Associative Memory Model Based in ICA Approach to Human Faces Recognition

Celso Hilario, Josue-Rafael Montes, Teresa Hernández, Leonardo Barriga and Hugo Jiménez
CIDESI- Centro de Ingeniería y Desarrollo Industrial México

1. Introduction

The human-like activities have been representing a research topic in several areas, which try to understand the internal processes involved. However, the complexity and the diversity of this situation has allowed to propose approaches of different areas. These approaches pretend to imitate via simulation/emulation particular behaviors. Human perception as information of an acquiring process and brain data manipulation process approach represent two open tasks in pattern recognition analysis. Human Face Recognition (HFR) involves two approaches. The main overcome is given by a set of adequate features to characterize a human face image and multiple situations involved in the recognition process.

HFR is typically performed by the use of previous well-known features and a set of classifiers; both of them used to define a criterion for clustering and classifying each set of features. However, these features are not completely invariant on different environment conditions as well as changes of perspective, luminance conditions and shadows generation. The majority of existent approaches are limited and conditioned to specific scenario conditions, where face features are well-behaved. In this sense, there are approximations as Santini & Jain (1999); Zhang & Zhang (2010); Zhao et al. (2003) ,where the authors show several criteria, focused on the human face features and its invariance to different scenario conditions.

On the other hand, the classifiers used for clustering human face characteristics are deeply dependable on the feature behavior; i.e. over the space of features, faces with high degree of similarity are spread in cumulus; then these cumulus are feasible for clustering via any grouping criterion. In this scenario, the most representative approaches include Abdelkader et al. (2001); Aggarwal & Cai (1999); Collins et al. (2002); Duda et al. (2000), where different clustering criteria for human face recognition are shown. Note the foundations and paradigms used are different, and consequently the results obtained are distinct in similar scenarios. Different approaches are similar in the point of each one proposed a new way to discriminate information that results independent[1] one of the other. One well-accepted classifier described above are the associative memories . The associative memory approach is a connective one; which uses the linearity expressed in data set as well as a linear transformation. This approach usually is refereed as a kind of Neural Network. In several scenarios it could represent a robust

[1] Independent term is referred to the disposition of different measures that results different among them under certain well-defined operators

approach because it supports the interference of noise in the data Duda et al. (2000); Minsky & Papert (1987). Some of the most distinctive works done around associative memories include Hopfield (1982); Kosko (1998); Polyn & Kahana (2008). The majority of current Associative Memories approaches use a binary representation or some discrete scenarios, where the coding process consists on symmetric representation models. This situation might be a limit to express robust classifiers.

Finally, other approaches make emphasis in the mixture of a set of features and clustering approaches. Some of the most significant works are Ben et al. (2002); Bray (2005); Giese & Poggio (2000), where the authors explicitly show criteria to define human features and a cluster approach to group each class.

In this work, we present a different way to figure out the problem of face recognition. The proposal consists on consider the problem of HFR as an associative task. Face variations represent the data mixed with noise. Then, the proposal consists of an heterogeneous process, where each different codification face represents a class and its variations the data versions mixed with noise. The associative process expands traditional approaches, due to this it uses two ways of independence: the linear independence and the probabilistic independence. Both of them are deeply based on the superposition property of signal analysis Books (1991). Finally, the proposal has been tested with a data base of faces. This data base considers different face position, luminance conditions and face gesticulation.

2. Foundations

In this chapter, we describe main concepts, where our proposal is based. In the first stage a description of the independence data concept viewed on different areas and which characteristics are important to take care of. In second part is introduced the concept of linear independence and its main properties. Finally, in the third part, the statistical independence is introduce and its main properties.

2.1 Data independence

The concept of independence in several areas is related to the idea of analyzing when certain events or objects are exclusive; i.e. there is no affectation in the behavior of any interacting objects. Typically, the independence, according to the area of research is oriented of the theoretical foundations used to express and measure it. The independence is close related with its opposite, the dependence concept. Both definitions are completely dependable of the information representation and the way of operate it. This fact implies, the way of measure independence/dependence is strictly related of what characteristic or form to manipulate information is used. For instance, two algebraic expression may be linear dependentTomasi (2004); but it does not imply being statistical dependentDuda et al. (2000), nor grammatically dependent too. Analyzing several dependence definitions, there are common characteristics surrounding the concept. These characteristics are:

1. A well-defined domain , which is usually mapped to an order/semiorden relationship[2].
2. A representation of data structure.
3. An operator to determine the independence/dependence.

[2] It could be discrete or continuous.

4. An operator to mix independent data.

The first one is achieved with the aim to declare an explicit order in the work space. This order establishes basic foundations to other operators. The second property is focused to any data needs of a representation to figure out certain behaviors, which defines the interdependence and dependency of the data.The next property consists on an explicit operator or set of properties which define, for a particular space, which is the criteria of dependence/independence using the last two points above mentioned. Finally, the last property represents an operator as a Cartesian product like, which defines the rules of mixing two independent datum.

2.2 Linear independence

Firstly, a first kind of independence is discussed. It is linear independence. Linear algebra is a branch that studies vector spaces (also called linear spaces) along with linear maps, mappings between vector spaces that preserve the linear structure. Because vector spaces have bases, matrices can be used to represent both vectors and linear transformations, this facilitates computation and makes vector spaces more concrete. Linear algebra is commonly restricted to the case of finite-dimensional vector spaces, while the peculiarities of the infinite-dimensional case are traditionally covered in linear functional analysis.

