AUTHOR QUERY FORM

ELSEVIER	**Journal: ESWA** **Article Number: 8484**	**Please e-mail or fax your responses and any corrections to:** **E-mail: corrections.esch@elsevier.sps.co.in** **Fax: +31 2048 52799**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Click on the 'Q' link to go to the location in the proof.

Location in article	Query / Remark: click on the Q link to go Please insert your reply or correction at the corresponding line in the proof
Q1	Please confirm that given names and surnames have been identified correctly.
Q2	Please note that the reference style has been changed from a Numbered style to APA style as per the journal specifications.
Q3	Please check the DOI number.
Q4	Please provide the significance for the bold characters provided in the Table 1.

Please check this box if you have no corrections to make to the PDF file ☐

Thank you for your assistance.

Expert Systems with Applications xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Feature selection for face recognition based on multi-objective evolutionary wrappers

Q1 Leandro D. Vignolo [a,*], Diego H. Milone [a], Jacob Scharcanski [b]

[a] Grupo de Investigación en Señales e Inteligencia Computacional, Departamento de Informática, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, CONICET, Argentina
[b] Instituto de Informatica and Dept. de Engenharia Eletrica, Universidade Federal do Rio Grande do Sul, Caixa Postal 15064, 91501 970 Porto Alegre, RS, Brazil

ARTICLE INFO

Keywords:
Wrappers
Multi-objective genetic algorithms
Feature selection
Face recognition

ABSTRACT

Feature selection is a key issue in pattern recognition, specially when prior knowledge of the most discriminant features is not available. Moreover, in order to perform the classification task with reduced complexity and acceptable performance, usually features that are irrelevant, redundant, or noisy are excluded from the problem representation. This work presents a multi-objective wrapper, based on genetic algorithms, to select the most relevant set of features for face recognition tasks. The proposed strategy explores the space of multiple feasible selections in order to minimize the cardinality of the feature subset, and at the same time to maximize its discriminative capacity. Experimental results show that, in comparison with other state-of-the-art approaches, the proposed approach allows to improve the classification performance, while reducing the representation dimensionality.

1. Introduction

Face recognition has received significant attention due to its promising applications in security systems and human–computer interaction, which has motivated important new developments in research areas such as image processing and artificial intelligence. In general, the methodologies are developed for face images acquired under controlled conditions, but in practical situations, face recognition systems usually must also deal with changing conditions like variations in pose, expression and illumination, which introduce intra-class variability in the extracted features with respect to the training data (Li and Jain, 2011; Milborrow and Nicolls, 2008; Wen, 2012). In a face recognition problem, a given face image is classified into K previously known face classes. This is usually done using a model trained with the feature vectors extracted from a database of face images (Cevikalp and Triggs, Q2 2010; Oh et al., 2013).

Two main approaches exist in face recognition, those which are based on holistic methods and the others based on analytic techniques (Kong et al., 2005). Holistic methods, such as eigenfaces (Turk and Pentland, 1991), use global characteristics of the face images. On the other hand, analytic techniques, like the Active Shape Models (ASM) (Cootes et al., 1995; Wang et al., 2013), extract face features related to the eyes, the nose, the mouth, etc.

In facial modeling with ASM, a number of points (i.e. image locations) are selected from an input image, but only some of these points are useful for characterizing the face, since the others have small contributions to discrimination, or are noisy. As the training of ASM converges towards salient edges, if these edges are distorted by noise or some other artifact, like local illumination variation, erroneous feature matchings might arise (Behaine and Scharcanski, 2012). Despite recent improvements made to ASM techniques, the matching errors may be undesirably high at some face locations (Hill et al., 1996; Kim et al., 2007). Even after some new implementations that improve the landmark location accuracy, the detection of facial features with varying pose and illumination is still challenging (Milborrow and Nicolls, 2008; Zheng et al., 2008). Usually, once a set of face image locations (i.e. points) is selected by the ASM method, a number of features describing each face location is extracted. Then, the resulting feature vectors representing the faces are usually of high dimensionality, which makes the classification task more difficult (Bishop, 2007). Also, large feature sets are prone to overfitting and, hence, to achieve poor generalization performance (Handl and Knowles, 2006).

In Behaine and Scharcanski (2012), the authors proposed to improve the ASM performance in face recognition by weighting the facial features according to a method based on adjusted mutual information. As the authors shown, this criterion allowed the selection of the most relevant landmark points, in order to improve the face classification results. However, the flexibility provided by the full set of features obtained by the ASM approach has not yet been fully explored by means of feature selection techniques, in order to

* Corresponding author. Tel.: +54 342 4575233x191; fax: +54 342 4575224.
E-mail addresses: ldvignolo@fich.unl.edu.ar (L.D. Vignolo), d.milone@ieee.org (D.H. Milone), jacobs@inf.ufrgs.br (J. Scharcanski).

0957-4174/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.eswa.2013.03.032

reduce the dimensionality of the representation while improving the face classification results. On the other hand, significant progresses have been made with the application of different artificial intelligence techniques for feature selection. In particular, many works rely on evolutionary algorithms for feature subset optimization (Chatterjee and Bhattacherjee, 2011; Hsu et al., 2011; Li et al., 2010; Pedrycz and Ahmad, 2012), and for the search of optimal representations (Charbuillet et al., 2009; Vignolo et al., 2011a,b; Vignolo et al., 2013). In Vignolo et al. (2012) a genetic wrapper was proposed for the selection of the most relevant features for improving the accuracy of face recognition. Nevertheless, this wrapper was focused on classification accuracy improvement, which limits the proposed method since it overlooks other important issues in face classification (e.g. feature space dimensionality and class overlap).

