-like tumors were of TNBC phenotype [18], [19], [22].

The treatment for TNBC is adjuvant chemotherapy and radiotherapy. Unfortunately response to chemotherapy does not correlate with overall survival. In addition, most recurrences are observed in TNBC during the first and third years after therapy, and most deaths take place in the first five years. The survival decreases after the first metastatic event [17]. Therefore, in this heterogeneous group of tumors, new identification and classification techniques are necessary to establish a better diagnosis, prognosis to establish appropriate therapies [8].

The main objective is to design a biomedical robot able to help to physicians on the diagnosis of different subgroups of TNBC in order to optimize their treatment protocol. The first aim of this research is to analyze the possibility of performing an automatic histological grade prediction using different biometric attributes of TNBC images and also a list of currently-used pathological and immunohistochemical variables. The methodology used in this paper is inspired in [9], [10] previous research works. The preliminary conclusion of this study is that the use of both pieces of information (immunohistochemical markers and histological images) might improve the accuracy of TNBC histological grades classification and survival.

II. Database Description

A. Histological Images

A cohort of 105 Caucasians women diagnosed in the Hospital Universitario Central de Asturias(Spain) with TNBC and ages between 30 and 94 years were enrolled in this study, which was developed in accordance with the Helsinki Declaration of 1975. Tumor samples were obtained from surgical resection. They were fixed in 10% formaldehyde and paraffin embedded, then cut $4\mu m$ thick, mounted on treated slides, and stained with H&E stain (Hematoxylin and eosin stain). Finally, the sections were studied and photographed at two different resolutions (100X and 400X) using an Olympus light microscope. Most of the cancers in this cohort were classified in histological degrees 2 (20 samples) and 3 (89

samples), and only two samples were in degree 1. Also, a few samples have a histological degree which is unknown. This methodology will be used in the future to asses the histological grade of the TNBC samples, and analyzing the possibility of predicting the patients survival.

B. Pathological and Clinical Variables

The pathological and clinical variables description is important to understand the different classification problems involved in this analysis, and the way the histological grade of the TNBC is established in medical practice.

Grade The Nottingham Histologic Score system (the Elston-Ellis modification of Scarff-Bloom-Richardson grading system) has been applied to establish the histological grades of the TNBC cancers. This system is based on the ability of the tumor to form structures similar to the ducts where the tumor is originated, on the similarity between the cancer cells and the original benign cells and finally on its proliferative activity. The histological grade will be used as the class for the different machine learning classification problems presented in this paper. In the grade score, several factor are taken into account:

1. Tubular structure formation: the score increases with the percentage of tumor area forming glandular/tubular structures, as follows: score 1: $> 75\%$ of tumor area forming glandular/tubular structures; score 2: 10% to 75% of tumor area forming glandular/tubular structures; score 3: $< 10\%$ of tumor area forming glandular/tubular

2. Nuclear pleomorphism: the score increases with varaition of size and shape of cells, as follows: score 1: Small nuclei with little increase in size in comparison with normal breast epithelial cells, regular outlines, uniform nuclear chromatin, little variation in size; score 2: Cells larger than normal with open vesicular nuclei, visible nucleoli, and moderate variability in both size and shape; score 3: Cells with vesicular nuclei, often with prominent nucleoli, exhibiting marked variation in size and shape.

3. Mitotic rate: The mitotic count score depends on the field diameter of the microscope used by the pathologist. The pathologist will count how many mitotic figures are seen in 10 high power fields. Using a high power field diameter of 0.50 mm, the criteria are as follows: score 1: less than or equal to 7 mitoses per 10 high power fields; score 2: 8-14 mitoses per 10 high power fields; score 3: equal to or greater than 15 mitoses per 10 high power fields.

The final total score is used to determine the histological grade of a TNBC sample is as follows:

- Histological grade 1: tumors with a total score between 3 and 5;
- Histological grade 2 tumors with a total score between 6 and 7;
- Histological grade 3 tumors with a total score between 8 and 9.

In the present case most TNBC samples were classified with grades 2 (20 samples) and mainly 3 (89 samples). Higher scores are usually correlated with the worse prognostics.

Other currently used variables include:

TNM stage: The pathologic stage of breast cancer takes into consideration the tumor size (T) and the presence of any lymph nodes metastases (N) or distant organ metastases (M).

Differentiation: it is combined measure of the tubular formation and the pleomorphism. It is a descriptor provided by the medical experts based on visual inspection of the histological images. This histological variable is expected to be highly correlated in medical practice to the histological grade of the different TNBC samples.

Vascular and perineural invasion: binary variable indicating the presence or absence of tumor cells inside the vessels and nerves, respectively.

Necrosis: binary variable indicating the presence of death cells. This variable is correlated with the TNBC aggressiveness.

In situ component: binary variable indicating the absence of invasion of tumor cells into the surrounding tissue. Most of these variables are provided by the pathologist by visual inspection of the TNBC images.

C. Immunohistochemical variables

The following immunohistochemical variables are also currently monitored:

1) **Estrogen receptors** (ER), Progesterone receptors (PR) and Androgen receptors (AR) nuclear expression: Hormone receptor status is important because these variables serve to decide whether the cancer is likely to respond to hormonal therapy or other treatments.
2) **Human epidermal growth factor receptor 2** (HER-2): HER2 testing is performed to assess prognosis and to determine suitability for trastuzumab therapy.
3) **Ki67 expression**: The percentage of Ki-67 ($< 2\%$, $2-20\%$ and $> 20\%$) can be used to aid in assessing the proliferative activity of normal and neoplastic tissue.
4) **Bcl-2 expression**: Bcl-2 is specifically considered as an important anti-apoptotic protein and classified as an oncogene. Bcl-2 expression is associated with a better prognosis [6].
5) **E-cadherin expression**: Reduction or loss of E-cadherin expression is associated with invasive carcinoma and possibly metastasis in a variety of carcinomas [21].
6) **P53 expression:** is also a maker used in breast cancer, but its significance in predicting clinical outcome remains controversial.
7) **CK-5/6** and **CK-14 expression**: are helpful markers in the identification of breast cancer with a basal phenotype.
8) **COL11A1**: is a stromal marker of invasion [12].

These variables but COL11A1 have been described in the literature as useful for TNBC description.

III. Machine learning using pathological and immunohistochemical variables

The aim of this section is to analyze the most discriminatory pathological and immunohistochemical variables of the histological grade. Data preprocessing include in this case the imputation of the clinical variables that have not been measured for some patients in real practice and the normalization of these variables in the interval $[0, 1]$ according to their own empirical cumulative distribution. Figure 1 shows the normalized pathological and immunohistochemical variables. The samples are arranged by their histological grades (2 to 3) beginning by the top of the image. It can be observed the high variability of these variables within the different classes (histological grades).

To perform machine learning we have first used feature selection methods to finding the minimum-size list of most discriminatory variables. For that purpose we defined the Generalized Fisher's ratio of the attribute j, for a binary classification problem, as follows [11]:

$$FR_j = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2},$$

where is μ_{j1} a measure of the center of the distribution of the attribute j in class i, and σ_{ji} is a measure of its dispersion within the class i. The Fisher's ratio (GFR) can be also generalized for multiclass classification as follows:

$$GFR_j = \sum_{i=1}^{N_c} \sum_{k=i+1}^{N_c} \frac{(\mu_{ji} - \mu_{jk})^2}{\sigma_{ji}^2 + \sigma_{ji}^2},$$

where N_c is the number of classes, j is the attribute index and i, k the classes indices.

This feature selection method looks for attributes that are homogenous within each class (low intra-class dispersion) and show a high separation between the center of the corresponding distributions (inter-class distance). Most discriminatory attributes correspond to higher Fisher's ratios.

The algorithm to find the minimum-size list of features is based on Recursive Feature Elimination, that is:

1) Attributes are ranked by the decreasing value of their Fisher's ratio.
2) Beginning by the tail of the list we calculate the accuracy of the different set of attributes, that are formed by dropping one attribute at each time. The set with the optimum accuracy and minimum size is therefore selected.
3) Finally, the accuracy of this reduced set of the attributes in the class prediction is based on Leave-One-Out method, using the average distance on the reduced set of attributes. The class with the minimum distance is assigned to the sample test. The average accuracy is calculated by iterating over all the samples.

The classification problem is linearly separable when this simple algorithm provides high accuracies. In the case where these accuracies decrease, other nonlinear classification algorithms should be used instead. If despite all these modifications, there is no improvement in the accuracy, this would mean that the data set (data and class) is noisy.

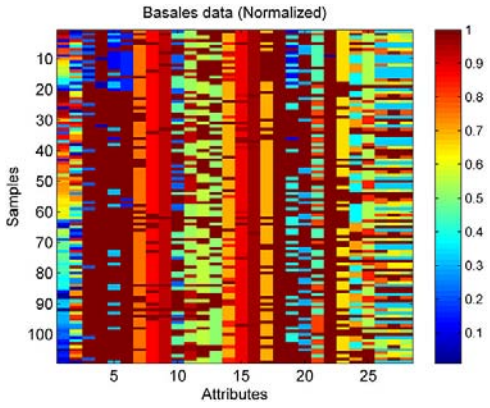


Figure 1: Normalized pathological and immunohistochemical array.

TABLE I: List of most discriminatory attributes and their corresponding Fisher's ratios using the median.

Attributes	FR
Mitotic count (10HPF)	4.56
Differentiation	4.55
AR expression	2.60
Tubular Formation	2.46
Insitu	2.06
T	2.05
N	1.99
Ki67 expression	1.78

TABLE II: List of most discriminatory attributes and their corresponding Fisher's ratios using the mean.

Attributes	FR
Differentiation	2.41
Mitotic count (10HPF)	2.00
Ki67 expression	0.94
AR expression	0.64
Tubular formation	0.56
pro-CollA1 intensity	0.24
Bcl2 expression	0.22
pro-Coll1A1 Score	0.21

Tables 1 and 2 show the list of attributes selected by the Fisher's ratio (FR) analysis using the mean and/or the median to describe the center of the distribution. These attributes are ranked by decreasing discriminatory power (FR). Five of the eight attributes in these lists are in common: Mitotic count, Differentiation, AR expression, Ki67 expression and Tubular Formation. In the first case (Table 1) the reduced base of features with the highest accuracy $(96, 4\%)$ is composed by the four first markers (Mitotic count, Differentiation, AR expression, and Tubular Formation). Only four patients are wrongly predicted using these attributes. In the second case (Table 2) the two first markers provide an accuracy of 94.4%.

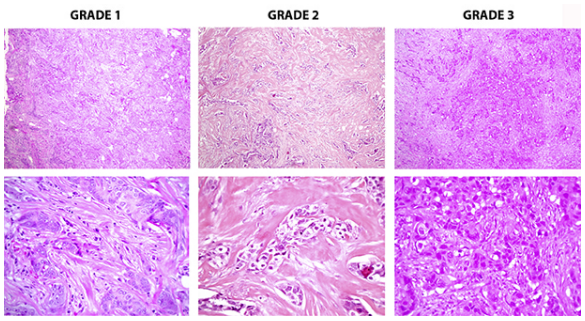


Figure 2: Basal images at two resolutions (100x and 400x magnification) for TNBC with three different histological grades

Also, other list of features with high accuracy (93.6%) includes HER2, PR expression, nipple and/or skin invasion, ER, AR and Ki67 expressions. Interesting the main attributes in these lists coincide with those used by medical experts to asses the histological grade of TNBC samples, nevertheless, these lists also show other attributes that are important for this automatic classification and were not directly used by the pathologists.

We have also analyzed the possibility of predicting the median survival of the different patients. This analysis has shown that the best markers to predict survival (with 78% of accuracy) are: E-cad expression, tumor size, perineural invasion, tubular formation, differentiation, and TNBC subtype. The accuracy of this prediction is lower than in the former case (histological grade), showing that this problem is not very well linearly separable using these attributes. Other additional variables, such as the kind of treatment followed by the patient, should be also used. The use of a nonlinear neural network classifier (extreme learning machine) [15] improved the accuracy of the prediction till 84%.

IV. Machine learning using histological images

The second aim of this research is to analyze the possibility of performing an automatic histological grade prediction using different biometric attributes of TNBC images corresponding to different histological grades taken at two different resolutions. These pattern images have been chosen by expert pathologists in this field.

Figure 2 shows different histological images at two different resolutions for cancers in degrees 1, 2 and 3. It can be observed that the main differences reside on the histological variables that have been described in the previous section, that are visually assessed by medical experts. The question resides in the possibility of capturing these characteristics using image processing techniques and machine learning.

A. The automatic image classification problem

The automatic image recognition problem consists in assigning a class to a new incoming image $I \notin B_d$, given a database of TNBC color training images:

$$B_d = \left\{ I_k \in S_{(n,m,3)}\left(\mathbf{N}\right) : k = 1, \ldots, N \right\},$$

that are characterized by a set of histological grades (labels) annotated by medical experts:

$$C_d = \left\{ c_k \in \{1, 2, ..., N_c\}, k = 1, \ldots, N \right\}.$$

In this definition $S_{(n,m,3)}$ is the space of color images of size $m \times n$, and N_c is the number of classes (3 in this particular case). To perform the classification it is necessary to construct a learning algorithm for the class prediction:

$$C^* : S_{(n,m,3)} \rightarrow C_d.$$

The classification is based on a nearest neighbor algorithm:

1) First, finding the image $I_k \in B_d$ such as:
$$d\left(I, I_k\right) = \min d\left(I, I_j\right), I_j \in B_d,$$
where d is a suitable distance (or norm) criterium defined over $S_{(n,m,3)}$.
2) Once this image has been found, assigning the class as follows: $C_I^* = C_{I_k} = C_k$.

The images are represented by a feature vectors calculated for each individual method of analysis (or attribute). Naming $\mathbf{v}_i^k \in \mathbf{R}^{s_k}$ the feature vector of image I_i according to the attribute k, the distance between two images I_i and I_j is defined as follows:

$$d\left(I_i, I_j\right) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \tag{1}$$

where p is a certain norm defined over the k-attribute space $(\mathbf{R}^{s_k})$.

The final classification will be performed by consensus:

1) From every individual non-supervised classifier built using the different attributes, we retain the first N_f images that are closer to I.
2) Based on this classification a matrix $M \in M_{Nf \times Na}$ is built, containing the N_f image candidates for each of the N_a attributes and their corresponding histological grades.
3) The score of image I according to the attribute j $(j = 1, N_a)$ to belong to the class k $(k = 1, N_c)$ is established as follows:
$$s_{jk} = \frac{1}{f_k} \frac{N_{jk}}{N_f},$$
where f_k is the sampling frequency of class k in the training database (examples) and N_{jk} is the number of images belonging to class k within the N_f candidates found for attribute j.
4) The final score for a new incoming image I to belong to class k is calculated as follows:
$$S_k = \sum_{j=1}^{N_a} s_{jk} w_j = \mathbf{s}_{jk} \cdot \mathbf{w}, \;\; k = 1, N_c$$
where $\mathbf{s}_{jk}$ is the score assigned by attribute j to class k and $\mathbf{w}$ is a vector of weights corresponding to the trust factors assigned to any individual classifier (attribute).
5) After calculating the scores for all the classes, the final classification of the test image I is performed by selecting the class with the major score.