In linear algebra, a family of vectors is linearly independent if none of them can be written as a linear combination of finitely many other vectors in the collection. A family of vectors which is not linearly independent is called linearly dependent. This is, two or more functions, equations or vectors $f_1, f_2, \ldots, f_n$ which are not linearly dependent can not be expressed in the form

$$a_1 f1_+ a_2 f_2 + \ldots + a_n f_n = 0 \tag{1}$$

with $a_1, \ldots, a_n$ constants which are not zero values.

The linear dependence of vectors is defined from basic operators in the algebra: Summation of two vectors and scalar product. Both of them in combination of basic structure elements derive in concepts (like rank, determinant, inverse, Gauss diagonalizing), used to test the dependence of two vectors.

Associative Memories encode information as matrices; which make emphasis in linear dependent methods to group classes, and linear independent to discriminate among vectors.

2.3 Probabilistic independence

Probability independence is the second independence criterion. Probabilistic theory is the branch of mathematics concerned with probability, the analysis of random phenomena. The central objects of the probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

In probability theory, to say that two events are independent intuitively means that the occurrence of one event makes it neither more nor less probable that the other occurs. Similarly, two random variables are independent if the conditional probability distribution of either given the observed value of the other is the same as if the other's value had not been observed. The concept of independence extends to dealing with collections of more than two events or random variables. Formally, this definition says: Two elements A and B are independent if only if

$$Pr(A \bigcap B) = Pr(A)Pr(B) \tag{2}$$

Here $A \bigcap B$ is the intersection of A and B, that is, it is the event that both events A and B.

This kind of independence is oriented to the data probability of occurrence in the data domain. Note the probabilistic independence analyzes the form and behavior of the probabilistic density function (pdf) of a given data; which results totally different to the linear assumption of a linear combination before described. This kind of independence is focused on match the range domain distribution. In pattern recognition it usually becomes useful, because a pdf of an event represent the data variation representation of events with small affectations.

3. The proposal

The actual work is focused in the case of Human Face Recognition (HFR), where a new way of classification and recognition based on the concept of independence is proposed. In this section is described the process of the information coding for distinctive feature identification and the associative model used to classify the different faces.

3.1 Information coding

Decision process given a set of evidence of a cluster of classes is dependable of data coding. The capabilities of clustering are deeply dependable of the information coding process and the expressiveness of information encoded. Several authors has proposed different methods for classifying the information Chaitin (2004); Shannon & Weaver (1949). This classification usually is based on a numerical criterion to define an order relationship over a descriptor space; which is conformed with the characteristic measured. Typically the clustering consists on define any distance function and the establishment of a radius-like criterion; to choose which elements belongs to a particular class Duda et al. (2000). However they are limited by the distribution of the information coding and the expressiveness of information coding.

The problem of face recognition should be viewed as a pattern recognition process; however as it was comment above, it consists on select a previously well-known descriptor, carrying the limitations described in above paragraphs. Then, generalizing, we need to define a clustering criterion without explicit descriptors. Consequently, the descriptors must be located without explicit knowledge of scenario. According to our proposes, we use a set of descriptors which contains the normalized distances of features. These features result of estimate the derivatives of order n of the distance matrix as is described as follows.

Given an image $I(\mathbf{x})$, indexed by the vector position $\mathbf{x}$, such that it contains a face. Image will be operated with gradient operator $\nabla_k^{(n)} I(\mathbf{x})$. The k parameter denotes the parameters of the derivative approach used to estimate it, and (n) is the order of the operator. Furthermore, the derivative is normalized from $m \times n$ size to $m' \times n'$, which is represented by $I'(\mathbf{x}) = \nabla_k^{(n)} I(\mathbf{x})$. The dimension $m' \times n'$ will be fixed by subsequent images to be analyzed. Using $I'(\mathbf{x})$, a distance matrix is built, representing the derivatives as a long vector of $m'n'$ dimension by linking of each row of the image derivative as follows

$$M_d(i,j) = d_k(\mathbf{I}'(i), \mathbf{I}'(j)) \tag{3}$$

for all i, j positions in $\mathbf{I}'(x)$ which is the version as vector of $I'(\mathbf{x})$; d_k is any given distance function defined and M_d is a square matrix of $m'n' \times m'n'$. Matrix M_d is used as a set of descriptors of each face.

Gradient operator $\nabla_k^{(n)}$ provides information of pixel's intensities variations, which they indicates pixel the degree of texture and border information in the images. Note this operator results invariant at diffuse light sources. Additionally, matrix distance M_d is dependable of d_k distance function based on L_k norm; i.e. values of k less than 1 increases the sparseness the data on M_d, and values greater than 1 for k decreases the sparseness of data.

3.2 Associative memory

In this section we describe the proposal of a new kind of associative memory based in linear and statistical independence.

3.2.1 The principles

Associative models consist on build a model $\mathcal{M}$ such that for a pair of set $\mathcal{A}$ and $\mathcal{B}$ create a relationship $\mathcal{R} : \mathcal{A} \to \mathcal{B}$. Being strictly the $\mathcal{R}$ relation has a property where a pair of elements $(a, b) \in \mathcal{A} \times \mathcal{B}$, and the elements with almost a distance criteria d_a in $\mathcal{A}$ and d_b in $\mathcal{B}$ are related too; i.e. elements with a small similarity with a pair $(a, b) \in \mathcal{A} \times \times \mathcal{B}$ are related in the same way. Typically the memories are classified according to the nature of associated sets: when $\mathcal{A} = \mathcal{B}$ memory is named as auto associative; when $\mathcal{A} \neq \mathcal{B}$ is named as hetero associative; when $|\mathcal{A}| > |\mathcal{B}|$ is considered as a classifier and finally, when $|\mathcal{A}| \leq |\mathcal{B}|$ is considered as a transducer Knuth (1998) or codifier Shannon & Weaver (1949).