In order to guide the search within the space of feasible face classification solutions, here we propose the use of a Multi-Objective Genetic Algorithm (Coello Coello et al., 2007). This method allows to overcome the above mentioned limitations by maximizing the face classification accuracy, while minimizing the number of features and the mutual information. Two different strategies for the representation of the candidate solutions are proposed and compared, and the generalization performance of the feature subset selection is assessed using an independent data set.

The organization of this paper is as follows. First a brief introduction to the use of ASM for face modeling is given in Section 2, and next our multi-objective wrapper for the selection of features for face classification is presented in Section 3. Section 4 describes our experiments and discuss the results obtained for face classification. Finally, our conclusions and ideas for future work are presented in Section 5.

2. Active shape models for facial recognition applications

The ASM approach is used to represent shapes and their expected ways of deforming as learned from a training set. For this, it uses flexible point distribution models (PDM), based on the positioning of selected points in the face image examples (Hill et al., 1996). This PDM iteratively deforms to fit the shape of an object, constrained to vary in the way learned from a set of training examples. When applied to face recognition, the ASM is trained on a set of sample faces, and N points are used to represent the shape of each face within its class (see Fig. 1(a)).

Nevertheless, matching errors may arise in the location of the PDM points, often called *landmarks*, in a face image (see Fig. 1(b)) (Behaine and Scharcanski, 2012). Then, considering a training image set with K face classes, each class $k = 1, \ldots, K$ is represented by N landmark points $S_{k,\epsilon} = \{p_i(x_i + \epsilon_{x_i}, y_i + \epsilon_{y_i})^k\}$, where $i = 1, \ldots, N$, (x_i, y_i) are the coordinates of the landmark point p_i and $(\epsilon_{x_i}, \epsilon_{y_i})$ are the respective location errors. Every relevant facial characteristic (e.g. eye centers, mouth contours, etc.) is represented by a set of landmarks p_i, and the particularities of each point in the image are described by Q features (e.g. chrominance, texture, etc.). The features at landmark p_i will be denoted $\{F_{j,i}\}$, with $j = 1, \ldots, Q$.

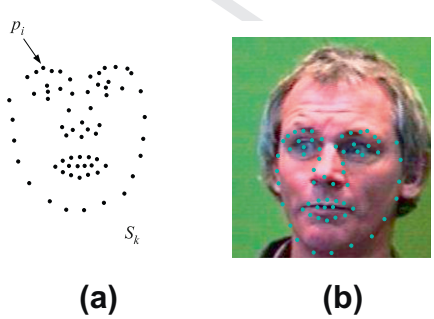

Fig. 1. Illustration of the landmark points used to model a face (a) and their location on an image (b) (Behaine & Scharcanski, 2012).

In order to describe each one of the N landmark points p_i, the mean $\mu_{F_{j,i}}$ and the variance $\sigma^2_{F_{j,i}}$ of the measurements of each feature j taken within a defined neighborhood of that point are commonly used (Behaine and Scharcanski, 2012). These are computed for all features $F^m_{j,i}$, with $m = 1, \ldots, M$, where M is the number of training image samples,

$$\mu_{F_{j,i}} = \frac{1}{w^2}\sum_{r=1}^{w}\sum_{q=1}^{w}\mu_{j,i}(r,q), \tag{1}$$

$$\sigma^2_{F_{j,i}} = \max_{r,q\in W}\left\{\sigma^2_{j,i}(r,q)\right\}, \tag{2}$$

where (r,q) are the pixel coordinates within the window W (of size $w \times w$), centered at the landmark point p_i (Behaine and Scharcanski, 2012), $\mu_{j,i}(r,q) = \frac{1}{M}\sum_{m=1}^{M}F^m_{j,i}(r,q)$ and $\sigma^2_{j,i}(r,q) = \frac{1}{M}\sum_{m=1}^{M}\left(F^m_{j,i}(r,q) - \mu_{j,i}(r,q)\right)^2$.

To consider the feature variability within the $w \times w$ neighborhood of landmark p_i, the maximum window variance was used in (2). The window size was set to $w = 2\max\{\sigma_\epsilon\}$, where σ_ϵ is the standard deviation of landmark location errors, measured during ASM training. The probability density of location errors at each landmark point is assumed to be approximately Gaussian (Shi et al., 2006).

In this work, the face detector proposed by Demirel and Anbarjafari (2009) is used, and the process applied to the database of face images in order to obtain the ASM-based set of features is described in detail in Behaine and Scharcanski (2012), Vignolo et al. (2012).

3. Multi-objective wrapper for face feature selection

Genetic algorithms (GAs) are meta-heuristic optimization methods, inspired on the process of natural evolution, that are capable of finding global optima in complex search spaces (Youssef et al., 2001). These optimization algorithms need to evaluate a problem-dependent objective function to guide the search. However, in most real-world problems we may be interested in satisfying more than one objective, and the optimization of one objective may conflict with the other objectives. In general, the solution of a multi-objective optimization problem is not a single point, but a set of points known as the Pareto optimal front (Kim and Liou, 2012).

Different modifications to the traditional GAs were proposed in order to tackle multi-objective problems (Fonseca and Fleming, 1993). One generic approach is to combine the individual objective functions into a single aggregative function, or to consider all but one objective as constraints. Another generic approach is to determine a Pareto optimal, or nondominated set of solutions. This means, a set of solutions for which none of the objective values can be improved without detriment in some of the other objective functions. This approach takes advantage of the population-based nature of GAs, which allows the generation of several elements of the Pareto set in a single run (Coello Coello et al., 2007).