Eventually, the number candidate images N_f, the sampling frequencies f_k and the trust factors $\mathbf{w}$ can be optimized (supervised learning) using global algorithms, such as PSO.

V. Image Attributes

In this paper we have used the following list of attributes, statistical based (histogram and variogram), spectral (discrete cosine transform), and image segmentation/regional descriptors (edges, texture and Zernike Moments). In this case all attributes will be calculated as global descriptors since TNBC image comparison should not be pixel-based.

A. PCA analysis using of attributes of the histological images

In this section we analyze the possibility of discriminating the different histological grades of the TBNC samples by means of unsupervised classification using the Principal Component Analysis (PCA).

PCA finds the orthogonal basis by diagonalizing the experimental covariance matrix of training images [16]:

$$S = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^T, \tag{2}$$

where $X_k \in \mathbf{R}^{Npixels}$ transformed into $1-D$ column vectors, $\mu = \frac{1}{N}\sum_{k=1}^{N} X_k$ is the images sample mean, N is the number of sample images contained in the learning database, and N_{pixels} is the number of pixels of each image. The eigenfaces u_k are the eigenvectors of S, corresponding to the largest eigenvalues. The dimensionality reduction from N_{pixels} to q parameters, is obtained by retaining the q first eigenfaces $\mathbf{u}_k$, spanning most of the database variability.

Figure 3 shows the PCA plot in two dimensions (two first PCA coordinates) of the different TNBC images at 10X resolution. We also show the TNBC samples that have positive androgen receptors (AR). This information is important since it implies a different type of TNBC (apocrine carcinoma). Apocrine carcinoma is a subtype of TNBC that expresses androgen receptor (AR), but often lacks estrogen receptor (ER) and progesterone receptor (PR). It is possible to observe that TNBC samples with $HG = 2$ are mainly located on three different clusters, surrounded by samples with $HG = 3$. Also, most of the $HG2$ samples correspond to apocrine type. Taking this fact into account, it seems that non-apocrine $HG2$ samples are only located in very restricted areas of the PCA diagram.

Figure 4 shows the 10 first mean PCA coefficients for TNBC images with histological grades 2 and 3. It can be observed that biggest differences occur high-order harmonics (4th). This attribute is expected to have a medium discriminatory power on TNBC images.

B. Color Histograms

An image histogram describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image [25].

For a gray-scale digital image I the histogram represents the discrete probability distribution of the gray-levels in the image. For this purpose the gray-scale space ($[0, 255]$ for an

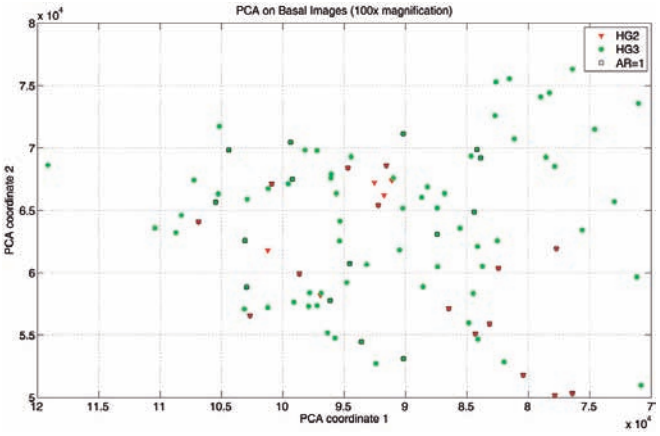


Figure 3: PCA plot of basal images at 100x magnification

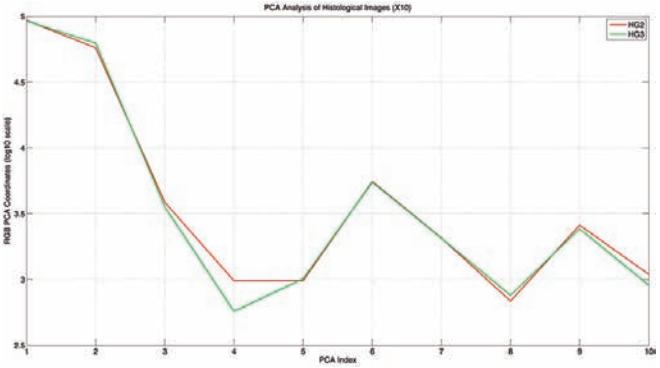


Figure 4: Mean PCA coefficients for HG2 and HG3 images.

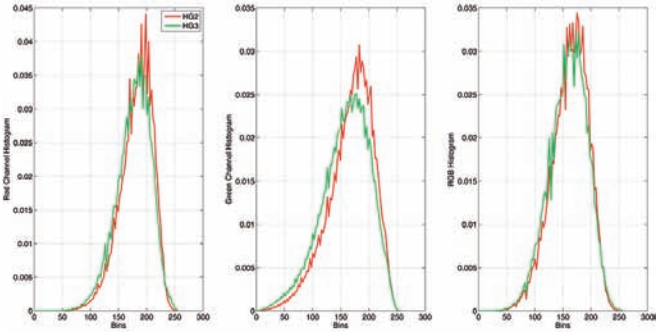


Figure 5: Mean histograms for HG2 and HG3 images.

`8-bit` image) is divided into L bins, and the number of pixels in each class n_i, $(i = 1, L)$ is calculated. In this case the attribute vector has dimension L:

$$H_I = (n_1, ..., n_L).$$

Relative frequencies can be also used by dividing the absolute frequencies n_i by the total number of pixels in the image.

In the case of RGB images the histogram is calculated for each color channel I_R, I_G and I_B, and then all the channels histograms are merged together, as follows:

$$H_I = (H(I_R), H(I_G), H(I_B)). \tag{3}$$

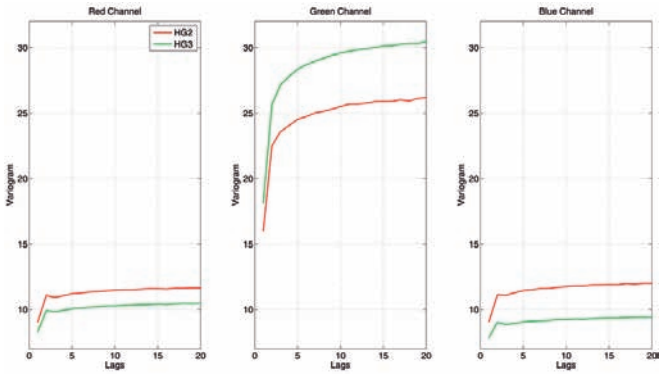


Figure 6: Mean variograms for HG2 and HG3 images.

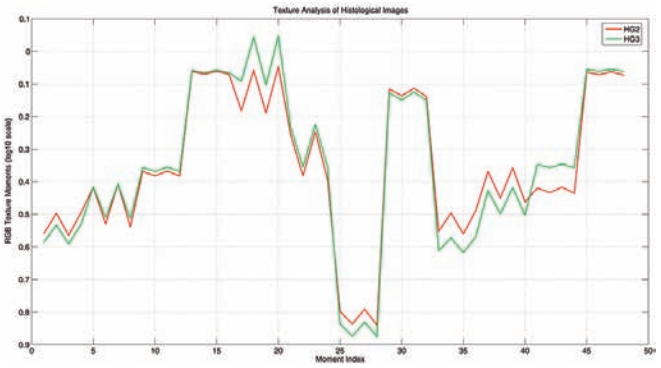


Figure 7: Mean texture coefficients for HG2 and HG3 images.

Color image histograms are widely applied in image processing and medical image classification.

Figure 5 shows the relative histograms of the color channels for TNBC images with HG2 and HG3. It can be observed that the major differences occur in the green channel, being its relative frequency lower than those of the red and blue channels. This attribute is expected to have a medium discriminatory power for TNBC images.

C. Variogram

The variogram of an image describes the spatial distribution in each color channel. In spatial statistics [13] the variogram describes the degree of spatial dependence of a spatial random field or stochastic process, the gray-scale in this case. For a given value of vector h, defined by a modulus and direction, the variogram is an index of dissimilarity between all pairs of values separated by vector h.

The omnidirectional p-variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^p. \qquad (4)$$

Usually $p = 2$. To compute the variogram each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$. Variograms are usually used to analyze spatial continuity and anisotropies. Therefore it can be considered as an additional texture attribute. The sill of the variogram is related to the channel variability, its range to the spatial continuity and its nugget (origin value) to the image microvariabilities.

Figure 6 shows the omnidirectional variograms of the three color channels for TNBC images with HG2 and HG3. It can be observed that the major differences occur in all the channels, being the blue and green the most discriminatory with respect to this attribute. The green channel also shows the biggest nugget. This attribute is expected to have a high discriminatory power for TNBC images.

D. Texture Analysis

Texture analysis of an image consists in analyzing regular repetitions of a pattern. In this paper, we use the spatial gray level co-occurrence matrix to describe an image texture. The gray level co-occurrence matrix (GLCM), or spatial dependence matrix of an image I is an estimate of the second-order joint probability function $P_{d,\theta}(i, j)$ of the intensity values of two pixels i and j located at a distance d apart (measured in number of pixels) along a given direction θ. Typically the GLCM is calculated for different pairs of d and θ. Different statistical moments can be calculated from the GLCM matrix, such as contrast, homogeneity, squared energy, correlation and entropy [3]. In the present case we have used a lag $d = 1$ for the directions. Figure 7 shows the texture moments of the three color channels for TNBC images with HG2 and HG3. The conclusions are similar than in the case of variogram. This attribute is expected to have a high discriminatory power for TNBC images.

E. Edges Detection

Edges are determined by sets of pixels where there is an abrupt change in intensity. If a pixel's gray level value is similar to those around it, there is probably not an edge at that point. However, if a pixel has neighbors with widely varying gray levels, it may represent an edge. Thus, an edge is defined by a discontinuity in the gray-level values [25]. More precisely, we can consider an edge as a property associated to a pixel where the image function $f(x, y)$ changes rapidly in the neighborhood of that pixel. In this case the function $f : \mathbf{R}^2 \to \mathbf{R}$ represents the pixels intensities. Related to f, an edge is a vector variable with two components: magnitude and direction. The edge magnitude is given by the gradient and its direction is perpendicular to the gradient's direction:

$$|\nabla f(x, y)| = \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2}, \qquad (5)$$

$$\theta(x, y) = arctg\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right) \pm \frac{\pi}{2}. \qquad (6)$$

In the case of digital images, these expressions are approximated by differences schemes. In this case we have used the Canny edge detection operator [5], which is one of the most commonly used in image processing, due to its property of

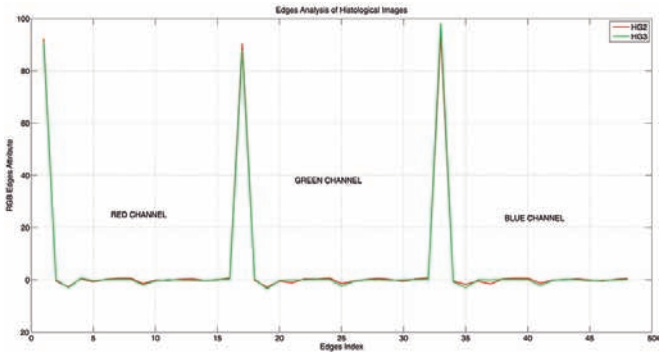


Figure 8: Mean edges vectors for HG2 and HG3 images.

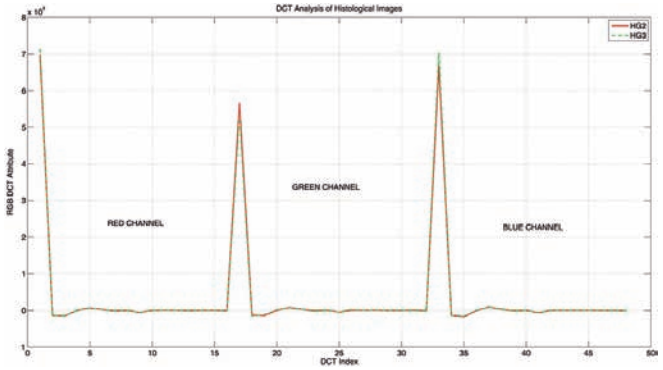


Figure 9: Mean DCT coefficients for HG2 and HG3 images.

detecting edges in a very robust manner in the case of noisy images.

The edge detection algorithm provides an image of the same size than the original image on which this analysis is performed. To produce the edge attributes we use a compression of edge image using the DCT. In the present case this analysis provides an attribute vector of dimension 48 for each image.

Figure 8 shows the DCT-edges moments of the three color channels for TNBC images with HG2 and HG3. The main differences occur for the first coefficients in each color channel. This attribute is expected to have a low/medium discriminatory power for TNBC images.

F. Discrete Cosine Transform (DCT)

DCT is a free-covariance model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image into cosines of increasing frequency [14].

DCT is a discrete Fourier transform that expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes. For an image I_k the DCT is defined as follows:

$$D(u,v) = c(u)c(v)\sum_{i=0}^{s-1}\sum_{j=0}^{n-1} D_{(i,j)} \tag{7}$$

where

$$D_{(i,j)} = I_k(i,j)\cdot \cos\frac{\pi(2i+1)u}{2s}\cos\frac{\pi(2j+1)v}{2n},$$

$u = 0, ..., s-1$, and $v = 0, ..., n-1$, being

$$c(\alpha) = \begin{cases} \frac{1}{\sqrt{N}}, & if \quad \alpha = 0, \\ \sqrt{\frac{2}{N}}, & if \quad \alpha \neq 0. \end{cases} \tag{8}$$

N is either the number of rows (s) or columns (n) of the image. The DCT can be expressed in matrix form as an orthogonal transformation

$$D_{CT} = U_{DC} I_k V_{DC}^T,$$

where matrices U_{DC} and V_{DC} are orthogonal. This transformation is separable and can be defined in higher dimensions. The feature vector of an image I_k is constituted by the $q_1 - q_2$ block of D_{CT}, $D_{CT}(1 : q_1, 1 : q_2)$, where q_1, q_2 are determined by energy reconstruction considerations using the Frobenius norm of the image I_k.