Model $\mathcal{M}$ is build usually with a few samples (commonly named as learning or training samples). Being strictly, there is no particular expressions which decide over related elements; instead, the process is well-based in theoretical foundations. $\mathcal{M}$ is build attending the main theoretical foundations which define the class of the memory used. The nature of the majority models $\mathcal{M}$ are based in connective approaches, and consequently, it express a lineal mixture among inputs and outputs. The quality of learning process depends on the capabilities of associate with fewer errors the training samples and it will be used as good estimators for non-considered pair of elements in the relationship Minsky & Papert (1987); Trucco & Verri (1998).

However, even the learning process results robust, there are situations where the linearity expressed by the inputs and outputs results insufficient for establish a relationship between them. In sections described above we speak about the independence concept. Then, define a theoretical framework which uses several independence criteria should be beneficial to develop better models $\mathcal{M}$ and consequently better associative models.

The aim contribution of this work consists on a new model of Associative Memory based on real domain and the mixing of two different approaches of independence: the lineal independence and statistic independence. The proposal works under the assumption that two signal can be mixed/unmixed if we know the structure of the distribution of each signal. Our approach works with large vectors, such that the distribution of the data inputs that conform the event encoded can be estimated.

An associative Memory have at least two operators: similarity operator, which indicates the radio of match with some class previously learned; and belonging operator, which verifies the data previously encoded in the memory. Additionally, an scheme to learn and estimate each class is needed. In further paragraphs the proposal is showed, which one mixes two kind of independences: the linear and statistic. Both of them are used to define a framework which associates inputs and outputs.

Given a set of information sources $\{S_1, \dots, S_n\}$, the data contained in all signals should be mixed. This fact becomes true whenever sensing process is related and should be affected by the same external variables. A first consideration is, the *true* variables of the system are not perceived directly; but we can assume that they are the result of any combination of the measured signals. For simplicity, this combination is viewed as a linear combination. Then, from the sources $\{S_1, \dots, S_n\}$, we can estimate each *true* variable as any linear combination as follows

$$U_i = \sum_{j=1}^{n} w_{ij} S_j \tag{4}$$

Consequently, to obtain simplicity the expression above can be rewriting as the dot product of $\mathbf{w}_i$ vector with a vector $\mathbf{S} = [S_1, S_2, \dots, S_n]$, as $U_i = \mathbf{w}_i \mathbf{S}^T$. After, as it is appreciated, in n sources, there are a maximum of U_n variables, it leads to $\mathbf{U} = W_{n \times n} \mathbf{S}^T$, for a particular time stamp t.

Then, unmixing and mapping this sources to well-behaved space becomes an overcome that could be estimated as Independent Component Analysis problem; i.e. the real source measurement must be considered as linear and statistically independent among each component. Under these assumptions, one way to estimate the independent variables could be done with and ICA approach. The approach consists on estimate a W matrix such that mix/unmix the sources to orthogonal variables as follows

$$U = W_{n \times n} X_{n \times m} \tag{5}$$

where $X = [S_1, S_2, \dots, S_n]^T$ is a matrix composed with all information sources; W a square matrix with the form of $W = [\mathbf{w_1}, \dots, \mathbf{w_n}]$ for unmixing the sources in X and U represents a set of linear and statistical independent variables. The values of W are estimated iteratively via fast-ICA algorithm as it is showed in Table 1. Fast-ICA algorithm consists on detect the orthogonal protections which maximize the information. The last one is measured via neg-entropy, as an approach to the real measurement result from the calculus of system entropy. As it can be appreciated the algorithm is non-deterministic starting from random values for each one $\mathbf{w}_i$ projection; which each iteration is tunning in direction of maximum information.

Unfortunately, one of the greatest disadvantages consists on the transformation W separates the mixed data, but the output are not sorted. This cause, the use of any component sort of returned by ICA is totally dependable of the phenomenon nature. However, for our purposes

```
Pseudo-Code
Estimate for each component $\mathbf{w}_i$ in $W$ as follows
1. Initialize $\mathbf{w}_i$ with random numbers.
2. Let $\mathbf{w}_i^+ \leftarrow E\{\mathbf{x}g(\mathbf{w}_i^T\mathbf{x})\} - E\{\mathbf{x}g'(\mathbf{w}_i^T\mathbf{x})\}\mathbf{w}$.
3. Let $\mathbf{w}_i \leftarrow \frac{\mathbf{w}_i^+}{||\mathbf{w}_i^+||}$
4. If the convergence is not achieved, go back to 2.
```

Table 1. Pseudo-code of *W* estimation

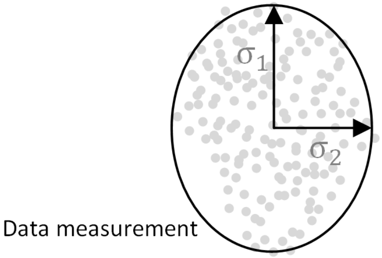


Fig. 1. Data Measurement Sparseness; as it is appreciated the eigen-values are located in the principal orthogonal axis.

this property could be used for developing a similarity criterion as is described in further paragraphs.