Particularly, the Multi-Objective Genetic Algorithm (MOGA) is a variation of the classical GA, in which the rank of an individual is the number of chromosomes in the population by which it is dominated (Fonseca and Fleming, 1993). This technique addresses the search toward the true Pareto front, while maintaining diversity in the population (Konak et al., 2006). A problem that arises in Pareto based multi-objective evolutionary algorithms is the diffi-

culty to preserve diversity among Pareto optimal solutions. The population tends to scatter around the existing optima forming stable sub-populations, or niches. One approach to overcome this difficulty, which is based on the concept of niching around promising points, makes use of a sharing function as proposed by Fonseca and Fleming (1993). Fitness sharing allows the MOGA to maintain the population diversity while encouraging the search for solutions in unexplored sections of a Pareto front. This is accomplished by reducing the fitness of solutions in densely populated areas of the search space (Kim and Liou, 2012). The MOGA, as other fitness sharing techniques, uses the parameter σ_s to define the size of the niche around a point in the Pareto front (Konak et al., 2006). In this way, the nearby solutions are penalized in order to maintain population diversity, and to promote the search around all the salient peaks in the domain of feasible solutions.

Here we propose and study three different wrappers for feature selection in face recognition applications. The first wrapper is a classical GA, in which each individual represents a particular selection of the set of facial features extracted from an input image by means of ASM. The second wrapper that we propose is a multi-objective GA with an aggregative fitness function, which combines classification accuracy and the number of features in a single equation. Finally, we propose a third wrapper which consists of a MOGA, with the same objective functions considered for the second alternative. Additionally, in this case we also use mutual information as an additional objective, in order to minimize the interdependence of the selected features. The proposed multi-objective wrapper method is described as a diagram in Fig. 2.

The selection of individuals is done considering the set of coefficients represented by each chromosome, using the tournament selection scheme. This consists on choosing a few individuals at random from the population in order to run a competition, from which the winner is selected for reproduction. To evaluate a particular individual, a set of images is used to compute the objective functions. In order to perform the evaluation, first, the feature vectors that represent the images are assembled with the coefficients indicated by the chromosome.

The classical mutation and one-point crossover are used, and an elitist replacement strategy is applied in order to maintain the best individual for the next generation.

3.1. Fitness functions

In the proposed multi-objective wrappers, one of the target functions evaluates the feature set suggested by a given chromosome, providing a measure of the face classification accuracy. Therefore, a classifier is used as the first objective function, so that the success classification rate is considered for each evaluated individual. In order to guide the search, while maintaining a low computational cost, a simple classifier algorithm was considered. This classifier assigns the test face image, represented by its feature vector, to the class with the closest prototype (mean feature vector). The mean is first computed based on the feature vectors in the training set, and the Euclidean norm was used as distance in our experiments. Then, after an optimized solution is found, the k-nearest neighbors (KNN) classifier (Bishop, 2007) is used to evaluate the classification performance on the test set.

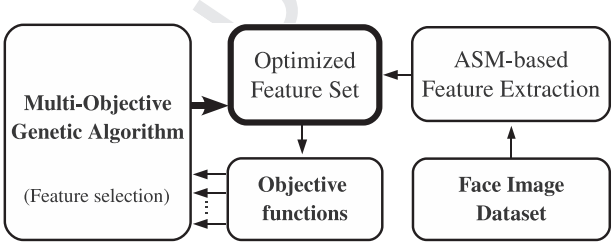

Fig. 2. General scheme of the proposed multi-objective wrapper.

It shall be observed that it is also beneficial to obtain a face image representation containing the smallest number of coefficients, which should be help in face image classification task, as discussed next.

3.1.1. Aggregative fitness function

For the aggregative approach we used a fitness function that combines classification accuracy and the number of features in a single equation. The proposed aggregative fitness function is:

$$F_a = \alpha F_1 + \frac{1-\alpha}{F_2}, \tag{3}$$

where α is a parameter that assigns a relevance to each objective. The first term of F_a corresponds to the prediction accuracy, F_1 (the fitness function used in the standard GA), and F_2 accounts for the number of selected features. In our experiments we adjusted α between 0.7 and 0.9. The second objective function is defined as

$$F_2 = 100\left(1 - \frac{n}{L}\right), \tag{4}$$

so that we obtain a number in the same range as the classification rate. Here, n is the number of coefficients selected by the chromosome, and L is the length of the chromosomes.

3.1.2. Proposed multi-objective approach

For the proposed MOGA we used the objective functions F_1 and F_2, defined in (3) and (4), respectively. Also, we used an additional objective function designed to minimize the mutual information (MI) of the selected coefficients. We computed the MI for every pair of coefficients on the training data using the method proposed by Peng et al. (2005). We defined this third objective function as

$$F_3 = \frac{M^*}{1 + LM/n}, \tag{5}$$

where M^* is the sum of the MI calculated for all the available features (taken in pairs), and M is the sum of the mutual information calculated for the features selected by a chromosome.

Considering the three proposed target functions, all steps in the evaluation of a population by the proposed multi-objective wrapper are detailed in the Algorithm 1.

Algorithm: Population evaluation in the proposed wrapper.

for *each individual in the population* **do**
- Re-parameterize the face images using the features selected by the chromosome
- (given the complete set of ASM features obtained using (1))
- Train the classifier with the training set
- Test the classifier with the validation set
- Assign classification rate as the current value for F_1
- Assign the current value for F_2, based on the number of features (4)
- Assign the current value for F_3, based on MI (5)
- In the case of the aggregative AG, compute the total fitness (3)

3.2. Chromosome codification

In this work, the mean of the color chrominance channels C_r and C_b of the YCbCr color space were used as features for describing

each of the 68 ASM landmark points, meaning that the dimensionality of a complete feature vector is $N \times Q = 136$ (Behaine and Scharcanski, 2012). We considered two different approaches for coding the chromosomes, yielding search spaces of significantly different sizes. In the first case, each gene represents a particular feature, independently of the landmark point associated to it. Thus, in this approach the chromosome size is 136, and each feature associated to a given landmark point can be selected individually and independently. In the second chromosome coding alternative, each gene in a chromosome represents one of the ASM landmark points, so the chromosome value indicates whether the corresponding features are used or not, and hence the chromosome size is reduced to 68. In both coding alternatives, the initialization consists on a random selection of the genes (values) in the chromosomes, since no restriction was applied to the re-combinations of features.