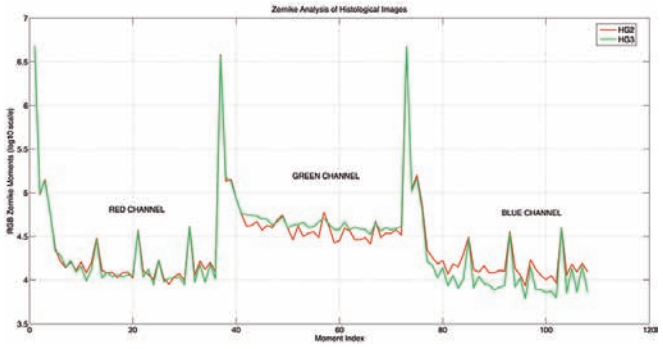


Figure 10: Mean Zernike moments for HG2 and HG3 images.

Figure 9 shows the DCT coefficients of the three color channels for TNBC images with HG2 and HG3. As in the previous case the main differences occur for the first coefficients in each color channel. This attribute is expected to have a low/medium discriminatory power for TNBC images.

G. Zernike Moments

Zernike polynomials are a sequence of polynomials that are orthogonal on the unit disk (referencia) and are widely used as basis functions for image analysis. Due to the orthogonality of Zernike polynomials, Zernike moments are image descriptors used in many applications due to their properties of orthogonality and rotation invariance. In biomedical applications, Zernike moments have been used as shape descriptors to classify benign and malignant breast masses [26].

Figure 10 shows the Zernike moments for the TNBC images of degree 2 and 3 for polynomials of order 10. Although Zernike moments has been previously applied as shape descriptors to classify benign and malignant breast tissues [26], the differences do not seem very important in this particular case and occur mainly for higher order polynomials in the green and blue channels. This attribute is expected to have a medium discriminatory power. Finally as a compendium of all this analysis, figure 11 shows the PCA plots in 2D (similar to figure 3) of all the attributes that have been commented. It

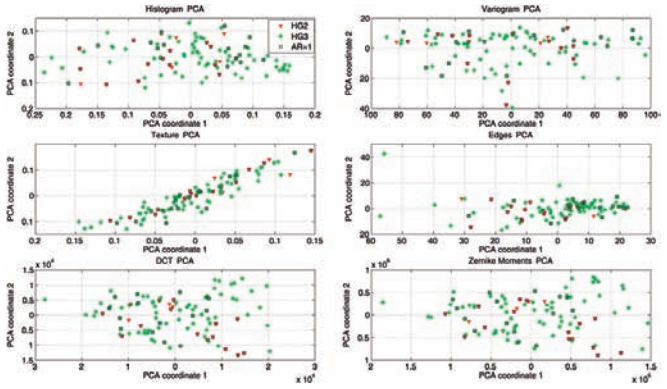


Figure 11: PCA plot of the used attributes for HG2 and HG3 images.

can be observed that the HG2 and HG3 samples are located differently in each of these diagrams.

The unsupervised machine learning algorithm commented in section IV-A provided an accuracy of 86.8% which is slightly higher than the majority voting algorithm (80%).

VI. Conclusions and future research

In this paper we have explored the possibility to design a biomedical robot able to assess physicians on the kind of histological grade/survival of different subgroups of TNBC samples in order to optimize their diagnosis/treatment and prognosis. Very promising premilimary results are shown using pathological and immunohistochemical variables and histological images of a cohort with 105 patients.

Future research will include the possibility of using other supervised learning techniques and global optimization methods (PSO) to optimize the machine learning parameters and improve the accuracy of the classification. Also it is expected that the use of both pieces of information (pathological and immunohistochemical variables and histological images) will provide complementary information, improving to improve the accuracy in the classification of TNBC samples (histological grade and survival).

Acknowledgment

The authors want to acknowledge *Anatomía Patológica* from the *Hospital Central de Asturias* for providing all the facilities to acquire and treat the TBNC samples.

References

[1] S. Aksoy and R. M. Haralick. Content-based image database retrieval using variances of gray level spatial dependencies. In *In Proc. of IAPR Intl. Workshop on Multimedia Information Analysis and Retrieval*, pages 3–19, 1998.

[2] K. R. Bauer, M. Brown, R. D. Cress et al. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. In *Cancer* 2007; New York. Published for the American Cancer Society by J. Wiley, 109: 17211728.

[3] M. H. Bharati, J.Jay Liu, and J. F. MacGregor. Image texture analysis: methods and comparisons. *Chemometrics and Intelligent Laboratory Systems*, 72(1):57 – 71, 2004.

[4] P. J. Boland, Majority system and the Condorcet jury theorem, *Statistician*, vol. 38, no. 3, pp. 181189, 1989.

[5] J. Canny, A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679 – 698, 1986.

[6] S. J. Dawson, N. Makretsov, F. M. Blows, et al. BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received. In *British Journal of Cancer.* Volume 103(5), pages: 668-75, August 2010.

[7] S. Edge, D. Byrd, C. Compton, et al., eds. AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 2010, pp 347-76.

[8] O. Fadare, F. A. Tavassoli, Clinical and pathologic aspects of basal-like breast cancers. In *Nat Clin Pract Oncol.* 2008 Mar;5(3):149-59. doi: 10.1038/ncponc1038.

[9] J. L. Fernández-Martínez., A. Cernea. Numerical analysis and comparison of spectral decomposition methods in biometric applications. *International Journal of Pattern Recognition and Artificial Intelligence*, Accepted for publication in 2013.

[10] J. L. Fernández-Martínez, A. Cernea, E. García-Gonzalo, et al. Aligned PSO for Optimization of Image Processing Methods Applied to the Face Recognition Problem, In *Swarm, Evolutionary, and Memetic Computing (SEMCCO), Springer Berlin Heidelberg*, Lecture Notes in Computer Science, volume 8297, 2013.

[11] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen*, 7:179–188, 1936.

[12] C. García-Pravia, J. A. Galván, N. Gutiérrez-Corral, L. Solar-García, E. García-Pérez, M. García-Ocaña, J. Del Amo-Iribarren, P. Menéndez-Rodríguez, J. García-García, J. R. de Los Toyos, L. Simón-Buela, L. Barneo. Overexpression of COL11A1 by cancer-associated fibroblasts: clinical relevance of a stromal marker in pancreatic cancer. In *PLoS One.* 2013 Oct 23;8(10):e78327. doi: 10.1371/journal.pone.0078327. eCollection 2013.

[13] P. Goovaerts. *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press, Incorporated, 1997.

[14] Z. M. Hafed and M. D. Levine. Face recognition using the discrete cosine transform. *Int. J. Comput. Vision*, 43(3):167–188, July 2001.

[15] Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew Extreme learning machine: Theory and applications. *Neurocomputing* 70 (2006) 489501.

[16] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 12(1):103–108, January 1990.

[17] C. Liedtke, C. Mazouni, K. R. Hess, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. Journal of Clinical Oncology 26 (8): 1275-81, 2008.

[18] C. M. Perou, T. Sørlie, M. B. Eisen, et al. Molecular portraits of human breast tumours. In *Nature* 406 (6797): 747-52, 2000.

[19] E. A. Rakha, J. S. Reis-Filho, F. Baehner, et al., Breast cancer prognostic classification in the molecular era: the role of histological grade. In *Breast Cancer Research* , Volume 12, Issue 14, 12:207, 2010.

[20] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review, 33, 1-2, 1-39.* 2010.

[21] S. M. Siitonen, J. T. Kononen, et al. Reduced E-cadherin expression is associated with invasiveness and unfavorable prognosis in breast cancer. *American Journal of Clinical Pathology,* 1996;105:394–402.

[22] T. Sørlie, C. Perou, R. Tibshirani, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. In *Proc Natl Acad Sci USA* 98 (19): 10869–10874, 2001.

[23] G. N. Srinivasan and Dr. G. Shobha. Statistical texture analysis. In *Proceedings Of World Academy Of Science, Engineering And Technology*, volume 36, pages 1264–12699, 2008.

[24] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, January 1991.

[25] S. E. Umbaugh. *Computer Vision and Image Processing: A Practical Approach Using CVIPtools*. Number ISBN 0-13-264599-8. Prentice Hall Professional Technical Reference, 1998.

[26] A. Tahmasbi, F. Saki, H. Aghapanah, S.B. Shokouhi, A Novel Breast Mass Diagnosis System based on Zernike Moments as Shape and Density Descriptors, *Proceeding of 18th Iranian Conference of Biomedical Engineering (ICBME)* , vol., no., pp.100,104, 14-16 Dec. 2011.

B.2. Aligned PSO for Optimization of Image Processing Methods Applied to the Face Recognition Problem

Authors:
Juan Luis Fernández Martínez, Ana Cernea, Esperanza García-Gonzalo, Velasco, Julian and BijayaKetan Panigrahi

Swarm, Evolutionary, and Memetic Computing - Lecture Notes in Computer Science series
Vol. 8297, pp. 642-651
2013

The paper presents a supervised ensemble learning classification methodology for the face recognition problem, based on a novel PSO version, the Aligned PSO (aPSO), applied to the face recognition problem.

The aPSO algorithm is a deduced from the PSO continuous model by adopting a centered discretization in acceleration, a backward discretization in velocity and an average discretization in trajectories using the same discrete position terms that appear in the velocity. This approach is currently used in diffusion-convection problems. First and second order stochastic stability analysis is also performed following the methodology that was used in the past for other PSO family members.

The first part of the article is dedicated to the presentation of the stochastic stability and convergence analysis of Aligned PSO. The numerical experiments using benchmark functions showed that the Aligned PSO version provides similar results to PSO, although it seems to be more exploratory.

In the second part this optimization algorithm is applied to the face recognition problem using the ORL database. Aligned PSO is used to optimize the parameters of a supervised learning technique that is designed to solve the face recognition problem. Images are represented, as usual, through their attributes vectors, which determine the individual classifiers. In this paper we have used the following list of attributes, statistical based (histogram and variogram), spectral (discrete cosine transform and discrete wavelet transform), and image segmentation/regional descriptors (texture).

The supervised classification that was performed here is based on the learning method L^* which depends also on a set of parameters $\mathbf{m} = (w_1, ..., w_{N_a}, N_{first})$ that are optimized via Aligned PSO. Since this article is previous to our third JCR article presented in the previous section, the exploitation of the whole uncertainty space of equivalent models, generated by the Aligned PSO was not contemplated. Therefore, the optimum model of parameters was chosen for the ensemble classification, obtaining very promising results.

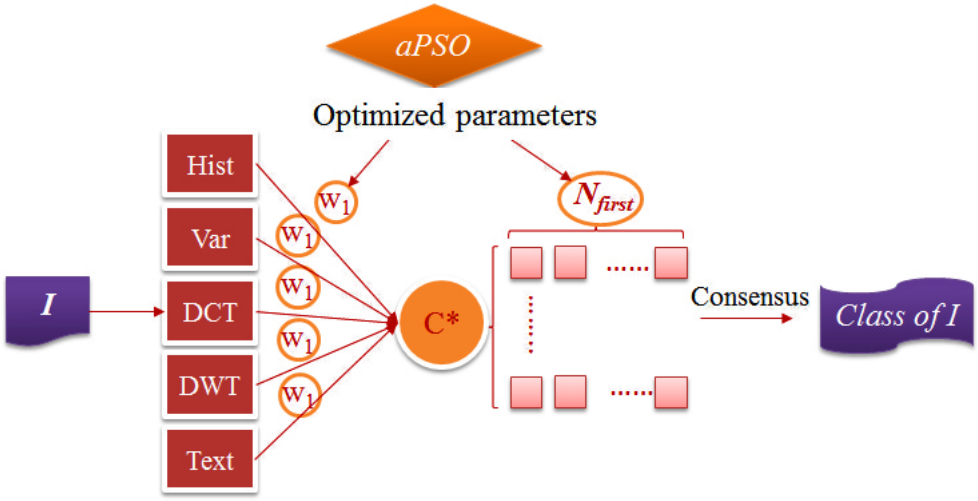


Figure B.1: Work-flow of the supervised classification with optimized parameters via aPSO.

The algorithm of supervised classification of a test images I used in this paper is illustrated in the workflow represented in figure B.1.

Its application to the ORL database provides 100% median identification accuracy over 100 independent runs. Nevertheless, other exploratory PSO versions could be used instead to perform this task. Further research will be devoted to this subject in the near future.

Aligned PSO for Optimization of Image Processing Methods Applied to the Face Recognition Problem

Juan Luis Fernández-Martínez[1], Ana Cernea[1], Esperanza García-Gonzalo[1], Julian Velasco[1], and Bijaya Ketan Panigrahi[2]

[1] Mathematics Department, Oviedo University, Spain
[2] Department of Electrical Engineering. IIT Delhi, India

Abstract. This paper is devoted to present the stochastic stability analysis of a novel PSO version, the aligned PSO, and its application to the face recognition problem using supervised learning techniques. Its application to the ORL database provides 100% median identification accuracy over 100 independent runs.

1 Introduction

Particle Swarm Optimization (PSO) is a bio-inspired algorithm that tries to mimic the the behavior of birds flocks or fish schooling. It was first proposed by Kennedy and Eberhart [9]. A swarm of n individuals or particles search the solution space and moves according to the equations

$$\begin{array}{l} \mathbf{v}_i^{k+1} = \omega \mathbf{v}_i^k + \phi_1^k(\mathbf{g}^k - \mathbf{x}_i^k) + \phi_2^k(\mathbf{l}_i^k - \mathbf{x}_i^k), \\ \mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \mathbf{v}_i^{k+1}, \end{array} \tag{1}$$

where $\omega \in \mathbb{R}$ is called the inertia weight, and ϕ_1 and ϕ_2 are random variables uniformly distributed in the intervals $[0, a_g]$ and $[0, a_l]$ with $a_g, a_l \in \mathbb{R}$. The first term in the velocity update is called the inertia damping; the second term is the social contribution and it is a stochastic discrete gradient between the positions and the global best $\mathbf{g}^k$, that is, the model with the lowest misfit found; finally the third term is called the cognitive contribution and it is also a stochastic discrete gradient with the best position found for each particle, $\mathbf{l}_i^k$.

This equation can be considered as a particular discretization (centered in acceleration and backward in velocity) of the continuous spring-mass system [3]:

$$\begin{cases} \mathbf{x}_i''(t) + (1-\omega)\,\mathbf{x}_i'(t) + \phi \mathbf{x}_i(t) = \phi_1 \mathbf{g}\,(t) + \phi_2 \mathbf{l}_i\,(t)\,, \\ \mathbf{x}_i\,(0) = \mathbf{x}_{i0}, \\ \mathbf{x}_i'\,(0) = \mathbf{v}_{i0}. \end{cases} \tag{2}$$

2 The Novel Aligned PSO Version

In this paper we propose to adopt the same PSO discretizations for the velocities and accelerations:

$$\begin{array}{l} \mathbf{x}_i''\,(k) \simeq \mathbf{x}_i^{k+1} - 2\mathbf{x}_i^k + \mathbf{x}_i^{k-1}, \\ \mathbf{x}_i'\,(k) \simeq \mathbf{x}_i^k - \mathbf{x}_i^{k-1}, \end{array} \tag{3}$$

B.K. Panigrahi et al. (Eds.): SEMCCO 2013, Part I, LNCS 8297, pp. 642–651, 2013.

and to approximate the position by an average of $\mathbf{x}_i^k$ and $\mathbf{x}_i^{k-1}$ as follows:

$$\frac{\mathbf{x}_i^k + \mathbf{x}_i^{k-1}}{2}.$$

This approximation is commonly used in convection-diffusion problems.