3.2.2 Similarity criterion

The similarity criterion for the proposal is based on the information given in the matrix W. The orthogonal components expressed in W can not be sorted, but it is possible to weight the contribution of each orthogonal component.

The weight process is done by analyzing the eigen-values of the matrix W. The relative magnitude of each eingen-value is used as a contribution of each orthogonal axis. This is, factorizing W as $U\Sigma V^T$, where Σ is an square matrix and the diagonal vector $[\lambda_1, \ldots, \lambda_n]$ has the singular values of W. The eigen-values Σ of W are related with the rank and the sparseness of each orthogonal component in the source space (see Figure 1). The most simple case of similarity using ICA is defined for two different signals as the proportion between each one of different eigen-values of the unmixing matrix W; i.e. formally is defined as

$$d(s_1, s_2) = \frac{\lambda_2}{\lambda_1} \tag{6}$$

where $W = U \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} V^T$, and W is estimated with the algorithm described in Table 1. This similarity criterion is given for a pair of signals and can be interpreted as a degree of orthogonality existent in two components lineal and statistical independent; i.e. for a pair of signals s_1 and s_2 both of them are *similar* if and only if both s_1 and s_2 are linear and statistical dependent. The degree of independence is measured with the proportion of the second and the first eigen-value. If the proportion is near to 0 means, means s_1 and s_2 are linear and statistical dependent; i.e. they are similar. The proportion $\frac{\lambda_2}{\lambda_1}$ provides a normalized measure, and λ_1 is a normalized factor. In other scenarios we can use directly λ_2 as the degree of similarity between s_1 and s_2. Note, for a not normalized distance $d(s_1, d_2) = \lambda_2$, the expression is a metric Santini & Jain (1999).

Next,the belonging operator must be defined using a similarity function (see Equation 6), as a function $\epsilon_\lambda : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \{true, false\}$, and its equation is defined as

$$\epsilon_\lambda(s_1, s_2) = d(s_1, s_2) \geq \lambda \tag{7}$$

where s_1 and s_2 represent the data information of encoded event. This operator is performed with the aim to decide when a given reference data s_1 and testing data s_2, correspond to the same information encoded.

3.2.3 Memory architecture

In last section the similarity criterion and the belonging operator has been defined. In this section we describe the architecture of an Associative Model based on last definition. The focus consists on use a set of representative data vectors, which represent the different classes of the memory. These vectors are matched with the data for a classification process using the belonging operator. Learning and discovering process are discussed in follows section. After this, we make the assumption that the representative classes are known.

The architecture of the Associative Memory proposed consists on given k classes denoted by $\Psi_1, \Psi_2, \ldots, \Psi_n$. Each class Ψ_i is denoted by a set of signals with the form $\Psi_i = \{s_1, s_2, \ldots, s_k\}$. Each signal s_i represents the events coding as a large vector fixed vectors. The process of encoding must warranty the best descriptors are encoded. For HFR purposes, it is assumed, the coding process described above has been used. We point out, the quality and properties of encoding process will affect the association capabilities. To get adequate results we emphasize that the information representation follows the properties of linear and statistical independence in the created relationships in the memory. The most representative information, of each class, is used as reference to match for discovering which class corresponds. For each Ψ_i there is a representative element denoted by the operator $E_i(\Psi_i)$. This operator is defined in general terms as the expected information of class Ψ_i. The nature of this operator is stochastic, but it might change, according to the data nature and distribution of coding data.

Then, the Associative Memory is defined by a matrix $\Xi = \begin{bmatrix} E(\Psi_1) \\ \vdots \\ E(\Psi_k) \end{bmatrix}$, such that it is conformed by all expected elements of each class Ψ_i. The matrix Ξ represent the Associative Memory.

Next, the memory needs an operator to verify when it contains a particular data. The operator of belongings aforementioned, is extended to verify if a given data s_i expressed as vector, is contained in any class of a memory Ξ. The operator is extended in two ways: the first variation return which elements has a certain degree of similarity with the testing data s_i. This operator will be useful when someone need to know what classes have closed similarity to the data s_i.

The operator is expressed as

$$\begin{aligned} \in_{\Lambda}(\Xi, s_j) &= \{\epsilon_{\Lambda}(E(\Psi_i), s_j) = true\} \\ \text{for } j &= 1, 2, \ldots, k. \end{aligned} \tag{8}$$

Note that it return a set with the most similar classes. Second variation return the most similar class to our testing data s_j. Then a variation of last operator consists on returning the class which has the minimum distance to the representative information of the class. Finally, the operator is defined as

$$\in^{*}_{\Lambda}(\Xi, s_j) = \min\{d_i(E(\Psi_i), s_j)\} \tag{9}$$

$$\text{for all } E(\Psi_i) \in \epsilon_{\Lambda}(E(\Psi_i), s_j). \tag{10}$$

Both operators are used to determine when a given data s_j belongs to any learned class. As final comments, the value of Λ is a threshold of belongings, and it is a parameter dependable of the scenario.

3.2.4 Learning process

Learning process, in an Associative Memories, consists on discover the relationships between $\mathcal{A}$ and $\mathcal{B}$ sets, and it usually is viewed as a class discovering process under a metric criterion defined by the belonging operator (see Equation 7).

The class discovering process, in the associative memory, is achieved with a learning process that needs a belongings threshold Λ. Note that the signal used has been normalized in frequency and length. Both assumptions help to operate them. For general purposes, it has been considered only the automatic learning process, which results useful in several scenarios. For a given set of signals $\{S_1, \ldots, S_n\}$, the process of discovering the classes is performed as follows.