4. Experimental results and discussion

A set of face images from the Essex Face Database was used in our experiments (Vision Group, 2007), which contains a significant diversity of individuals and expression changes. In order to make a comparative evaluation of our experimental results with respect to other approaches available in the literature, 100 face classes were used. Five face images per class were randomly selected for training and other fifteen face images per class were separated for the test set (Behaine and Scharcanski, 2012).

As stopping criteria for the optimization, we considered a maximum of 500 generations, and convergence was assumed after 100 generations without fitness improvement. After the optimization step, the classification performance with the selected feature subsets was evaluated on the test set, which was not used for the feature selection process. That is, the data from the test set was not used for the fitness evaluation during the optimization, which allowed to estimate the generalization performance of the optimized feature subsets. This test was performed employing a KNN classifier (with $k = 1$). We carried out several optimization experiments, considering different alternatives and combination of parameters, and here we discuss the most relevant.

The experimental results are presented and discussed next. Section 4.1 discusses the experiments performed with the most simple approach studied in this paper, using the single-objective wrapper. Then, in Section 4.2, the results obtained with the proposed multi-objective strategies (i.e. aggregative GA and MOGA) are addressed. Finally, Section 4.3 presents a comparative analysis of the obtained results.

4.1. Single-objective optimization

In this Section, we first describe the experiments that involve chromosomes of length 136 (as explained before), which will be referred to as GA-136. The classifier described in Section 3.1 was used in the evolution, which was evaluated on the training data set in order to compute the fitness of each candidate solution. The GA population consisted of 30 individuals, and crossover and mutation probabilities were set to 0.8 and 0.025, respectively. In this case, the proposed GA converged to a set of 62 features, and the KNN classifier achieved an accuracy of 97.20% on the test data set.

Another set of experiments were conducted with single-objective optimization and GA-136 chromosomes. In order to obtain a better generalization performance, we enlarged the training data set using the Smoothed Bootstrap Resampling (SBR) method (Young, 1990). When the amount of data is not enough to ensure statistically significance, this method can be used to create new samples by adding noise to the feature values of the original samples. In particular, zero mean Gaussian noise with $\sigma = 0.1$ was used in our experiments, since this value allowed to preserve the variance of the original train data. Accordingly, in the next experiments (GA-136 + SBR), 20 SBR examples were generated for each class in order to perform the fitness evaluations. After the convergence of the GA, 68 features were selected, which allowed the classifier to achieve an accuracy of 97.40% on the test data. Therefore, we can infer that the resampling of the training data allows better generalization.

However, compared with the previous case, a larger subset of features was selected. A plot of the maximum fitness value obtained as the number of generations is increased is shown in Fig. 3(a). Note that the convergence of the GA required about 220 generations in this experiment.

The following approach tested, as explained in Section 3, consisted in reducing the length of the chromosomes to the number of landmark points (68). This means that, within each chromosome, the selection of a given landmark implies that both of the corresponding features are used. As a result of this experiment, referred to as GA-68 + SBR, we obtained a reduced feature set of size 56. With this feature set we obtained 98.0% of classification accuracy on the test data set, suggesting that the reduction of the chromosome size simplified the search space, making the search easier for the GA. Fig. 3(b) shows the evolution of the fitness value, and it can be verified that the best solution was found after only 63 generations. When compared to Fig. 3(a), it suggests that the codification strategy with smaller chromosomes, in addition to the resampled training data set, allowed a faster convergence of the GA.

4.2. Multi-objective optimization

In this section, we discuss the experimental results obtained by using the simultaneous optimization of multiple objectives. We first used a classical GA with the aggregative fitness function given in (3), taking into account the number of features besides the classification accuracy. As in the previous case, we studied both the codification alternatives with chromosome lengths 136 and 68, and used SBR samples for training.

Figs. 4(a) and (b) show the convergence plots for the optimizations using chromosomes of length 136, and the aggregative fitness function with $\alpha = 0.8$ and $\alpha = 0.85$, respectively. In the first case, GA-Aggre-2ob-136 + SBR ($\alpha = 0.8$), the GA converged to a set of only 32 features, and the KNN classifier achieved an accuracy of 97.40% on the test data set. With a similar experiment but using $\alpha = 0.85$, we obtained a set with ten additional features (42), which lead to a small improvement on classification accuracy of the test set (97.80%).

On the other hand, conducting the same experiments indicated above, but using chromosomes of length 136, we obtained a subset of 46 features with classification accuracy of 97.60%, and a subset of 56 features giving an accuracy of 97.80% on the test set, with $\alpha = 0.8$ and $\alpha = 0.85$, respectively. For these experiments, the fitness behaviors for different generations are shown in Figs. 4(c) and (d). It is noticeable that the convergence of the GA takes a longer time to optimize two objectives simultaneously, in contrast to the optimizations with a single objective discussed before.

The last group of experiments consists in using a MOGA to optimize two and three objectives simultaneously. In addition to classification accuracy and the number of features, in these experiments we also considered the minimization of the mutual information between selected features as a third objective. For the problem in hand, we obtained the most interesting results when σ_s was set to 0.09 and 0.1.