Then, the new PSO algorithm, named aligned PSO, becomes:

$$\begin{aligned} \mathbf{v}_i^{k+1} &= \omega \mathbf{v}_i^k + \tfrac{\phi_1^k}{2}(\mathbf{g}^k - \mathbf{x}_i^k) + \tfrac{\phi_2^k}{2}(\mathbf{l}_i^k - \mathbf{x}_i^k) + \\ &\quad + \tfrac{\phi_1^k}{2}(\mathbf{g}^{k-1} - \mathbf{x}_i^{k-1}) + \tfrac{\phi_2^k}{2}(\mathbf{l}_i^{k-1} - \mathbf{x}_i^{k-1}), \\ \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k + \mathbf{v}_i^{k+1}. \end{aligned} \tag{4}$$

It can be observed that a new term, involving the attractors $\mathbf{g}^{k-1}$ and $\mathbf{l}_i^{k-1}$, appears in the velocity update. This fact affects the center of attraction for each particle, that now becomes:

$$\mathbf{o}_i^k = \frac{\phi_1^k \left(\mathbf{g}^k + \mathbf{g}^{k-1}\right) + \phi_2^k \left(\mathbf{l}_i^k + \mathbf{l}_i^{k-1}\right)}{2\left(\phi_1^k + \phi_2^k\right)},$$

instead of:

$$\mathbf{o}_i^k = \frac{\phi_1^k \mathbf{g}^k + \phi_2^k \mathbf{l}_i^k}{\left(\phi_1^k + \phi_2^k\right)},$$

in the case of the standard PSO version (1)

2.1 The First Order Stability Region

In this section we briefly analyze the stochastic stability of the aligned PSO version, following the methodology shown in [3,4,5]. The first order dynamical system associated with the algorithm (4) is:

$$\begin{pmatrix} E\left(\mathbf{x}_i^{k+1}\right) \\ E\left(\mathbf{x}_i^k\right) \end{pmatrix} = A_\mu \begin{pmatrix} E\left(\mathbf{x}_i^k\right) \\ E\left(\mathbf{x}_i^{k-1}\right) \end{pmatrix} + c_\mu, \tag{5}$$

where

$$A_\mu = \begin{pmatrix} E\left(A\right) & E\left(B\right) \\ 1 & 0 \end{pmatrix},$$

and

$$\begin{aligned} A &= \frac{2\omega - \phi^k + 2}{2}, \\ B &= -\frac{\phi^k + 2\omega}{2}. \end{aligned} \tag{6}$$

The iteration matrix A_μ depends on ω and $\overline{\phi} = E\left(\phi\right)$. System (5) is stable when the spectral radius of A_μ is less than one. The first order stability region turns out to be the triangular region given by:

$$S_1 = \left\{\left(\omega, \overline{\phi}\right) : -1 < \omega < 1, 0 < \overline{\phi} < 2\left(1 - \omega\right)\right\}. \tag{7}$$

Also the parabola

$$\overline{\phi} = 6 - 4\sqrt{2 + 2\omega} + 2\omega, \tag{8}$$

marks the limit between the real and complex eigenvalues of A_μ. Figure 1a shows the isolines of the first order spectral radius.

2.2 The Second Order Stability Region

The second order stability dynamical system associated with (4) is:

$$\begin{pmatrix} E\left(\left(\mathbf{x}_i^{k+1}\right)^2\right) \\ E\left(\mathbf{x}_i^{k+1}\mathbf{x}_i^{k}\right) \\ E\left(\left(\mathbf{x}_i^{k}\right)^2\right) \end{pmatrix} = A_\sigma \begin{pmatrix} E\left(\left(\mathbf{x}_i^{k}\right)^2\right) \\ E\left(\mathbf{x}_i^{k}\mathbf{x}_i^{k-1}\right) \\ E\left(\left(\mathbf{x}_i^{k-1}\right)^2\right) \end{pmatrix} + c_\sigma, \tag{9}$$

with

$$A_\sigma = \begin{pmatrix} E\left(A^2\right) & 2E\left(AB\right) & E\left(B^2\right) \\ E\left(A\right) & E\left(B\right) & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

where matrix A_σ depends on ω and $\overline{\phi} = E\left(\phi\right)$ and $E\left(\phi^2\right)$. Taking into account that the distributions of ϕ_1 and ϕ_2 are uniform and naming $\beta = a_g/a_l$, we have

$$\overline{\phi} = \overline{\phi}_1 + \overline{\phi}_2 = \frac{a_g}{2} + \frac{a_l}{2} = \left(1+\beta\right)\frac{a_g}{2}, \tag{10}$$

$$\mathrm{Var}\left(\phi\right) = \frac{a_g^2}{12} + \frac{a_l^2}{12} = \frac{a_g^2}{12}\left(1+\beta^2\right) = \frac{1+\beta^2}{3\left(1+\beta\right)^2}\overline{\phi}^2,$$

$$E\left(\phi^2\right) = \mathrm{Var}\left(\phi\right) + \overline{\phi}^2 = \frac{2\left(2+3\beta+2\beta^2\right)}{3\left(1+\beta\right)^2}\overline{\phi}^2.$$

Now the iteration matrix A_σ depends on ω, $\overline{\phi}$ and β. The second order system (9) is stable in region

$$S_2 = \left\{\left(\omega, \overline{\phi}, \beta\right) : -1 < \omega < 1, 0 < \overline{\phi} < m_\beta\left(1-\omega\right)\right\}, \tag{11}$$

where

$$m_\beta = \frac{3\left(1+\beta\right)^2}{2+3\beta+2\beta^2},$$

is the slope of the second order upper stability limit. This region is a triangle embedded in S_1. Figure 1b also shows the second order spectral radius isolines. The upper stability limit intersects the ω axis in $\omega = 1$. Its slope is maximum when $\beta = 1$, being its value $m_\beta = 12/7$. This slope is minimum when $\beta \to 0$ and when $\beta \to \infty$ being its value $m_\beta = 3/2$. The geometry of the first and second order stability regions for the aligned PSO is much simpler than the ones corresponding to the standard PSO version.

2.3 Numerical Experiments

We have performed several numerical experiments using different benchmark functions currently used in practice. We had calculated the median logarithmic error over 50 simulations for a grid of $\left(\omega, \overline{\phi}\right)$ points covering the first order stability region. We have used 20, 40 and 100 particles for dimensions 10, 30 and

50 respectively. We have used as objective functions the Rosenbrock, Griewank, Rastrigin and Sphere benchmark functions. Figure 2 shows the results that have been obtained for 30 dimensions with the Rosenbrock and Griewank functions with the PSO and aPSO algorithms. It can be observed that the low misfit $\left(\omega, \overline{\phi}\right)$ points are located in a boomerang type zone that includes the upper border of the aPSO second stability region. In the aPSO case the boomerang part tend to vanish when the dimension of the search space increases. Also, PSO and aPSO provide similar minimum misfits.

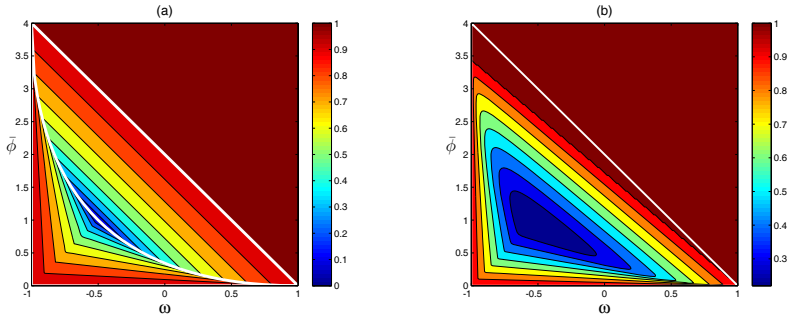


Fig. 1. Spectral radius isolines for (a) first and (b) second order stability regions for $\beta = 1$

3 Application to the Face Recognition Problem

3.1 Image Classification

The automatic image recognition problem consists of classifying a given probe image I providing a database of training images. Mathematically the problem can be formulated as follows: given a database of training images

$$B_d = \left\{ I_k \in S_{(n,m)}\left(\mathbb{N}\right) : k = 1, \ldots, N \right\}, \tag{12}$$

characterized by a set of labels

$$C_d = \left\{ C_k \in \{1, 2, ..., q\}, k = 1, \ldots, N \right\}, \tag{13}$$

and a new incoming image $I \notin B_d$, the problem consists in estimating its class C_I^*. In this definition $S_{(n,m)}$ is the space gray-scale images of size $m \times n$. In this problem the learning database typically contains N_p poses of each of the q individual classes, that is $N = q \cdot N_p$. To perform the classification it is necessary to construct a learning algorithm for the class prediction:

$$C^* : S_{(n,m)} \to C_d. \tag{14}$$

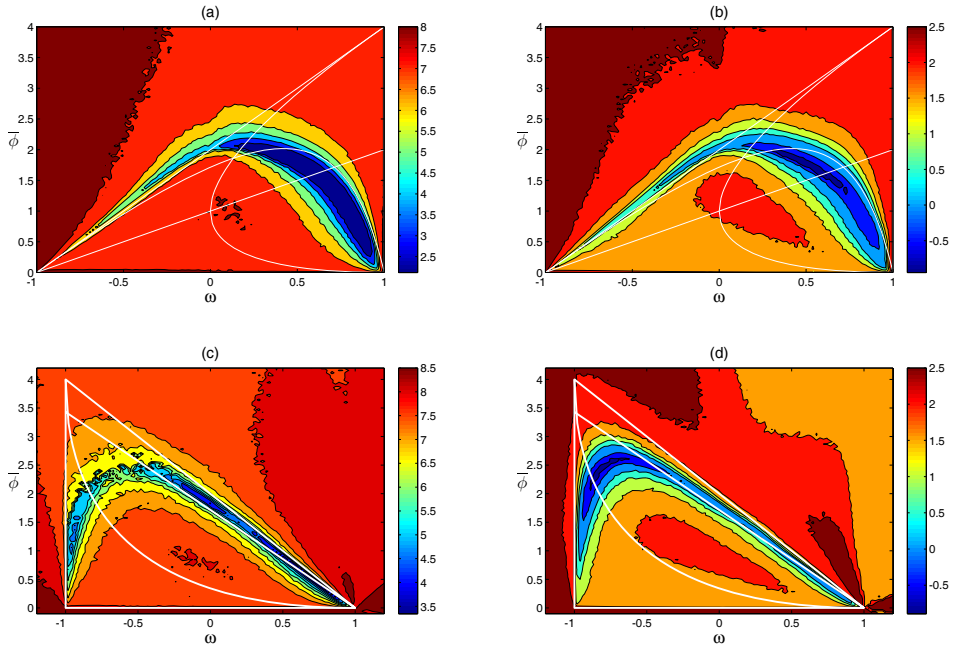


Fig. 2. Logarithmic median error in 30 dimensions for (a) PSO with Rosenbrock function, (b) PSO with Griewank function, (c) aPSO with Rosenbrock function and (d) aPSO with Griewank function

In the case of non-supervised learning the methodology is as follows:

1. First finding the image $I_k \in B_d$ such as:

$$d\left(I, I_k\right) = \min_{j \in B_d} d\left(I, I_j\right), \tag{15}$$

   where d is a suitable distance (or norm) criterium defined over $S_{(n,m)}$.
2. Once this image has been found: $C_I^* = C_{I_k} = C_k$.

The images are represented by a feature vector calculated for each individual method of analysis (or attribute). Naming $\mathbf{v}_i^k \in \mathbb{R}^{s_k}$ the feature vector of image I_i according to the attribute k, the distance between two images I_i and I_j is defined as follows:

$$d\left(I_i - I_j\right) = \|\mathbf{v}_i^k - \mathbf{v}_j^k\|_p, \tag{16}$$

where p is a certain norm defined over the k-attribute space ($\mathbb{R}^{s_k}$). The success of the non-supervised classification implicitly depends on the relationship between the k-attribute, the adopted criteria (norm or cosine) and the class information. Not all the attributes used in this paper perform equally. Attributes are described in section 3.2.

In the case of supervised classification the learning method C^* depends also on a set of parameters $\mathbf{m}$ that has to be tuned using class information coming from a subset of images of the database (testing database). The algorithm of supervised classification of a test images I used in this paper is as follows:

1. From every individual non-supervised classifier built using the different attributes, we retain the first N_{first} images that are closer to I.
2. Based on this classification a matrix $M \in M_{Nfirst \times Na}$ is built, containing the N_{first} image candidates for each of the N_a attributes ($N_a = 5$).
3. The array $\mathbf{c} \in \mathbb{R}^{Nd}$ containing the N_d different candidates is formed. For any image candidate $I_c \in \mathbf{c}$, its score according to the set of positions (i, j) in matrix M is calculated as follows:

$$s(I(i,j)) = (N_{first} - j + 1) \cdot \mathbf{w}(i),$$

   where $\mathbf{w}$ is a vector of weights corresponding to the trust factors assigned to any individual classifier (attribute).
4. After calculating the scores for all images in matrix M, the final classification of the test image I is performed by selecting the image with the major score among all the candidates of M matrix.

The aligned PSO version is used to optimize the parameter N_{first} and the set of weights $\mathbf{w}$. For that purpose the learning database will be divided in two different parts, one for learning and the other for optimizing. The results of the supervised classifiers will be obtained with a validation database of images that were not used in the supervised learning process.

3.2 Image Attributes

In this paper we have used the following list of attributes, statistical based (histogram and variogram), spectral (discrete cosine transform and discrete wavelet transform), and image segmentation/regional descriptors (texture). All these attributes can be calculated for gray scale and color images, both, locally or globally. In the case of global analysis the attribute features are calculated over the whole size of the image, meanwhile in the case of local features, the image is divided into blocks. For each block the local attributes are computed and the final feature vector is formed by merging all the local attributes into an unique vector, always computed in the same order. In this paper we have used a partition of the images into 8×4 blocks, nevertheless finer subdivisions could be also adopted.

Color Histograms. An image histogram describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image [11].