The simplest case is whenever the process involves only one class. At the beginning, class Ψ_1 is an empty set. First signal S_1 is added to the first class $\Psi_1 \leftarrow \Psi_1 \bigcup \{S_1\}$ and consequently, $\Xi = [\Psi_1]^T$. For further signals S_i, they are added to set Ψ_1 if and only if $\in^{*}_{\Lambda} (\Xi, S_i) = 1$.

Generalizing for k classes, whenever $\in^{*}_{\Lambda} (\Xi, S_i) = 0$, a new class Ψ_k is created, such that $\Xi \leftarrow [\Xi, \Psi_k]^T$, where Λ is a distance threshold among classes. The number of classes Ψ_k added represents the different orthogonal concepts in the memory. Note that, if we had previously information of k classes, it became as a supervised learning process. Analogous case, if we had not previously information, it became as dynamic learning process. Additional to mention, if learning process is always computed, the expected value $E(\Psi_i)$ for each class should be changed, adapting dynamically to new data evidence.

In this point we need to define the $E(\Psi_i)$ operator. In several cases the expected operator is defined as the average; which is computed as the average of each component of all elements in Ψ_i. Note that we make the assumption of Ψ_i elements are sparse over the feature space uniformly. However, when the data distribution does not follows an uniform distribution, it will be insufficient. In this sense $E(\Psi_i)$ becomes as the expected value. The computation can be easily computed and estimated for each component. In this situation, the expected value is computed for each vector position. Note for the estimation of expected value, we need to have enough evidence to approximate component distribution and estimate global maximum. The expected value needs that the learning data must be sparse uniformly in the feature space.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**Total**
# Pictures	24	11	23	10	25	26	18	15	30	10	24	40	14	32	302

Table 2. Number of postures per sample of data base used.

The Two last approaches discussed above would result useful; however data in Ψ could result affected by noise or might not being sampled uniformly. In this cases is needed a method to dismiss this variations. One approach consists on eliminate the least significant orthogonal components in Ψ_i. Then, the filter process would be performed via PCA (Principal Component Analysis) approach. PCA approach consists on reconstruct the original information only taking most significant linear components. The advantage of eliminate small components is that eliminate several redundancies making more compact and better sparseness the learning evidence in the feature space.

For Ψ_i signal set is constructed a matrix D with each data vector transposed as follows

$$D = [S_1^T S_2^T \dots S_n^T] \text{ for } S_1, S_2, \dots, S_n \text{ signals in } \Psi_i \tag{11}$$

For dismissing the noise effects and sparse better the feature space a matrix D^* is build factorizing D as $U\Sigma V^T$ and build a matrix D^* without least significant data. D^* is reconstructed with as

$$D^* = U\Sigma^* V \tag{12}$$

where Σ^* is equal to Σ, but with the difference that the lasted singular values $\sigma_l^*, \sigma_{l+1}^* \dots \sigma_n^*$ are zero values and l value is estimated considering a percentage of original information. Proportion between summation of l principal component and summation of all components represents noise/signal ratio. Then election of l must define a percentage function of total information used in data. The percentage of information represented with l singular values is computed with $\% = \frac{\sum_{i=1}^{l} \sigma_i}{\sum_{j=1}^{n} \sigma_j}$. The election of l value, must be defined covering at least α percentage of information as follows

$$I(\alpha) = \max\arg\{l | 1, 2, \dots, n\} \tag{13}$$
$$\text{such that } \frac{\sum_{i=1}^{l} \sigma_i}{\sum_{j=1}^{n} \sigma_j} \leq \alpha$$

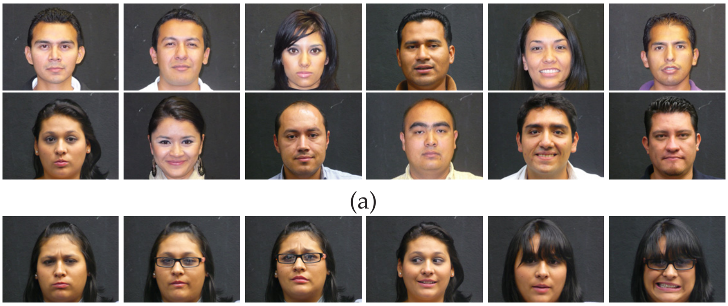
Finally $D^* = [(S_1^*)^T (S_2^*)^T \dots (S_n^*)^T]$ represents filtered data and it will be applied any scheme to estimate the calculus of expected value for Ψ_i.

4. Experimental results

In this section, we describe an experimental method for validating the proposal. The validation process consists on developing an Associative Memory to classify and to recognize Human Faces. The implementation details are given and described in follows sections.

4.1 Experimental model

Our proposal was tested with the development of Human Face recognition. The information consists on a data base of Human Faces. The data base includes different faces and different faces gesticulations. Figure 2 (a) shows samples of the data base. Each face in data base has several poses and gesticulation. The variation of each face is important, because it is used in



(a)

(b)

Fig. 2. Face data base: (a) some different samples contained in the data base; where (b) for each sample, there is different face gesticulation.

the learning process, for extraction of the main features which characterize each face. Pictures were taken with a Panasonic Lumix Camera at 7.1 Megapixels (3072×2304) in RAW format. The number of photos and persons involved are appreciated in Table 2.