Please cite this article in press as: Vignolo, L. D., et al. Feature selection for face recognition based on multi-objective evolutionary wrappers. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.03.032

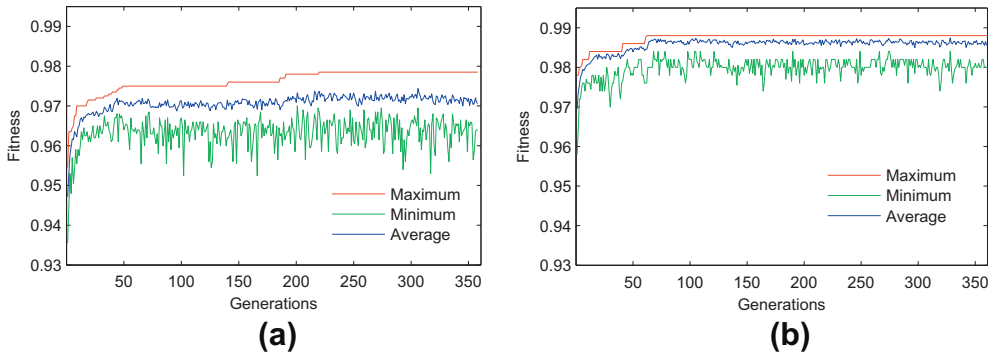

Fig. 3. Convergence of the GA in the experiments: (a) GA-136 + SBR and (b) GA-68 + SBR.

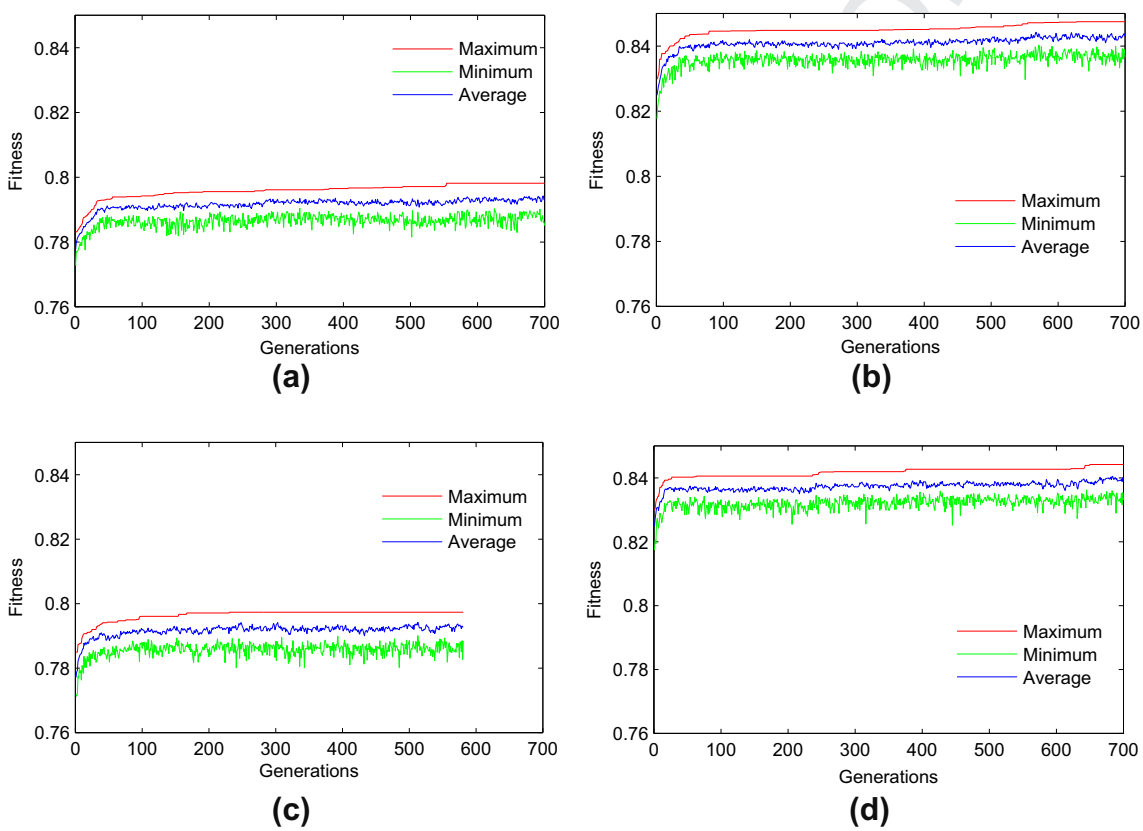

Fig. 4. Convergence of experiments: (a) GA-Aggre-2ob-136 + SBR with $\alpha = 0.8$, (b) GA-Aggre-2ob-136 + SBR with $\alpha = 0.85$, (c) GA-Aggre-2ob-68 + SBR with $\alpha = 0.8$, and (d) GA-Aggre-2ob-68 + SBR with $\alpha = 0.85$.

Several optimization experiments were conducted with the MOGA, first combining classification accuracy and the number of features, and then also including the mutual information measurement. Performing the optimization with two objectives (MOGA-2ob, as with the aggregative GA) and chromosomes of length 136, we obtained a subset of 37 features ($\sigma_s = 0.09$) giving an accuracy of 97.30% on the test set, and subset of 32 features ($\sigma_s = 0.1$) giving an accuracy of 96.67% on the test set. With chromosomes of length 68, we obtained a subset of 38 features giving an accuracy of 97.53%, and a subset of 30 features giving an accuracy of 97.30% on the test set. In this way, we compare the MOGA and the aggregative GA, showing that the performances of both are similar, except for a slight improvement of the MOGA in the later case.