For a gray-scale digital image I the histogram represents the discrete probability distribution of the gray-levels in the image. For this purpose the gray-scale space ($[0, 255]$ for an `8-bit` image) is divided into L bins, and the number of pixels in each class n_i, $(i = 1, L)$ is calculated. In this case the attribute vector has dimension L:

$$H_I = (n_1, ..., n_L).$$

Relative frequencies can be also used by dividing the absolute frequencies n_i by the total number of pixels in the image.

In the case of RGB images the histogram is calculated for each color channel I_R, I_G and I_B, and then all the channels histograms are merged together, as follows:

$$H_I = \left(H\left(I_R\right), H\left(I_G\right), H\left(I_B\right)\right).$$

Variogram. The variogram of an image describes the spatial distribution in each color channel. In spatial statistics [6] the variogram describes the degree of spatial dependence of a spatial random field or stochastic process, the gray-scale in this case. For a given value of vector h, defined by a modulus and direction, the variogram is an index of dissimilarity between all pairs of values separated by vector h.

The omnidirectional p-variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^p. \tag{17}$$

Usually $p = 2$. To compute the variogram each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$.

Texture Analysis. Texture analysis of an image consists in analyzing regular repetitions of a pattern [10]. In this paper, we use the spatial gray level co-occurrence matrix to describe an image texture.

The gray level co-occurrence matrix (GLCM), or spatial dependence matrix of an image I is an estimate of the second-order joint probability function $P_{d,\theta}(i, j)$ of the intensity values of two pixels i and j located at a distance d apart (measured in number of pixels) along a given direction θ [2,1]. Typically the GLCM is calculated for different pairs of d and θ. Figure 3 shows the spatial relationships between a pixel and its adjacent pixels, and the corresponding displacement vector (d, θ). Different statistical moments can be calculated from the GLCM ma-

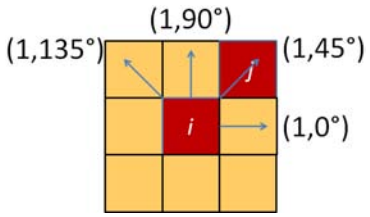


Fig. 3. Spatial relationships of a pixel i with its adjacent pixels

trix, such as contrast, homogeneity, squared energy, correlation and entropy [2]. In the present case we have used a lag $d = 1$ for the directions $0^o, 45^o, 90^o, 135^o$. This analysis provides an attribute vector of dimension 20 for each image.

Discrete Cosine Transform (DCT). DCT is a free-covariance model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image into cosines of increasing frequency. DCT has been used by Hafed and Levine [7] in face recognition, showing that DCT applied to normalized images is very robust to variations in geometry and lightning.

Mathematically, DCT is a discrete Fourier transform that expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes. For an image I_k the DCT is defined as follows:

$$D(u,v) = c(u)c(v) \sum_{i=0}^{s-1} \sum_{j=0}^{n-1} D_{(i,j)}$$

where

$$D_{(i,j)} = I_k(i,j) \cdot \cos \frac{\pi(2i+1)u}{2s} \cos \frac{\pi(2j+1)v}{2n},$$

$u = 0, ..., s-1$, and $v = 0, ..., n-1$, being

$$c(\alpha) = \begin{cases} \frac{1}{\sqrt{N}}, \ if \ \alpha = 0, \\ \sqrt{\frac{2}{N}}, if \ \alpha \neq 0. \end{cases}$$

N is either the number of rows (s) or columns (n) of the image. The DCT can be expressed in matrix form as an orthogonal transformation

$$D_{CT} = U_{DC} I_k V_{DC}^T,$$

where matrices U_{DC} and V_{DC} are orthogonal. This transformation is separable and can be defined in higher dimensions. The feature vector of an image I_k is constituted by the $q_1 - q_2$ block of D_{CT}, $D_{CT}(1 : q_1, 1 : q_2)$, where q_1, q_2 are determined by energy reconstruction considerations using the Frobenius norm of the image I_k.

Discrete Wavelet Transform (DWT). Wavelets are compact functions (defined over a finite interval) with zero mean and some regularity conditions (vanishing moments). The Wavelet transform converts a function into a linear combination of basic functions, called wavelets, obtained from a prototype wavelet through dilatations, contractions and translations. DWT was applied to face recognition by Kakarwal and Deshmukh [8].

The discrete wavelet transform (D_{WT}) of an image $I \in \mathcal{M}(m,n)$ is defined as follows:

$$D_{WT} = U_W^T I V_W,$$

where U_w and V_w are two orthogonal matrices constructed as follows:

$$U_W = \left[\frac{H}{G}\right]_m^T, V_W = \left[\frac{H}{G}\right]_n^T,$$

where H represents a low pass or averaging portion of the wavelet filter, and G is the high pass or differencing portion. In the present case we have used the transform having a maximum number of vanishing moments: the Daubechies-2 family.

4 Numerical Results

To perform the numerical analysis we have used the ORL database of faces provided by AT&T Laboratories Cambridge. The ORL database contains 400 grey scale images, ten different poses of 40 distinct individuals taken during a period of two years. All the images were taken against a dark homogeneous background, varying the lighting, facial expressions and facial details. The database provides upright and frontal poses. The size of each image is 92x112 pixels, with 256 grey levels per pixel. In all the experiments over ORL, the learning database is composed of five poses of each individual, that are randomly selected. In the supervised learning procedure three of the poses are used to learn and the other two are used to optimize the learning parameters. The rest of the poses in the database are used as probe images for establishing the accuracy of the classification for each spectral technique, using both, global and local features. For each attribute the classification is performed 100 different times, randomly choosing the learning database and the set of probe images (200 images). Table 1 shows shows the median accuracy obtained for each of the individual attributes. The higher accuracies are obtained for the local histogram (98%), the DWT (95.5%) and the DCT (95.25%). The variogram and the texture analysis provides lower median accuracies. Using the supervised learning technique explained above a median accuracy of 100% is obtained. Table 2 shows the parameters provided by the aligned PSO, to perform the classification.

Table 1. Accuracies of each individual attribute

Classifier (Attribute)	Histogram	Variogram	Texture	DWT	DCT
Median Accuracy	98.00	90.00	90.75	95.50	95.25

Table 2. Parameters for supervised learning algorithm optimized by aPSO

Attribute	Histogram	Variogram	Texture	DWT	DCT	Nfirst
Weights	65.21	61.84	80.15	80.02	62.14	6.00

5 Conclusions

In this paper we present a novel PSO version, the aligned PSO that is deduced from the PSO continuous model adopting for the trajectories a discretization that it is in accord to the velocity discretization terms. This approach is currently used in diffusion-convection problems. First and second order stochastic stability analysis is also performed following the methodology that was used in the past for other PSO family members. Numerical experiments with benchmark functions show that performing aligned PSO parameter sets of inertia and global and local accelerations are located close to the upper limit of the second order stability region where exploration capabilities are very high. Finally the aligned PSO is applied to optimize the parameters of a supervised learning algorithm to solve the face recognition problem. Its application to the ORL database provides 100% median identification accuracy over 100 independent runs.

References

1. Aksoy, S., Haralick, R.M.: Content-based image database retrieval using variances of gray level spatial dependencies. In: Ip, H.H.-S., Smeulders, A.W.M. (eds.) MINAR 1998. LNCS, vol. 1464, pp. 3–19. Springer, Heidelberg (1998)
2. Bharati, M.H., Jay Liu, J., MacGregor, J.F.: Image texture analysis: methods and comparisons. Chemometrics and Intelligent Laboratory Systems 72(1), 57–71 (2004)
3. Fernández-Martínez, J.L., García-Gonzalo, E.: The generalized PSO: a new door to PSO evolution. J. of Artif. Evol. and Appl. 2008, 1–15 (2008)
4. Fernández-Martínez, J.L., García-Gonzalo, E.: The PSO family: deduction, stochastic analysis and comparison. Swarm Intell. 3(4), 245–273 (2009)
5. Fernández-Martínez, J.L., García-Gonzalo, E.: Stochastic stability analysis of the linear continuous and discrete PSO models. IEEE Trans. Evol. Comput. 15(3), 405–423 (2011)
6. Goovaerts, P.: Geostatistics for natural resources evaluation. Applied geostatistics series. Oxford University Press (1997) (Incorporated)
7. Hafed, Z.M., Levine, M.D.: Face recognition using the discrete cosine transform. Int. J. Comput. Vision 43(3), 167–188 (2001)
8. Kakarwal, S., Deshmukh, R.: Wavelet transform based feature extraction for face recognition. International Journal of Computer Science and Application Issue 1, 100–104 (2010)
9. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings IEEE International Conference on Neural Networks (ICNN 1995), Perth, WA, Australia, vol. 4, pp. 1942–1948 (November-December 1995)
10. Srinivasan, G.N., Shobha, G.: Statistical texture analysis. In: Engineering And Technology Proceedings of World Academy of Science, vol. 36, pp. 1264–12699 (2008)
11. Umbaugh, S.E.: Computer Vision and Image Processing: A Practical Approach Using CVIPtools. Prentice Hall Professional Technical Reference (1998)

Appendix C

Book Chapter

We also presented a chapter for the book *Nature Inspired Computing: Theory and Industrial Application* to be published by Springer in the Studies in Computational Intelligence series, entitled *Image Processing Methods for Biometric Applications*. This chapter summarises part of the research presented in these three papers, and it is included in the Appendix of this thesis.

C.1. Image Processing Methods for Biometric Applications

Image Processing Methods for Biometric Applications

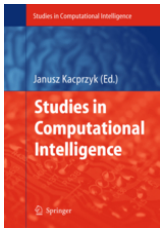
Authors:
Ana Cernea and Juan Luis Fernández Martínez

Book Chapter
Accepted to be published as a chapter on the edited "Nature Inspired Computing : Theory and Industrial Application" to be published by Springer in the *Studies in Computational Intelligence series*
2013

To whom it may concern

Dear Ana Cernea and J. L. Fernández-Martínez

Your book chapter on " Image Processing Methods for Biometric Applications" is accepted to be published as a chapter on the edited Book titled "Nature Inspired Computing : Theory and Industrial Application" to be published by Springer in the Studies in Computational Intelligence series.



With Best Regards

On behalf of

Guest Editors

Bijaya Ketan Panigrahi
Department of Electrical Engineering
IIT Delhi, Delhi, INDIA - 110016
E-mail : bijayaketan.panigrahi@gmail.com

Zhihua Cui
Complex System and Computational Intelligence Laboratory,
Taiyuan University of Science and Technology,
No.66, Wa-liu Road, Wanbailin District,
Taiyuan, Shanxi, P.R.China, 030024
E-mail: cuizhihua@gmail.com

Editorial Board / Review Board Members

Biometry is one of the most important fields of application of image processing techniques and have a great interest in science and technology. Security authentication and biomedical applications are fields where biometric applications are widely used.

This book chapter is a general overview of the main biometric applications, developing in more detail the face recognition problem and presenting some of the most successfully used image processing methods.

The image attributes revisited and analyzed in this paper can be divided into statistical attributes, segmentation attributes, and spectral decomposition methods, with applications to the face recognition problem.

Classification is then performed using each image attribute as individual classifiers, using different norms and/or the cosine defined by the Euclidean scalar product in the space of image attributes. Also, two kinds of analysis (global and local) are carried out for each attribute. The numerical results of this analysis performed over the ORL database showed the following conclusions:

- The maximum median accuracy is provided by the local histogram (98% using the cosine criterion) followed by the local percentiles (95.5% using the Euclidean norm).
- The crucial point for all the methods shown in this paper lies in choosing the right parameters to perform the different analysis and to provide the different attributes of the images. By tuning these parameter correctly we should be able to improve the classification accuracy. This fact reinforced the use of supervised methods to learn these parameters and encouraged our posterior researches devoted to this subject, with applications in different fields.

Image Processing Methods for Biometric Applications

Ana Cernea and J. L. Fernández-Martínez

Abstract Automatic image interpretation and recognition is a challenging problem in artificial intelligence, pattern recognition and computer vision with applications in security authentication, biometry, and biomedicine, to name some. In this paper we revisit some of the main image processing methods such as statistical, segmentation and spectral decomposition methods applied to the face recognition problem. Classification is performed using different norms and/or the cosine defined by the Euclidean scalar product in the space of image attributes. The local histogram is the attribute that provides the highest median accuracy rate. No method is able to provide systematically 100% accuracy rate operating stand-alone. This fact outlines the need of supervised learning methods combining different kinds of attributes to improve the classification. Future research will be devoted to this important subject.

1 Introduction

Biometry is one of the most important fields of application of image processing techniques and have a great interest in science and technology. Security authentication [10] and biomedical applications [28] are fields where biometric techniques are widely used. One of the main challenges in dealing with very large databases of images is establishing a low-dimensional feature representation having enough discriminatory power to perform high accuracy classification and/or prediction with the corresponding uncertainty quantification. Many different methods of supervised and unsupervised classification have been already presented in the literature, but no numerical comparison between all of these methods has been performed using the same database and the same training conditions.

Mathematics Department, Oviedo University
C/ Calvo Sotelo s/n 33007 Oviedo, Spain, e-mail: cerneadoina@uniovi.es and e-mail: jlfm@uniovi.es

In this paper we present a numerical comparison between three different types of attributes used for image representation: statistical attributes, spectral attributes, and segmentation features and region description. Classification is performed using non-supervised methods in the attribute space. These algorithms are applied to the face recognition problem, using the ORL database which is publicly available and contains a set of 400 face images corresponding to 40 different individuals. These algorithms could be applied to other interesting biometric problems such as iris recognition, fingerprint identification, etc.

This paper is organized as follows: in section 2 we revisit some important biometric applications such as face recognition, iris recognition, fingerprint identification and signature analysis. In section 3 we introduce the face recognition problem. Section 4 is devoted to present some statistical methods for image processing. In section 5 we present some segmentation and region descriptors commonly used in the literature concerning this subject. In section 6 we explain why orthogonal transformations of images are very useful in the dimensionality reduction, and we present some important spectral attributes such as Principal Component Analysis (PCA), 2DPCA, Fisher's LDA and other free-covariance techniques such as the Discrete Cosine Transform (DCT), the Discrete Wavelet Transform (DWT) and the Discrete Walsh-Hadamard Transform (DWHT). These orthogonal transformations are very fast to compute for every individual image of the database. Finally, in section 7 we show the application of these algorithms to the face recognition problem outlining the conclusions. The need of ensemble supervised methods to improve the identification accuracy is one of the main conclusions of this analysis.

2 Biometric Applications

Biometrics is the automatic person identification using physiological or behavioral characteristics, such as fingerprint, face, iris, retina, palm-print, etc. Among all the biometric techniques, fingerprint recognition is the most popular method and is successfully used in many authentication applications [13]. In this section we briefly describe the main biometric techniques, their peculiarities and the methods that are used in each case to accomplish the corresponding identification problem.