In Section 3.1, a scheme for encoding was presented. This process is applied to the different faces in the data base. For practical purposes, derivative of order 1 has been used for encoding and characterizing each face. Derivative has been implemented with a symmetrical mask $d = [-10 + 1]$, which has assumed Gaussian. The parameters of Gaussian were fixed to $\mu = 0$ and $\sigma = 1$ with length 9, which define a Gaussian from -4 to 4 in normalized dimensions. Finally, the derivative filter was defined as $F = [-G(0,1)0 + G(0,1)]_{9\times 19}$. The Gaussian results beneficial due to, it allows dismiss noise effects in images.

Next, features descriptors were estimated with Equation (3). This equation needs a normalized version of ∇I with $m' \times n'$ dimensions. The amount of information encoded in a pattern is directly affected by dimension of normalized image. A first view, a fixed value is used to compute these patterns (32 × 32 pixels). In a second stage, dimensions of normalized version has changed in follows dimensions: $8, 16, 24, 32, 40, 48, 56, 64, 96, 128, 160, 192$ and 224. Then a matrix M of descriptors has been created as comment in Section 3.2.4. Image patterns represent the relationships among all pixels in the image border.

Process learning uses a set of patterns, as input. This process computes distinctive face patterns. In Section 3.2.4 were described two approaches for estimating these patterns. In our implementation, we only test with a single average. This consideration might be strictly simple; but for our approach the average results enough for implementations purposes. This is, superposition principle, states that, for all linear systems, the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually. So that if input A produces response X and input B produces response Y then input $(A + B)$ produces response $(X + Y)$; which for a simple average is the same, factorizing by $\frac{1}{n}$ each term involved in sum operator Books (1991).

To illustrate this process, in Figure 3 are shown some faces patterns. This patterns were computed with normalized images of 16×16. As is appreciated, vertical and horizontal lines provided information of distribution and behavior data for several faces. Note, image descriptor has a resolution of $16^2 \times 16^2$ resulted of distance image.

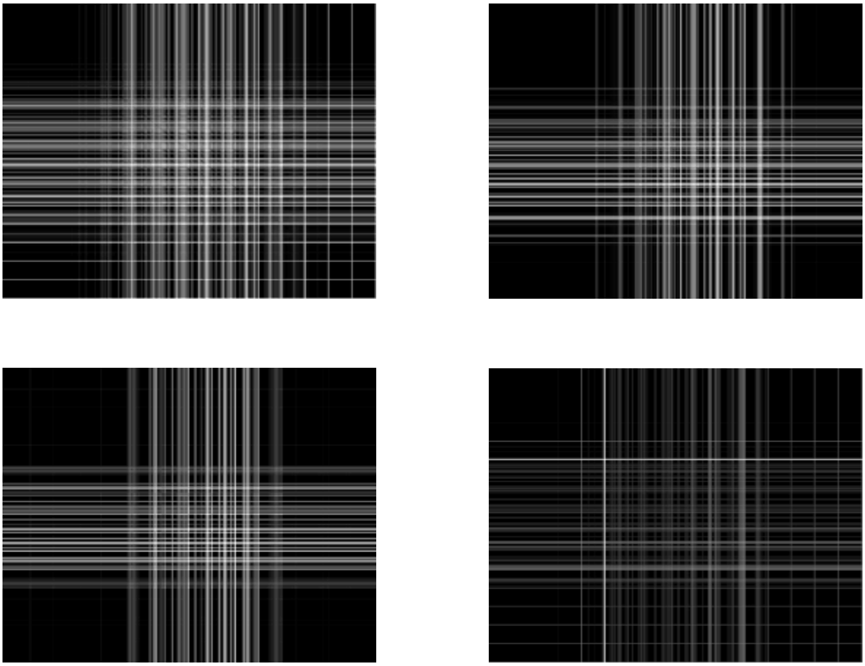


Fig. 3. Samples of patterns estimated in learning process. Visual patterns represent inner relations among features borders of each face.

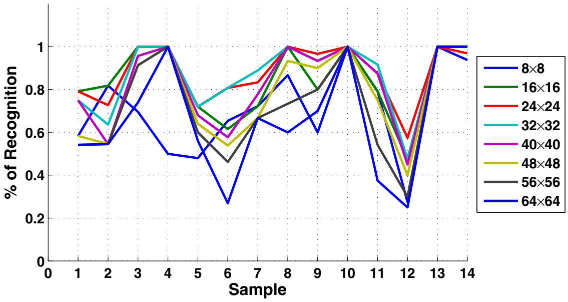


Fig. 4. Efficiency of recognition changing dimensions of normalized image.

4.2 Results and discussion

After computing a face pattern from each different sample, a memory is created as follows;

$$\mathcal{M} = \begin{bmatrix} E(\Psi_1) \\ \vdots \\ E(\Psi_n) \end{bmatrix}$$

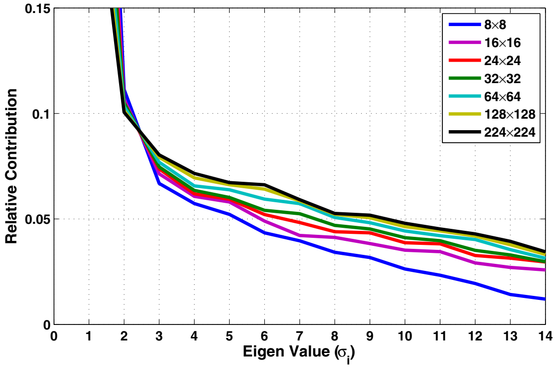


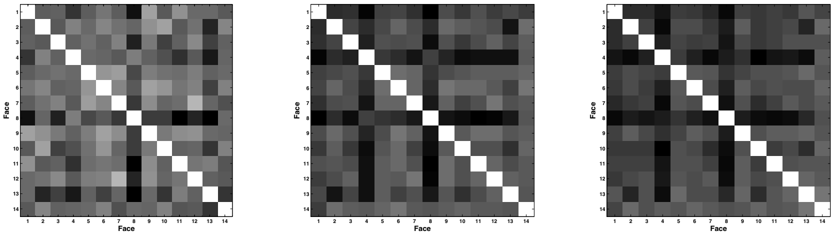
Fig. 5. Samples of patterns estimated in learning process. Visual patterns represent inner relations among features borders of each face.