On the other hand, when we also consider the minimization of mutual information (MOGA-3ob). We obtained a subset of only 26 features giving an accuracy of 97.00% ($\sigma_s = 0.09$), and a subset of 30 features which obtained 97.53% of accuracy on the test set ($\sigma_s = 0.09$), with chromosomes of length 136. Finally, with chromosomes of length 68, we obtained a subset of 36 features giving an accuracy of 97.93% ($\sigma_s = 0.09$), and a subset of 38 features giving 98.00% of accuracy on the test set ($\sigma_s = 0.09$).

4.3. Comparative analysis and discussion

Table 1 summarizes the results of the aforementioned experiments, and compares the performances obtained by the optimized subsets of features with two different approaches representing the state of the art. The second column shows the classification accuracy achieved by the different feature sets, obtained with the proposed wrapper optimization method on the test data set, and the third column shows the number of features involved. The last column exhibits the relative error reduction (RER) with respect to the

Table 1
Q4 Classification results obtained for the test data.

Method	Accuracy(%)	Number of features	Relative error reduction
DFBFR (Demirel & Anbarjafari, 2009)	93.73	2×100^2	–
Enhanced ASM (Behaine & Scharcanski, 2012)	95.33	54	(reference)
GA-136	96.93	62	34.26%
GA-136 + SBR	97.40	68	44.33%
GA-68 + SBR	**98.00**	56	**57.17**%
GA-Aggre-2ob-136 + SBR ($\alpha = 0.8$)	97.40	32	44.33
GA-Aggre-2ob-136 + SBR ($\alpha = 0.85$)	97.80	42	52.89
MOGA-2ob-136 + SBR ($\sigma_s = 0.09$)	97.30	37	42.18
MOGA-2ob-136 + SBR ($\sigma_s = 0.1$)	96.67	32	28.69
MOGA-3ob-136 + SBR ($\sigma_s = 0.09$)	97.00	**26**	35.76
MOGA-3ob-136 + SBR ($\sigma_s = 0.1$)	97.53	30	47.11
GA-Aggre-2ob-68 + SBR ($\alpha = 0.8$)	97.60	46	48.61
GA-Aggre-2ob-68 + SBR ($\alpha = 0.85$)	97.80	56	52.89
MOGA-2ob-68 + SBR ($\sigma_s = 0.09$)	97.53	38	47.11
MOGA-2ob-68 + SBR ($\sigma_s = 0.1$)	97.30	30	42.18
MOGA-3ob-68 + SBR ($\sigma_s = 0.09$)	97.93	36	55.67
MOGA-3ob-68 + SBR ($\sigma_s = 0.1$)	**98.00**	**38**	**57.17**

Enhanced ASM (Behaine and Scharcanski, 2012), meaning the percentage by which the error rate is reduced. As illustrated by this table, the optimized representations obtained by the evolutionary wrappers obtained better classification performances. It should be observed that these optimized representations provided larger feature sets when compared to the Enhanced ASM. However, the feature set provided by GA-68 + SBR improves the accuracy of the Enhanced ASM in more than 4%, with two more features.

The multi-objective approaches obtained significantly smaller feature subsets, specially if compared with the Enhanced ASM approach, with better classification performances. For instance, the smallest subset found consists of only 26 features and allows a significant reduction of the classification error, with respect to the Enhanced ASM, with RER 35.76%. Moreover, the solutions found by MOGA provided fewer features and, at the same time, produced accuracies that are similar to those obtained by the single-objective GA. For instance, the MOGA-3ob-68 + SBR allowed to reduce the classification error as much as our earlier GA-68 + SBR (RER 57.17%), using only 38 coefficients (features).

The aggregative multi-objective approach is useful to find small feature sets with reduced classification error, and the MOGA approaches provided better solutions (that is, almost the same accuracy with fewer features). Additionally, the minimization of the mutual information as a third objective provides solutions with a better compromise between classification error and the number of features. However, it is important to observe that in this experiments we favor solutions that provide high classification accuracy more than those with fewer features.

An interesting performance analysis can be obtained by changing the 100-class problem into a binary classification task, and then computing the ROC curve according to the methodology proposed in Bolle et al. (2005). For this binary classification task we took the 15 test patterns of a given class and assigned them as the *registered user* class, and all of the remaining test patterns, from the other 99 classes, were assigned to the *unregistered user* class. This was repeated for each of the 100 classes (each time a different class was labeled as registered) and the classification results obtained were averaged. As the unregistered users are unknown, the training patterns corresponding to this class were not used in the classification (we used only the patterns corresponding to the *registered user* class). Instead of using the KNN classifier, the rule to classify the test samples was based on the Euclidean distance to the training samples of the *registered user* class. This rule can be described simply as follows: if the distance from the test image to each of the training (registered) users is less than the threshold δ, it is labeled as *registered*; otherwise the test image is classified as *unregistered*.

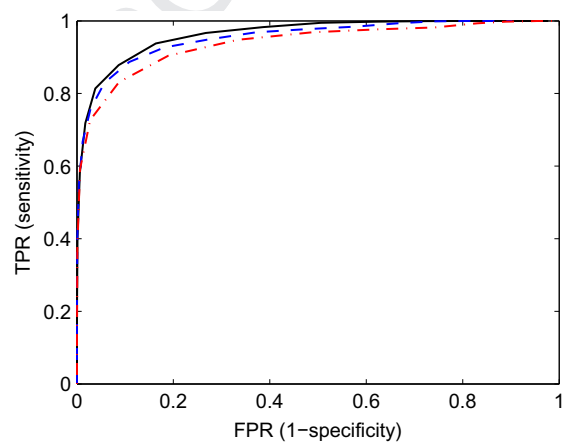

Fig. 5. ROC curve generated by varying the threshold δ in the binary classification task. The solid line corresponds to the MOGA-3ob-68 + SBR, the dashed line to GA-68 + SBR, and the dash-dot line to the complete feature set.