2.1 Face Recognition

The requirement for reliable personal identification in computerized access control has resulted in an increased interest in biometrics. Among all biometrics applications, face recognition has the benefit of being a passive, non intrusive system for verifying personal identity. A survey of popular face recognition techniques can be found in [33, 39].

Although it seems to be natural to identify persons according to their face, there are numerous factors that influence the performance of face recognition techniques such as, background and lighting conditions; intrinsic factors due to the physical nature of the face (age, facial expression, hair, glasses,etc); and finally the database size and composition concerning the different poses of each individual. Nevertheless, in practice the face recognition problem usually deals with images taken in similar conditions. Although each biometric application has its own techniques, the methodologies shown here for the face recognition problem can be applied to other biometric applications shown in this paper. In fact, in each case the problem to solve is similar: finding the optimum attribute description of the biometric image to accomplish at the same time the dimensionality reduction and the accuracy on the identification. Also in real authentication systems these methods have to be fast when applied to very large databases.

2.2 Iris Recognition

Iris recognition is an automated method of biometric identification that uses pattern-recognition techniques on images of the irides of an individual's eyes. Of all the biometric technologies used for human authentication, iris recognition is considered to be one of the most accurate non-invasive techniques, due to the following characteristics:

- stability - the unique pattern in the human iris is considered to remain unchanged throughout one's lifetime.
- uniqueness - the probability of finding two identical patterns is nearly impossible.
- reliability - since a distinctive iris pattern is not susceptible to false authentication (theft, loss, etc).

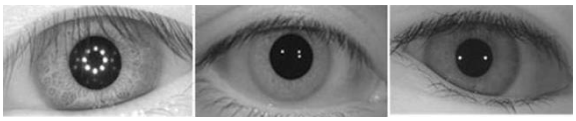


Fig. 1 Iris samples for iris recognition from Casia Iris database.

The first operational automatic iris recognition system was developed by Daugman in 1993 [12, 11], using 2D Gabor wavelets. The Hamming distance between the iris code of the test and the training iris images was used for recognition. An other important approach is proposed by Wildes and consists of computing a binary edge map followed by a Hough transform to detect circles. For matching two irises, Wildes applies a Laplacian of Gaussian filter at multiple scales to produce a template and computes the normalized correlation as a similarity measure [36]. A detailed survey of recent efforts on iris recognition can be found in [7].

2.3 Fingerprint Identification

Fingerprint identification is the process of comparing two instances of friction ridge skin impressions, from human fingers to determine whether these impressions could have come from the same individual.

The major fingerprint features (also called minutia) are: ridge ending, bifurcation, and short ridge (or dot). The ridge ending are points where ridges terminate. Typical fingerprint recognition methods employ feature-based image matching, where different minutia are extracted, both from the registered and the input fingerprint image [3, 25]. The number of corresponding pairings between these two images is used to recognize a valid fingerprint image [13].

Among the most widely used techniques for fingerprint identification are:

- Minutiae Extraction Technique which represent the fingerprint by its local features, like terminations and bifurcations [14]
- Pattern Matching or Ridge Feature - using the ridge feature in form of fingercode which is a vector obtained by applying a convolution to the original fingerprint image with a Gabor filter [22]
- combinations of the above techniques also with different image processing methods. Also, pattern based algorithms can be used, and they compare the basic fingerprint patterns (arch, whorl, and loop) between a previously stored template and a candidate fingerprint [26, 9].

Fig. 2 Fingerprints samples from Casia Fingerprints Database.

2.4 Signature Recognition

Signature recognition can be divided into two main areas depending on the method of data acquisition: on-line and off-line signature recognition [29], [40]. In off-line signature recognition, signature is available on a document which is scanned to get the digital image representation. The on-line signature recognition, also known as dynamic signature uses special hardware, such as a digitizing tablet or a pressure sensitive pen to acquire the signature in real time.

The most significant benefit of Signature Recognition is that it is reliable, since it is very difficult to imitate the behavioural patterns inherent to the process of signing.

Fig. 3 Signatures samples from Casia Signatures Database.

The most popular pattern recognition techniques applied for signature recognition are dynamic time warping algorithm and Gaussian Mixture Models, and Hidden Markov Models. Detailed surveys of the current techniques can be consulted in [31, 20]

3 The Face recognition Problem

The face recognition problem consists in, given a new incoming face image, identifying the individual corresponding to it, from a database of different face images of known individuals. The learning database (Bd) contains N face images

$$Bd = \{I_{k=1,\dots N} \in \mathscr{M}_{m\times n}(\mathbf{N})\},$$

corresponding to C individuals (or classes). We will assume that for each of these individuals the learning database contains n_p poses that are used to perform the classification of the new input images, and the total number of poses in Bd verifies $N = n_p \cdot C$. Typically we dispose of $n_p = 5$ poses for each individual.

In order to perform the classification, images are represented by a feature vector $\mathbf{v}_k \in \mathbf{R}^s$, obtained for each method of analysis. Faces are described by different kind of image attributes: statistical based features, spectral features, and image segmentation/regional descriptors features (texture-based features). The dimension s of $\mathbf{v}_k$ depends on the dimensionality reduction achieved by each technique. In the case of the spectral methods we have used the same energy threshold $\theta = 99.5\%$ to achieve the dimensionality reduction. All these attributes can be calculated for gray scale and color images, both, locally or globally. In the case of global analysis the attribute features are calculated over the whole size of the image. In the case of local features, this analysis is performed by dividing the images into blocks. For each block the local attributes are computed and the final feature vector is formed by merging all the local attributes into an unique vector, always computed in the same order. Figure 4 shows the sketch of this process. The use of local features increases the dimension of the attribute space, nevertheless it is also expected to increase the discriminative power of the analysis. In this paper we have used a partition of the images into 8×4 blocks, nevertheless finer subdivisions could be also adopted.

Fig. 4 Image partition in blocks to performing local analysis

The classification algorithm proceeds as follows:

- given a new image $I \in \mathscr{M}_{m\times n}(\mathbf{N})$, $I \notin Bd$, corresponding to a new pose of any of the C individuals of the database, characterized by its features vector $\mathbf{v}_I$, finding the most similar image in the database Bd, according to a criterion defined over the space of attributes $\mathbf{R}^s$.

  Two different criteria are used in this paper:

- Minimize the distance between the new image J and the database images in a certain norm p, defined over $\mathbf{R}^s$: $d(I,I_k) = \|\mathbf{v}_I - \mathbf{v}_k\|_p$. In this paper we have used the Euclidean norm ($p = 2$) and also other p norms such as $p = 3,5$.
- Maximize the cosine of the angle between the above mentioned images represented by their attributes vectors:

$$\cos(I,I_k) = \frac{\mathbf{v}_I \cdot \mathbf{v}_k}{\|\mathbf{v}_I\|_2 \|\mathbf{v}_k\|_2}.$$

  The cosine criterion is related to the existence of a dot product (for instance for p=2).

Although both criteria applied to the same image are equivalent, they provide different recognition accuracies when applied over the whole set of images (testing database). In the cases where $\cos(I,I_k)$ provides higher accuracies, that means this criterion performs better with normalized images $\mathbf{n}_I, \mathbf{n}_{I_k}$, since

$$cos(I,I_k) = \frac{\mathbf{v}_I}{\|\mathbf{v}_I\|} \cdot \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|} = \mathbf{n}_I \cdot \mathbf{n}_{I_k}.$$

The algorithm presented in this paper belongs to the class of non-supervised classification since the distance $d(J,I_k)$, and the cosine $\cos(J,I_k)$, are defined ad-hoc, that is, no information from the learning database classes is used to optimize the classification of the testing database.

4 Statistical Methods for Image Processing

4.1 Color Histograms

The image histogram is the simplest attribute that can be calculated and describes the frequency of the brightness in the image. The shape of the histogram provides information about the nature of the image. For example, a narrow histogram means a low contrast, meanwhile a histogram shifted to the right indicates a bright image; a histogram with two major peaks (called bimodal) implies the presence of an object that is in contrast with the background, etc [35]. Due to this fact it is expected that the histogram will have a great discriminative power. Also, this attribute is fast to be computed and can be precalculated for all the images in the learning database.

For a gray-scale digital image I the histogram represents the discrete probability distribution of the gray-levels in the image. For this purpose the gray-scale space ($[0,255]$ for an `8-bit` image) is divided into L bins, and the number of pixels in each class n_i, $(i = 1,L)$ is calculated. In this case the attribute vector has dimension L:

$$H_I = (n_1, ..., n_L).$$

Relative frequencies can be also used by dividing the absolute frequencies n_i by the total number of pixels in the image.

In the case of *RGB* images the same analysis can be performed considering the color channels I_R, I_G and I_B independently, and merging all the channels histograms together, as follows:

$$H_I = (H(I_R), H(I_G), H(I_B)).$$

The histogram can be calculated either globally over the whole image or locally. Global histograms do not capture the spatial distribution of the color channels. Local histograms provide this kind of information. Their calculation follows the general procedure for computing any local attribute explained above (see Figure 4).

4.2 Percentiles

The p-percentile of a color (or gray) channel c_i is defined as the number x_p such as

$$P(c_i \leq x_p) = p \in [0,1].$$

The p-percentiles are related to the cumulative probability functions of each color channel. In practice, we compute the percentiles $1\%, 99\%$ and 5% to 95% with 5% step of probability. This produces a feature vector of dimension 21 for each color channel.

Percentiles and histograms are related because the histogram is the density probability function $f(t)$ of each color channel, and the cumulative probability function $F(x)$ is the integral:

$$F(x) = \int_{-\infty}^{x} f(t)dt \Leftrightarrow f(x) = F'(x).$$

Nevertheless, histograms and percentiles provide different results for classification purposes.

4.3 Variogram

The variogram describes the spatial distribution in each color channel. In spatial statistics [17] the variogram describes the degree of spatial dependence of a spatial random field or stochastic process, the gray-scale in this case. For a given value of vector h, defined by a modulus and direction, the variogram is an index of dissimilarity between all pairs of values separated by vector h.

The omnidirectional p-variogram is the mean of the p-absolute difference between the color values of the $N(h)$ pairs of pixels that are located at the same distance h:

$$\gamma_i(h) = \frac{1}{N(h)} \sum_{k=1}^{N(h)} |c_i(x_k) - c_i(x_k + h)|^p. \quad (1)$$

Usually $p = 2$. To compute the variogram each color channel (matrix) is transformed into the corresponding color vector $c_i(x)$. Typically $N(h)$ is limited to one third of the total number of pixels. The number of classes that have been considered in this case was $N(h) = 100$.

The variogram could also be considered as an image texture descriptor. The texture attributes are described in the following section.

5 Segmentation features and region descriptors

5.1 Texture Analysis

Texture analysis of an image consists in analyzing regular repetitions of a pattern [16]. For texture analysis we use the spatial gray level co-occurrence matrix which is based on the joint probability distributions of pairs of pixels.

The gray level co-occurrence matrix (GLCM), or spatial dependence matrix of an image I is an estimate of the second-order joint probability function $P_{d,\theta}(i, j)$ of the intensity values of two pixels i and j located at a distance d apart (measured in number of pixels) along a given direction θ [6][2]. Typically the GLCM is calculated

for different pairs of d and θ. Figure 5 shows the spatial relationships between a pixel and its adjacent pixels, and the corresponding displacement vector (d,θ).

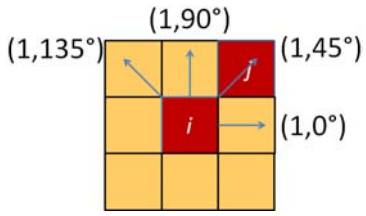


Fig. 5 Spatial relationships of a pixel i with its adjacent pixels.

To calculate the GLCM matrix for a given pair (d,θ) the algorithm proceeds as follows:

- First the matrix $F_{d,\theta}(i,j) \in \mathscr{M}_{(n\times n)}$ of absolute frequencies of pairs of pixels with gray levels i and j at a distance d in the direction θ is built. For instance, figure 6 shows the $F_{d,\theta}$ matrix for a 4×4 image with 5 gray levels, for the displacement vector $(1,0)$.
- Secondly, $F_{d,\theta}$ is normalized as follows:

$$P_{d,\theta}(i,j) = \frac{F_{d,\theta}(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} F_{d,\theta}(i,j)} \tag{2}$$

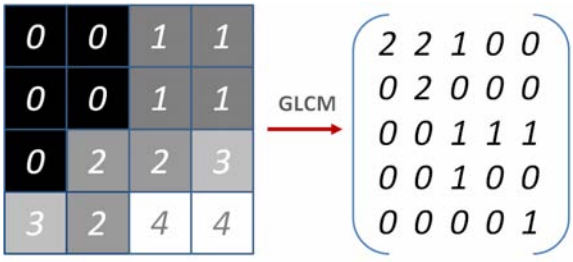


Fig. 6 $F_{d,\theta}$ for a 4×4 image with 5 gray levels.

Different statistical moments can be calculated from the GLCM matrix [6]:

- Contrast is a measure of local image variation, typically it means the intensity change between a pixel and its neighbor over the whole image.

$$C = \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j|^2 P_{d,\theta}(i,j).$$

Contrast is zero for a constant image.

- Homogeneity measures the proximity of the distribution of the GLCM matrix to its diagonal.
$$H = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{P_{d,\theta}(i,j)}{1+|i-j|}$$
- Squared energy
$$E = \|P_{d,\theta}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}P_{d,\theta}(i,j)^2$$
which is the square of the Frobenius norm of the GLCM matrix.
- Correlation
$$\rho = \frac{cov(i,j)}{\sigma_i\sigma_j} = \sum_{i=1}^{n}\sum_{j=1}^{n}P_{d,\theta}(i,j)\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i\sigma_j}$$
where
$$\mu_i = \sum_{i=1}^{n}\sum_{j=1}^{n}iP_{d,\theta}(i,j),$$
$$\mu_j = \sum_{i=1}^{n}\sum_{j=1}^{n}jP_{d,\theta}(i,j),$$
$$\sigma_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(i-\mu_i)^2P_{d,\theta}(i,j),$$
$$\sigma_j^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(j-\mu_j)^2P_{d,\theta}(i,j).$$
- Entropy
$$S = -\sum_{i=1}^{n}\sum_{j=1}^{n}P_{d,\theta}(i,j)\log P_{d,\theta}(i,j)$$

In the present case we have used a lag $d = 1$ for the directions $0, 45, 90, 135$. This analysis provides an attribute vector of dimension 20 for each image.