where $E(\Psi_1)$ corresponds to a distinctive pattern of Ψ. Hence, an associative memory is created. For putting the memory on line, it only need to verify an encoded face and infer, which class encoded in $\mathcal{M}$ has less similarity and estimate its class. Note, an associative model is developed with the aim to infer which class has less similarity, so one must be careful because the class with less similarity additionally, need to follow a threshold similarity criterion too. This is, even there is no exist a similar class, memory return most similar.

Then, the validation of memory is performed measuring the degree of accuracy of recognition any face, to corresponding class. To perform it, all faces $s_1, s_2, \ldots, s_n$ in data base are tested. Normalized image dimension has changed. Results are shown in Figure 4.

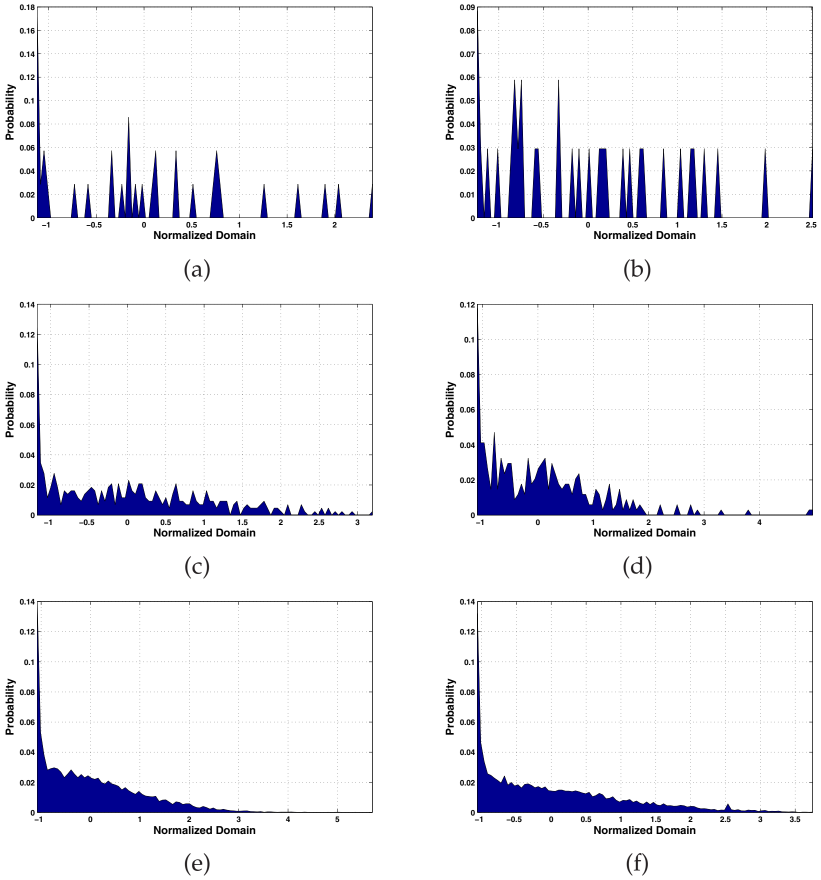
The level of recognition of our approach is over 85% with pattern sizes of 24×24 and 32×32. Note, for small sizes (8 and 16) of image patterns, and for considerable high dimensions, the memory miss classify (64 or more). This is, in the first case there is not enough information for choosing which class is more similar. To justify this fact, a matrix of similarity has been created with patterns in memory. The matrix has 14×14 dimensions and represent in pseudo color the similarity degree. White color represents high degree of similarity and black color represents no similarity. In fact, the face patterns are analyzed; hence, we expect the matrix similarity has only similitudes in the diagonal. Then, Figure 6 (a) shows that degree of similarity among classes is small and it might cause miss classification problems. But when dimensions become higher, the degree of independence usually increases too, as is appreciated in Figure 6 (b) and Figure 6 (c), for 32×32 and 224×224 respectively. Additionally this can be verified with Figure 5, where principal components of patterns are computed and its relative contribution. Note, in higher sizes, they become more linear.

Intuitively, someone expect to be more accurate with high dimensions. However, note, the face recognition approach is based on a distance matrix, and associative memory approach is founded in two kind of independences: linear and statistical. Whenever data become higher, distribution of face patterns becomes similar, being not possible to classify. This point is illustrated in Figure 7, where two pdf's of different classes are estimated with 8×8, 32×32 and 224×224 pattern size. Note all of them, are dissimilar at the beginning and become similar when dimension of pattern is increased. This is, this approach is suitable for classes well differentiated.



(a) (b) (c)

Fig. 6. Similarity matrix among classes for (a) normalized image of 8 × 8; 32 × 32 and 224 × 224.