Fig. 5 shows the ROC curves constructed with the true positive rate (TPR) and false positive rate (FPR) indexes, obtained by averaging the results for the 100 binary classification tests. The classification performance obtained with the feature set MOGA-3ob-68 + SBR (solid line), with the feature set GA-68 + SBR (dashed line), and with the complete feature set (dash-dot line), for different values of threshold δ (varying from 0 to 120) are shown. High TPR values indicate that most of the test samples that belong to the *registered* class are correctly classified. On the other hand, high values of FPR occur when *unregistered* samples are labeled as *registered*. As can be seen in Fig. 5, to obtain the highest TPR we need to tolerate a FPR different of zero. It is important to observe that our optimized feature sets allow to improve on the classification results obtained with the complete feature set, obtaining a higher TPR without increasing the FPR. Also, analyzing the ROC curves it can be noticed that the 38-feature set obtained with the MOGA shows a significant improvement in classification performance with respect to the 56-feature set obtained with the classical GA (the same observation applies to the complete feature set). This confirms our hypothesis that it is beneficial to minimize the size of the feature set. Also, it can be noticed that the use of the resampling method allowed to obtain better results.

5. Conclusions and future work

This paper presented and compared multi-objective wrappers, based on evolutionary computation techniques, designed to optimize the feature selection process in face image classification. The proposed wrappers provide feature sets of different sizes and face class discrimination capabilities, and the choice of the most appropriate wrapper should be guided by the requirements of the problem in hand (e.g. reduced feature set combined with a high classification accuracy, or just focus on high classification accuracy). The experiments were performed on a well known face image data set, where the face images were represented using the ASM approach. These experiments revealed that the optimized feature sets offer improved classification accuracy in comparison with other state of the art approaches. Probably because these optimized face representations provide better class separability in the feature space, while simplifying the classification task. Furthermore, the dimensionality of the ASM-based representation was significantly reduced, which also helps to avoid overfitting. Hence, the proposed strategy provides a valid alternative for the selection of relevant features for face recognition.

In the future, we plan to perform experiments with a larger data set, with increased variability of pose and illumination, and explore other options in terms of feature set optimization. We would like to explore other multi-objective optimization algorithms such as PESA-II or NSGA-II (Coello Coello et al., 2007), in order to compare the performance with the MOGA. Also, a measure of compactness (Stegmayer et al., 2012) could be also considered as objective function in order to improve the clustering of classes in our evolutive wrapper. Moreover, we would consider the use of other heuristic search methods, such as particle swarm (Kennedy and Eberhart, 1995; Tsai and Kao, 2011) and scatter search (Mart et al., 2006).

Acknowledgments

The authors would like to thank SPU (Secretaría de Políticas Universitarias, Ministerio de Educación, Argentina) and CAPES (Coordenadoria de Aperfeiçoamento de Pessoal de Ensino Superior, Brazil) for financial support, and the Vision Group from the University of Essex (UK), for providing the face image database. Also, the authors wish to acknowledge the support provided by Agencia Nacional de Promoción Científica y Tecnológica and Universidad Nacional de Litoral (with PAE-PICT 00052, CAID 012-72), and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) from Argentina.

References

Behaine, C., & Scharcanski, J. (2012). Enhancing the performance of active shape models in face recognition applications. *IEEE Transactions on Instrumentation and Measurement, 61*(8), 2330–2333. http://dx.doi.org/10.1109/TIM.2012.2188174.

Bishop, C. M. (2007). *Pattern recognition and machine learning* (first ed.). Springer.

Bolle, R., Connell, J., Pankanti, S., Ratha, N., Senior, A. (2005). The relation between the roc curve and the cmc. In *4th IEEE workshop on automatic identification advanced technologies* (pp. 15–20). http://dx.doi.org/10.1109/AUTOID.2005.48.

Cevikalp, H., Triggs, B. (2010). Face recognition based on image sets. In *IEEE conference on computer vision and pattern recognition (CVPR) 2010* (pp. 2567–2573). http://dx.doi.org/10.1109/CVPR.2010.5539965.

Charbuillet, C., Gas, B., Chetouani, M., & Zarader, J. (2009). Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification. *Speech Communication, 51*(9), 724–731 [Special issue on non-linear and conventional speech processing].

Chatterjee, S., & Bhattacherjee, A. (2011). Genetic algorithms for feature selection of image analysis-based quality monitoring model: An application to an iron mine. *Engineering Applications of Artificial Intelligence, 24*(5), 786–795. http://dx.doi.org/10.1016/j.engappai.2010.11.009.

Coello Coello, C., Lamont, G., & Van Veldhuizen, D. (2007). *Evolutionary algorithms for solving multi-objective problems. Genetic and Evolutionary Computation* (second ed.,). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-0-387-36797-2.

Cootes, T., Taylor, C., Cooper, D., & Graham, J. (1995). Active shape models – their training and application. *Computer Vision and Image Understanding, 61*(1), 38–59.

Demirel, H., & Anbarjafari, G. (2009). Data fusion boosted face recognition based on probability distribution functions in different colour channels. *EURASIP Journal on Advances in Signal Processing* (25), 25:1–25:10. http://dx.doi.org/10.1155/2009/482585.

Fonseca, C. M., & Fleming, P. J. (1993). Genetic algorithms for multiobjective optimization: Formulation discussion and generalization. In *Proceedings of the 5th International Conference on Genetic Algorithms* (pp. 416–423). San Francisco, CA, USA: Morgan Kaufman Publishers Inc..

Handl, J., & Knowles, J. (2006). Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal of Computational Intelligence Research, 2*(3), 217–238.

Hill, A., Cootes, T., & Taylor, C. (1996). Active shape models and the shape approximation problem. *Image and Vision Computing, 14*, 601–607. http://dx.doi.org/10.1016/0262-8856(96)01097-9 [6th British Machine Vision Conference].