5.2 Edges Detection

Edges are determined by sets of pixels where there is an abrupt change of intensity. If a pixel's gray level value is similar to those around it, there is probably not an edge at that point. However, if a pixel has neighbors with widely varying gray levels, it may represent an edge. Thus, an edge is defined by a discontinuity in the gray-level values [35]. More precisely, we can consider an edge as a property associated to a pixel where the image function $f(x,y)$ changes rapidly in the neighborhood of that pixel. In this case the function $f : \mathbb{R}^2 \to \mathbb{R}$ represents the pixels intensities. Related

to f, an edge is a vector variable with two components: magnitude and direction. The edge magnitude is given by the gradient and its direction is perpendicular to the gradient's direction:

$$|\nabla f(x,y)| = \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2},$$
$$\theta(x,y) = arctg\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right) \pm \frac{\pi}{2}.$$

Fig. 7 Canny edges detection method applied to a face sample from the ORL database.

In the case of digital images, these expressions are approximated by differences schemes. The gradient can be computed using different convolution masks, such as Robert, Sobel, Prewitt, Laplace, Canny, etc. To perform the classification we have used the Canny edge detection operator [8], which is one of the most commonly used in image processing, due to its property of detecting edges in a very robust manner in the case of noisy images. Figure 7 shows the edges of a face image from the ORL database, obtained by applying the Canny operator.

6 Spectral Attributes

In this section we briefly introduce the spectral methods that are used in this paper. Spectral decomposition methods consist in finding an orthonormal base which best separates the image projections and to reduce the dimensionality of the attributes space with respect to the pixels space $\mathbb{R}^{Npixels}$. Spectral methods can be divided into two categories: 1) those that act in the whole database of images (PCA, 2DPCA and Fisher's LDA); 2) those that act on single images, although they could be also applied to the whole training image database, such as: DCT, DWT and DWHT.

6.1 Linear Operators and Orthogonal Transformations

A grey digital image $I(m,n)$ can be regarded as the matrix of a linear operator between two linear spaces $\mathbb{R}^n$ and $\mathbb{R}^m$. In the case of a color image there exist three different linear operators, one for each color channel (R, G and B).

Given an image $I(m,n)$, it is possible to define in several ways, two different orthogonal transformations U, V, such as:

$$I = USV^T, \tag{3}$$

where U, V form orthonormal basis of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively.

The interests of the orthogonal transformations are the following:

- They conserve the energy of the image, that is the Frobenius norm of matrix I. The demonstration is as follows:

$$\|I\|_F = \sqrt{trace(I^T I)} = \sqrt{trace(VS^T SV^T)} = \sqrt{trace(S^T S)} = \|S\|_F.$$

- Matrix S is blocky diagonal if U and V contain as columns the left and right singular vectors provided by the singular value decomposition of I. In other cases, such as DCT and DWT, these orthogonal transformations serve to decorrelate the pixels of I by compressing its energy onto the first harmonics (spectral modes of the image). The pixel decorrelation is based on the fact that orthogonal transformations (in our case U and V) induce rotations to the principal axes of the image in $\mathbb{R}^m$ and $\mathbb{R}^n$. In the case of PCA, this rotation is induced by the orthonormal basis calculated through the experimental covariance matrix.
- Energy compression consists in finding the number of transformed pixels p, q of S (1), such as:

$$\|I - S(1:p, 1:q)\|_F < \theta$$

  where θ is a prescribed energy threshold, and $S(1:p, 1:q)$ represents the $p \times q$ upper block of S. The dimensionality reduction is achieved from $m \times n$ pixels to $p \times q$ frequency components. Interesting to remark that if the SVD were used only the q first singular values are needed, because S is blocky - diagonal.

Orthonormal transformations based on a set of training images $\{I_1, I_2, ..., I_N\}$, follow similar principles.

6.2 Covariance-based Spectral Methods

In this part we describe a set of spectral techniques for dimensionality reduction that are based on the experimental covariance of a set of images belonging to the learning database. Thus, these techniques are covariance-based, and the reduced base has to be recalculated when the learning database changes due to the addition or suppression of images. Both are the main drawbacks of these methods.

6.2.1 Principal Component Analysis (Eigenfaces)

Principal component analysis (PCA) is a popular unsupervised method which aims at finding a linear subspace in which the variance of the projected data is maximized. PCA applied to image analysis and face recognition [24, 33, 34] is an orthogonal transformation aimed at decorrelating the pixels of a given image, concentrating its energy in a small number of principal components, and reducing the face recognition problem dimensionality. Eigenfaces has been widely investigated and became one of the most popular approaches in face recognition [27]. PCA provides high accuracy recognition rates and it is a fairly robust method to lighting variations, nevertheless its performance degrades with scale changes [21].

PCA finds an orthonormal basis of the *pixel space* ($\mathbb{R}^{Npixels}$) by diagonalizing the mean centered empirical image covariance matrix, $S \in \mathscr{M}_{Npixels \times Npixels}$, constructed with the whole database as follows:

$$S = \sum_{k=1}^{N} (X_k - \mu)(X_k - \mu)^T = X_c X_c^T,$$

where $\mu = \frac{1}{N}\sum_{k=1}^{N} X_k$ is the images sample mean, N is the number of sample images contained in the learning database, N_{pixels} is the number of pixels of each image (we suppose that all images in the sample have the same dimensions), $X = [X_1, X_2, ..., X_N] \in \mathscr{M}_{Npixels \times N}$, where $X_i \in \mathbb{R}^{Npixels}$ are the database images transformed into $1-D$ column vectors, and $X_c = X - \mu$ is the centered image matrix. S is a symmetric semidefinite positive matrix, thus, it admits orthogonal diagonalization as follows:

$$S = U D_1 U^T,$$

where $U \in \mathscr{M}_{Npixels \times Npixels}$.

The dimensionality reduction is obtained by retaining the q first principal components $\mathbf{u}_k \in U$, such as:

$$|\sigma_{tot}^2 - \sum_{k=1}^{q} \lambda_k| < \theta,$$

where θ is the energy cutoff and λ_k are the non-null eigenvalues of D_1. Finally, defining $W = U(:, 1:q)$, each image X_i is projected onto the base of the eigenvectors of the first q eigenvalues of U, obtaining the feature vector $Y_i = W^T X_i$, and achieving a dimensionality reduction from N_{pixels} to the q first principal coordinates.

6.2.2 2DPCA

2DPCA for face recognition has been introduced by Yang et al. [37]. As opposed to conventional PCA, 2DPCA is based on $2D$ matrices, I_i, rather than $1D$ vectors, X_i. That is, each image I_i does not need to be previously transformed into a column vector and image covariance is constructed directly using the original image matrices. This covariance matrix is, in contrast with the scatter matrix of PCA, much smaller.

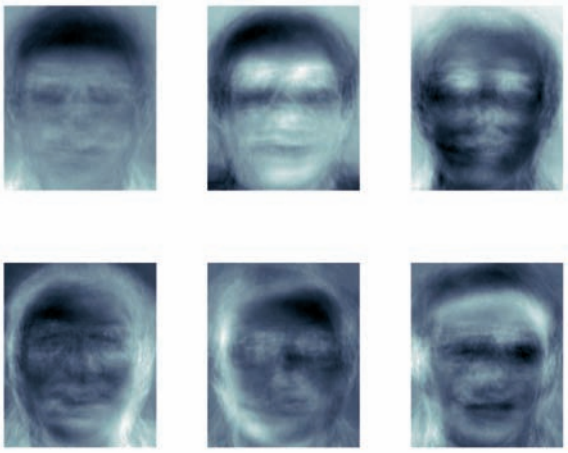


Fig. 8 The first six eigenfaces for the ORL training database

The 2DPCA diagonalizes the following mean centered covariance matrix to find the orthogonal projection basis:

$$S_M = \frac{1}{N}\sum_{i=1}^{N}(I_i - \bar{I})^T(I_i - \bar{I}),$$

where $\bar{I}$ is the mean image matrix calculated pixel by pixel over the learning database. As S_M is the sum of symmetric and semidefinite matrix, it admits orthogonal diagonalization like in the PCA case. The dimensionality reduction and the feature projection follows the same logic than for PCA.

6.2.3 Fisher's Linear Discriminant (Fisherfaces)

Fisher Linear Discriminant Analysis (LDA) [15], for short Fisherfaces has been applied to the face recognition problem by Belhumeur [5]. This technique is also based on a linearly projection W, which attempts to reduce dimensionality while preserving as much of the class discriminatory information as possible. More precisely, the transformation W is selected such as the projections of the sample database X have the maximum degree of separation. Fisher's solution to this problem is the calculation of W which maximizes the differences between classes, normalized by the scatter within-class. For that purpose, between- and within-class scatter matrices are defined:

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T,$$

$$S_W = \sum_{i=1}^{C} \sum_{X_k \in C_i} (X_k - \mu_i)(X_k - \mu_i)^T,$$

where C is the total number of classes in the database, N_i is the number of images in each class C_i, and μ_i is the mean of the images in class i. Numerical implementation of this method can be consulted in [38]. Numerical experimentation has shown that Fisherfaces have error rates lower than eigenfaces in face recognition [5]. This is due to the fact that Fisher's LDA is a supervised projection method.

Fig. 9 The first six Fisher's faces for the ORL training database

6.3 Covariance-free Spectral Methods

Conversely to the above described spectral methods, these techniques do not involve diagonalization, and the reduced bases (and the corresponding projections of the images) are calculated for each individual image. Thus, these techniques are more suitable to be implemented in large databases concerning real authentication systems.

6.3.1 Discrete Cosine Transform (DCT)

DCT is a free-covariance model reduction technique that attempts to decorrelate 2D images by projecting the rows and columns of the incoming image into cosines of increasing frequency. DCT has been used by Hafed and Levine [18] in face recognition, showing that DCT applied to normalized images is very robust to variations in geometry and lightning.

Mathematically, DCT is a discrete Fourier transform operating on real data, by expressing a signal in terms of a sum of sinusoids with different frequencies and amplitudes. For an image I_k the DCT is defined as follows:

$$D(u,v) = c(u)c(v)\sum_{i=0}^{s-1}\sum_{j=0}^{n-1} D_{(i,j)}$$

where

$$D_{(i,j)} = I_k(i,j)\cdot\cos\frac{\pi(2i+1)u}{2s}\cos\frac{\pi(2j+1)v}{2n},$$

$u = 0,...,s-1$, and $v = 0,...,n-1$, being

$$c(\alpha) = \begin{cases} \frac{1}{\sqrt{N}}, \ if \ \alpha = 0, \\ \sqrt{\frac{2}{N}}, \ if \ \alpha \neq 0. \end{cases}$$

N is either the number of rows (s) or columns (n) of the image. The DCT can be expressed in matrix form as an orthogonal transformation

$$D_{CT} = U_{DC} I_k V_{DC}^T,$$

where matrices U_{DC} and V_{DC} are orthogonal. This transformation is separable and can be defined in higher dimensions. The feature vector of an image I_k is constituted by the $q_1 - q_2$ block of D_{CT}, $D_{CT}(1:q_1, 1:q_2)$, where q_1, q_2 are determined by energy reconstruction considerations using the Frobenius norm of the image I_k. The energy methodology used to find the q_1, q_2 values is the same in all these spectral methods.

6.3.2 Discrete Wavelet Transform (DWT)

Wavelets are compact functions (defined over a finite interval) with zero mean and some regularity conditions (vanishing moments). The Wavelet transform converts a function into a linear combination of basic functions, called wavelets, obtained from a prototype wavelet through dilatations, contractions and translations. DWT was applied to face recognition by Kakarwal and Deshmukh [23].

The discrete wavelet transform (D_{WT}) of an image $I \in \mathscr{M}(m,n)$ is defined as follows:

$$D_{WT} = U_W^T I V_W,$$

where U_w and V_w are two orthogonal matrices constructed as follows:

$$U_W = \left[\frac{H}{G}\right]_m^T, V_W = \left[\frac{H}{G}\right]_n^T,$$

where H represents a low pass or averaging portion of the wavelet filter, and G is the high pass or differencing portion. In all the cases we have

$$D_{WT} = \begin{pmatrix} HIH^T & HIG^T \\ GIH^T & GIG^T \end{pmatrix} = \begin{pmatrix} B & V \\ H & D \end{pmatrix}.$$

B is the blur, V are the vertical differences, H are the horizontal differences and D are the diagonal differences. DWT can be applied several times to further reduce the dimension of the attribute vector. In the present case we apply DWT twice and we use the blur B as feature attribute to solve the face recognition problem. In the present case we have used the transform having a maximum number of vanishing moments: the Daubechies-2 family.

6.3.3 Discrete Walsh-Hadamard Transform (DWHT)

The Hadamard transform is the projection of a signal onto a set of square waves called Walsh functions. The Walsh functions are real and only take the values $+1$ and -1.

The Hadamard matrix of order n can be obtained by defining its element in the i^{th} row and j^{th} $(i, j = 0, 1, ..., 2^n - 1)$ column, as follows:

$$H_n(i, j) = (-1)^{\sum_{k=0}^{n-1} i_k j_k} = \prod_{k=0}^{n-1} (-1) i_k j_k,$$

where

$$i = \sum_{k=0}^{n-1} i_k 2^k = (i_{n-1} i_{n-2} ... i_1 i_0), (i_k = 0, 1),$$

$$j = \sum_{k=0}^{n-1} j_k 2^k = (j_{n-1} j_{n-2} ... j_1 j_0), (j_k = 0, 1),$$

are the binary representations of i and j, respectively [4].

The Hadamard matrix has the following orthogonality property

$$H_m H_m^T = 2^m I_{2^m}.$$

The DWHT of an image I with $M \times N$ pixels is defined as follows:

$$DWHT(u,v) = \sum_{j=1}^{N}\sum_{i=1}^{M} H_m(u,i)I(i,j)H_n(v,j) =$$
$$= \sum_{j=1}^{N}\sum_{i=1}^{M} I(i,j)(-1)^{\sum_{k=0}^{m-1} u_k i_k + \sum_{l=0}^{n-1} v_l j_l},$$

where u, v, i and j are represented in binary form, being $m = log_2 M$ and $n = log_2 N$ [30]. Furthermore, the DWHT can be written in matrix form:

$$DWHT(I) = H_m I H_n^T,$$

The DWHT performs an orthogonal, symmetric transform, that computationally is very fast and simple. Projecting an image onto the Walsh-Hadamard basis functions only involves multiplication of each pixel by ± 1 [19]. The DWHT has been used in data compression [1]. DWHT has been applied to face recognition problems by Rauf [32].

7 Numerical Results and Conclusions

To perform the numerical analysis we have used the ORL database of faces provided by AT&T Laboratories Cambridge. The ORL database contains 400 grey scale images, ten different poses of 40 distinct individuals taken during a period of two years. All the images were taken against a dark homogeneous background, varying the lighting, facial expressions and facial details. The database provides upright and frontal poses. The size of each image is 92x112 pixels, with 256 grey levels per pixel.