(a) (b)

(c) (d)

(e) (f)

Fig. 7. Probabilistic density functions of two image patterns (a),(c),(e) and (b),(d),(f) with dimension 8 × 8; 32 × 32; and 224 × 224.

At this point, the miss classification effects when data become increased might be considered as inconvenient; however, it is not true, because framework is operated under its basic assumptions, and it shows the limits of encoding process; i.e. we need to define a better scheme to extract the most suitable features for face recognition which warrant linear independence and probabilistic independence. Then, to define an optimums classifier we need to consider the amount of information and its distribution associated to these data; however this point escape of the scope of this paper.

Summarizing, through this work, we propose a scheme for face recognition based on simple texture features, which uses a new paradigm of associative machines based on ICA approach. Its main characteristics include the use of real domain for characterizing data, including two kinds of independence.

5. Conclusion

In this chapter, we discussed about the independence concept, as an approach to characterize information by different features properties. Additionally, we pointed out the importance of different features when they are used to classify or discriminate certain events encoded. This fact results important to the well-development of classifier and coding process. Typically, common approaches are based on linear independence or statistical independence. Both approaches are distinct and measure different structure over data.

In this chapter, we proposed a new approach for classifying which take the advantage of both kind of linearity. It becomes relevant because we offer a new approach that characterize more strictly the information features used for data identification and data grouping. In this order, this work proposes a new family of associative memories, based on the ICA approach. The main characteristics include a real domain in the data, tolerance to certain data variations, and the association is given by the possibility of express each class as an orthogonal component and probabilistic independent. Preliminary tests show the viability of use this approach as general classifier being applicable in several research areas.

To test and validate the proposal, we implement a face recognizer application. It uses a basic coding data based on the derivative of the image. This encoding analyses differences between face border and its texture. In our tested scenario, the proposal is capable to discover and define a correct correspondence between the face variations and its corresponding face. In tested scenarios and information used the conditions were varied with the aim to show the robustness of the approach.

6. Future directions

As comment above paragraphs, a new framework based on linear independence and statistical independence is presented. Its main contributions are focused to the development of new kind of classifiers. This work exhibits the basis for new classifiers and recognizers. Further researches are focused to the application of this framework on different areas as signal analysis, image analysis and areas where the amount of information and lack of predefined features might difficult its analysis. This involves tasks as information coding, which extract good features and make more feasible the data manipulation.

Parallel, other areas of interest are related with the characterization of accuracy and efficiency of this approach, and the theoretical foundations that express new operators based on this proposal to manage in a better way the data.

Finally, one last direction is focused on the analysis of the feasibility to implement this approach in hardware, making more efficient time process and its application in real time systems.

7. References

Abdelkader, C., Cutler, R., Nanda, H. & Davis, L. (2001). EigenGait: Motion-Based Recognition of People using Image Self-Similarity, *Audio- and Video-Based Biometric Person Authentication* pp. 284–290.

Aggarwal, J. K. & Cai, Q. (1999). Human Motion Analysis: A Review, *IEEE Journal on Computer Vision and Image Understanding* 73(3): 428–440.
URL: *citeseer.ist.psu.edu/aggarwal99human.html*

Ben, J., Wang, Z., Pandit, P. & Rajaram, S. (2002). Human Activity Recognition using Multidimensional Indexing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(8): 1091–1104.

Books, P. (1991). *The Penguin Dictionary of Physics*, Penguin Books, Valerie Illingworth, London.

Bray, J. (2005). Markerless Based Human Motion Capture: A Survey, *Technical report*, Brunel University, Department System Engineering.

Chaitin, G. (2004). *Algorithmic Information Theory*, Cambridge University Press. IBM Research Center.

Collins, R. T., Gross, R. & Shi, J. (2002). Silhouette-Based Human Identification from Body Shape and Gait, *IEEE International Conference on Automatic Face and Gesture Recognition* pp. 351–356.

Duda, R. O., Hart, P. E. & Stork, D. G. (2000). *Pattern Classification*, 2nd. edition edn, Wiley-Interscience.

Giese, M. & Poggio, T. (2000). Morphable Models for the Analysis and Synthesis of Complex Motion Patterns, *International Journal of Computer Vision* 38(1): 59–73.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences of the USA* 79(8): 2554–2558.

Knuth, D. (1998). *Art of Computer Programming*, Vol. 1, 2 edn, Addison-Wesley Professional.

Kosko, B. (1998). Bidirectional associative memories, *IEEE Transactions on Systems, Man, and Cybernetics* 8(11): 40–46.

Minsky, M. L. & Papert, S. A. (1987). *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*, second edition edn, The MIT Press; Expanded edition.

Polyn, S. & Kahana, M. (2008). Memory search and the neural representation of context., *Trends in Cognitive Sciences* 12: 24–30.

Santini, S. & Jain, R. (1999). Similarity Measures, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(9): 871.

Shannon, C. & Weaver, W. (1949). *The Mathematical Theory of Communication*, University of Illinois Press.

Tomasi, C. (2004). *Mathematical Modelling of Continuos Systems*, Duke University.

Trucco, E. & Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*, 1 edn, Prentice Hall.

Zhang, C. & Zhang, Z. (2010). A survey of recent advances in face detection, *Technical Report MSR-TR-2010-66*, Microsoft Research Microsoft Corporation.

Zhao, W., Chellappa, R., Phillips, P. & Rosenfeld, A. (2003). Face recognition: A literature survey, *ACM Computing Surveys* 35(4): 399–458.