In all the experiments over ORL, the learning database is composed of five poses of each individual, that are randomly selected. The rest of the poses in the database are used as probe images for establishing the accuracy of the classification for each spectral technique, using both, global and local features. For each attribute the classification is performed 100 different times, randomly choosing the learning database and the set of probe images (200 images). Nevertheless, once the database has been generated, it is used to perform classification with all the different image processing methods under the same numerical conditions. For instance, in all the cases the energy cutoff used in spectral decomposition for the reconstruction has been fixed to 99.5% of the Frobenius norm of the transform.

Two different kind of analysis are performed using global and local attributes. Although the use of local features increases the dimension of the attribute space, it is expected to increase its discriminative power. Thus, in the algorithm presented in this paper, no individual classification with posterior fusion of scores is performed. Finally, we provide statistics for the classification calculated over 100 different simulations of each classifier(attribute): minimum, maximum, median and mean accu-

racy, and also interquartile range and standard deviation, for the distance computed using different norms and the cosine criteria.

Attribute	Type of analysis	Criterion	min	max	median	mean	IQR	std
Histogram	Local	cos	92.00	100.00	**98.00**	97.67	1.50	1.30
Variogram	Local	L_2	84.50	94.50	90.00	90.03	3.00	1.89
Percentiles	Local	L_2	90.50	98.50	95.50	95.13	2.00	1.61
Texture	Local	cos	84.50	93.50	91.00	90.80	2.50	1.83
Edges	Local	cos	53.50	68.50	62.75	62.80	4.00	2.96
PCA	Local	L_2	91.00	98.50	95.00	94.79	2.50	1.77
Fisher	Local	L_2	90.50	98.50	95.00	94.63	2.50	1.72
DCT	Local	L_3	89.00	97.50	94.00	93.80	2.50	1.85
2DPCA	Local	L_2	90.50	98.00	95.00	94.64	2.50	1.69
DWT	Global	L_2	91.00	99.00	95.50	95.44	2.00	1.58
DWHT	Local	L_3	88.50	97.50	94.50	93.90	2.75	1.91

Table 1 Accuracies statistics for the selected image attributes.

The conclusions of this analysis are:

- The maximum median accuracy is provided by the local histogram (98% using the cosine criterion) followed by the local percentiles ($95,5\%$ using the Euclidean norm).
- The variogram, texture and edges analysis do not provide very high accuracies, nevertheless these methods could be used in combination with the other descriptors to improve the classification.
- The discrete wavelet transform (DWT) is the spectral method that provides the highest accuracies ($95,5\%$) using the L_2-norm for global analysis. The rest of the methods perform better for local analysis. The discrete cosine transform (DCT) and the discrete Walsh-Hadamard transform (DWHT) provide better results using the L_3-norm.
- All the spectral methods perform more or less similarly. The reason why this happens is that orthogonal transformations are isometries, and all the transformed images (matrices) are related by similarity relationships of the kind $I = UDV^T$, where U and V are orthogonal matrices. Thus, all the similar matrices (I and D) contain the same information if no model reduction is performed.
- In our opinion the crucial point for all the methods shown in this paper lies in choosing the right parameters to perform the different analysis and to provide the different attributes of the images. By tuning these parameter correctly we should be able to improve the classification accuracy. This fact reinforces the use of supervised methods to learn these parameters. Further research will be devoted to this subject.

References

1. A.N. Akansu and R. Poluri. Walsh-like nonlinear phase orthogonal codes for direct sequence cdma communications. *Signal Processing, IEEE Transactions on*, 55(7):3800 –3806, july 2007.
2. Selim Aksoy and Robert M. Haralick. Content-based image database retrieval using variances of gray level spatial dependencies. In *In Proc. of IAPR Intl. Workshop on Multimedia Information Analysis and Retrieval*, pages 3–19, 1998.
3. A. Batool and A. Tariq. Computerized system for fingerprint identification for biometric security. In *Multitopic Conference (INMIC), 2011 IEEE 14th International*, pages 102–106, 2011.
4. K.G. Beauchamp. *Applications of Walsh and related functions, with an introduction to sequency theory*. Microelectronics and signal processing. Academic Press, 1984.
5. Peter N. Belhumeur, Jo˜ao Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In Bernard Buxton and Roberto Cipolla, editors, *Computer Vision ECCV '96*, volume 1064 of *Lecture Notes in Computer Science*, pages 43–58. Springer Berlin Heidelberg, 1996.
6. Manish H. Bharati, J.Jay Liu, and John F. MacGregor. Image texture analysis: methods and comparisons. *Chemometrics and Intelligent Laboratory Systems*, 72(1):57 – 71, 2004.
7. Kevin W. Bowyer, Karen Hollingsworth, and Patrick J. Flynn. Image understanding for iris biometrics: A survey. *Comput. Vis. Image Underst.*, 110(2):281–307, May 2008.
8. John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, 1986.
9. Raffaele Cappelli, Matteo Ferrara, and Davide Maltoni. Minutiae-based fingerprint matching. In *Cross Disciplinary Biometric Systems*, volume 37 of *Intelligent Systems Reference Library*, pages 117–150. Springer Berlin Heidelberg, 2012.
10. M.A. Dabbah, W.L. Woo, and S.S. Dlay. Secure authentication for face recognition. In *IEEE Symposium on Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007.*, pages 121 –126, april 2007.
11. J. G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1148–1161, November 1993.
12. J.G. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, 1988.
13. Anil K. Jain Davide Maltoni, Dario Maio and Salil Probhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.
14. Huimin Deng and Qiang Huo. Minutiae matching based fingerprint verification using delaunay triangulation and aligned-edge-guided triangle matching. In Takeo Kanade, Anil Jain, and NaliniK. Ratha, editors, *Audio- and Video-Based Biometric Person Authentication*, volume 3546 of *Lecture Notes in Computer Science*, pages 270–278. Springer Berlin Heidelberg, 2005.
15. Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen*, 7:179–188, 1936.
16. G.N.Srinivasan and Dr.G.Shobha. Statistical texture analysis. In *Proceedings Of World Academy Of Science, Engineering And Technology*, volume 36, pages 1264–12699, 2008.
17. P. Goovaerts. *Geostatistics for natural resources evaluation*. Applied geostatistics series. Oxford University Press, Incorporated, 1997.
18. Ziad M. Hafed and Martin D. Levine. Face recognition using the discrete cosine transform. *Int. J. Comput. Vision*, 43(3):167–188, July 2001.
19. M. Hassan, I. Osman, and M. Yahia. Walsh-hadamard transform for facial feature extraction in face recognition. *Int. J. of Comp. and Communication Eng.*, 1(7):436–440, 2007.
20. D. Impedovo and G. Pirlo. Automatic signature verification: The state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(5):609–635, 2008.

21. Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, june 2009.
22. A.K. Jain, S. Prabhakar, L. Hong, and S. Pankanti. Filterbank-based fingerprint matching. *Image Processing, IEEE Transactions on*, 9(5):846–859, 2000.
23. S. Kakarwal and R. Deshmukh. Wavelet transform based feature extraction for face recognition. *International Journal of Computer Science and Application Issue*, –:–, 2010.
24. M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 12(1):103–108, January 1990.
25. Dushyant Khosla Mary Lourde R. Fingerprint identification in biometric security systems. *International Journal of Computer and Electrical Engineering*, 2(5):852–855, october 2010.
26. M.M. Min and Y. Thein. Intelligent fingerprint recognition system by using geometry approach. In *Current Trends in Information Technology (CTIT), 2009 International Conference on the*, pages 1–5, 2009.
27. Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.
28. A. Nait-Ali. Hidden biometrics: Towards using biosignals and biomedical images for security applications. In *7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), 2011*, pages 352 –356, may 2011.
29. Réjean Plamondon and Sargur N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):63–84, January 2000.
30. W.K. Pratt, J. Kane, and H.C. Andrews. Hadamard transform image coding. *Proceedings of the IEEE*, 57(1):58 – 68, jan. 1969.
31. K. R. Radhika, G. N. Sekhar, and M. K. Venkatesha. Pattern recognition techniques in on-line hand written signature verification - a survey. In *Multimedia Computing and Systems, 2009. ICMCS '09. International Conference on*, pages 216–221, 2009.
32. Rauf Kh. Sadykhov, Vladimir A. Samokhval, and Leonid P. Podenok. Face recognition algorithm on the basis of truncated Walsh-Hadamard transform and synthetic discriminant functions. In *FGR*, pages 219–222. IEEE Computer Society, 2004.
33. L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987.
34. Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, January 1991.
35. Scott E Umbaugh. *Computer Vision and Image Processing: A Practical Approach Using CVIPtools*. Prentice Hall Professional Technical Reference, 1998.
36. R.P. Wildes. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363, 1997.
37. Jian Yang, David Zhang, Alejandro F. Frangi, and Jing yu Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:131–137, 2004.
38. Hua Yu and Jie Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.
39. W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, December 2003.
40. Alessandro Zimmer and Lee Luan Ling. A hybrid on/off line handwritten signature verification system. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1*, ICDAR '03, pages 424–, Washington, DC, USA, 2003. IEEE Computer Society.

Complementary Bibliography

[1] J. P. Fernández Álvarez, J. L. Fernández-Martínez, and C. O. Menéndez Pérez. Feasibility analysis of the use of binary genetic algorithms as importance samplers application to a geoelectrical ves inverse problem. *Mathematical Geosciences*, 40:375–408, 2008.

[2] World Medical Association. Declaration of helsinki - ethical principles for medical research involving human subjects. *The Journal Of the American Medical Association (JAMA)*, 310(20):2191–2194, 2013.

[3] Kwontaeg Choi, Kar-Ann Toh, and Hyeran Byun. A random network ensemble for face recognition. In Massimo Tistarelli and MarkS. Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 92–101. Springer Berlin Heidelberg, 2009.

[4] M.A. Dabbah, W.L. Woo, and S.S. Dlay. Secure authentication for face recognition. In *IEEE Symposium on Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007.*, pages 121 –126, april 2007.

[5] Hossein Ebrahimpour and Abbas Kouzani. Face recognition using bagging knn. *International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 17–19, 2007.

[6] J. L. Fernández-Martínez and E. García-Gonzalo. The PSO family: deduction, stochastic analysis and comparison. *Swarm Intelligence*, 3(4):245–273, dicember 2009.

[7] J. L. Fernández-Martínez and E. García-Gonzalo. Stochastic stability analysis of the linear continuous and discrete PSO models. *IEEE Transactions on Evolutionary Computation*, 15(3):405–423, June 2011.

[8] J. L. Fernández-Martínez and E. García-Gonzalo. Stochastic stability and numercal analysis of two novel algorithms of PSO family: PP-PSO and RR-PSO. *International Journal on Artificial Intelligence Tools*, 21(3):1240011, 2012.

[9] J. L. Fernández-Martínez, E. García-Gonzalo, and J. P. Fernández-Álvarez. Theoretical analysis of particle swarm trajectories through a mechanical analogy. *International Journal of Computational Intelligence Research*, 4(2):93–104, 2008.

[10] J. L. Fernández-Martínez, E. García-Gonzalo, J. P. Fernández-Álvarez, H. A. Kuzma, and C. O. Menéndez-Pérez. Pso: A powerful algorithm to solve geophysical inverse problems. application to a 1d-dc resistivity case. *Journal of Applied Geophysics*, 71(1):13–25, 2010.

[11] J. L. Fernández-Martínez, M. Z. Fernández Mu niz, and M. J. Tompkins. On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics*, 77(1):W1–W15, 2012.

[12] J. L. Fernández-Martínez, Z. Fernández-Mu niz, J. L. G. Pallero, and L M. Pedruelo-González. From bayes to tarantola: New insights to understand uncertainty in inverse problems. *Journal of Applied Geophysics*, 98:62–72, November 2013.

[13] J. L. Fernández-Martínez, J. L. G. Pallero, Z. Fernández-Mu niz, and L. M. Pedruelo-González. The effect of noise and tikhonov's regularization in inverse problems. part i: The linear case. *Journal of Applied Geophysics*, 2014.

[14] J. L. Fernández-Martínez, J. L. G. Pallero, Z. Fernández-Mu niz, and L. M. Pedruelo-González. The effect of noise and tikhonov's regularization in inverse problems. part ii: The nonlinear case. *Journal of Applied Geophysics*, 2014.

[15] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen*, 7:179–188, 1936.

[16] Mehrtash Harandi, Brian Lovell, and Javid Taheri. *Ensemble Learning for Object Recognition and Tracking*, volume 1 of *1*, pages 261–278. Springer, September 2011.

[17] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, june 2009.

[18] A.K. Jain, S. Prabhakar, L. Hong, and S. Pankanti. Filterbank-based fingerprint matching. *Image Processing, IEEE Transactions on*, 9(5):846–859, 2000.

[19] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings IEEE International Conference on Neural Networks (ICNN '95)*, volume 4, pages 1942–1948, Perth, WA, Australia, November-December 1995.

[20] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics, New York, 772, 2011.

[21] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, May 2003.

[22] V. Kurkova. Supervised learning with generalization as an inverse problem. *Logic Journal of IGPL*, 13:551–559, 2005.

[23] J. Lu, K.N. Plataniotis, AN. Venetsanopoulos, and S.Z. Li. Ensemble-based discriminant learning with boosting for face recognition. *Neural Networks, IEEE Transactions on*, 17(1):166–178, Jan 2006.

[24] A. Nait-Ali. Hidden biometrics: Towards using biosignals and biomedical images for security applications. In *7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), 2011*, pages 352 –356, may 2011.

[25] Mehran Najafi and Mansour Jamzad. An ensemble based learning for face recognition with similar classifiers. In *World Congress on Engineering and Computer Science, San Francisco, USA, 24–26 October 2007.*, 2007.

[26] A. B. Owen. Monte carlo and quasi-monte carlo for statistics. In L'Ecuyer P. and Owen A. B., editors, *Proceedings of Monte Carlo and Quasi-Monte Carlo Methods (MCQMC) 2008, Montreal Canada*, pages 3–18. Springer-Verlag, 2009.

[27] R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45, 2006.

[28] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[29] L. Rosasco, Caponnetto A., De Vito E., De Giovannini U., and Odone F. Learning, regularization and ill-posed inverse problems. In *18th Conf. on Neural Information Processing Systems, NIPS2004*, 2004.

[30] Farokhi S., Shamsuddin S., Flusser J., Sheikh U., Khansari M, and Jafari-Khouzani K. Rotation and noise invariant near-infrared face recognition by means of zernike moments and spectral regression discriminant analysis. *Journal of Electronic Imaging*, 22(1):013030–013030, 2013.

[31] Terry Windeatt. Ensemble neural classifier design for face recognition. In *ESANN*, pages 373–378, 2007.

[32] F. Zernike. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, 1(7-12):689–704, 1934.

[33] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, December 2003.