Hsu, H.-H., Hsieh, C.-W., & Lu, M.-D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications, 38*(7), 8144–8150. http://dx.doi.org/10.1016/j.eswa.2010.12.156.

Kennedy, J., Eberhart, R. (1995). Particle swarm optimization. In *IEEE international conference on neural networks* (Vol. 4, pp. 1942–1948). http://dx.doi.org/10.1109/ICNN.1995.488968.

Kim, J., Çetin, M., & Willsky, A. S. (2007). Nonparametric shape priors for active contour-based image segmentation. *Signal Processing, 87*(12), 3021–3044. http://dx.doi.org/10.1016/j.sigpro.2007.05.026 [Special section: Information processing and data management in wireless sensor networks.].

Kim, H., & Liou, M.-S. (2012). New fitness sharing approach for multi-objective genetic algorithms. *Journal of Global Optimization*, 1–17. http://dx.doi.org/10.1007/s10898-012-9966-4.

Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering and System Safety, 91*(9), 992–1007. http://dx.doi.org/10.1016/j.ress.2005.11.018.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J., & Abidi, M. A. (2005). Recent advances in visual and infrared face recognition – a review. *Computer Vision and Image Understanding, 97*(1), 103–135. http://dx.doi.org/10.1016/j.cviu.2004.04.001.

Li, S. Z., & Jain, A. K. (2011). *Handbook of face recognition*. Springer.

Li, Y.-X., Kwong, S., He, Q.-H., He, J., & Yang, J.-C. (2010). Genetic algorithm based simultaneous optimization of feature subsets and hidden Markov model parameters for discrimination between speech and non-speech events. *International Journal of Speech Technology, 13*, 61–73. http://dx.doi.org/10.1007/s10772-010-9070-4.

Mart, R., Laguna, M., & Glover, F. (2006). Principles of scatter search. *European Journal of Operational Research, 169*(2), 359–372. http://dx.doi.org/10.1016/j.ejor.2004.08.004 [Feature cluster on scatter search methods for optimization].

Milborrow, S., & Nicolls, F. (2008). Locating facial features with an extended active shape model. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer vision – ECCV 2008. Lecture Notes in Computer Science* (Vol. 5305, pp. 504–513). Berlin/Heidelberg: Springer [10.1007/978-3-540-88693-8-37]. Q3

Oh, S.-K., Yoo, S.-H., & Pedrycz, W. (2013). Design of face recognition algorithm using PCA–LDA combined for hybrid data pre-processing and polynomial-based RBF neural networks: design and its application. *Expert Systems with Applications, 40*(5), 1451–1466. http://dx.doi.org/10.1016/j.eswa.2012.08.046.

Pedrycz, W., & Ahmad, S. S. (2012). Evolutionary feature selection via structure retention. *Expert Systems with Applications, 39*(15), 11801–11807. http://dx.doi.org/10.1016/j.eswa.2011.09.154.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238. http://dx.doi.org/10.1109/TPAMI.2005.159.

Shi, J., Samal, A., & Marx, D. (2006). How effective are landmarks and their geometry for face recognition? *Computer Vision and Image Understanding, 102*(2), 117–133. http://dx.doi.org/10.1016/j.cviu.2005.10.002.

Stegmayer, G., Milone, D. H., Kamenetzky, L., López, M. G., & Carrari, F. (2012). A biologically-inspired validity measure for comparison of clustering methods over metabolic datasets. *IEEE ACM Transactions in Computational Biology and Bioinformatics, 9*(3), 706–716.

Tsai, C.-Y., & Kao, I.-W. (2011). Particle swarm optimization with selective particle regeneration for data clustering. *Expert Systems with Applications, 38*(6), 6565–6576. http://dx.doi.org/10.1016/j.eswa.2010.11.082.

Turk, M., Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition CVPR'91* (pp. 586–591). http://dx.doi.org/10.1109/CVPR.1991.139758.

Vignolo, L. D., Milone, D. H., Scharcanski, J., Behaine, C. (2012). An evolutionary wrapper for feature selection in face recognition applications. In *IEEE international conference on systems, man, and cybernetics (SMC) 2012*, Seul (Corea) (pp. 1286–1290). http://dx.doi.org/10.1109/ICSMC.2012.6377910.

Vignolo, L. D., Milone, D. H., & Rufiner, H. L. (2013). Genetic wavelet packets for speech recognition. *Expert Systems with Applications, 40*(6), 2350–2359. http://dx.doi.org/10.1016/j.eswa.2012.10.050.

Vignolo, L. D., Rufiner, H. L., Milone, D. H., & Goddard, J. C. (2011a). Evolutionary splines for cepstral filterbank optimization in phoneme classification. *EURASIP Journal on Advances in Signal Processing Volume*. http://dx.doi.org/10.1155/2011/284791.

Vignolo, L. D., Rufiner, H. L., Milone, D. H., & Goddard, J. C. (2011b). Evolutionary cepstral coefficients. *Applied Soft Computing, 11*(4), 3419–3428. http://dx.doi.org/10.1016/j.asoc.2011.01.012.

Vision Group, University of Essex (UK), Face recognition data (2007). <http://cswww.essex.ac.uk/mv/allfaces/faces94.html>.

Wang, C., Li, Y., & Song, X. (2013). Video-to-video face authentication system robust to pose variations. *Expert Systems with Applications, 40*(2), 722–735. http://dx.doi.org/10.1016/j.eswa.2012.08.009.

Wen, Y. (2012). An improved discriminative common vectors and support vector machine based face recognition approach. *Expert Systems with Applications, 39*(4), 4628–4632. http://dx.doi.org/10.1016/j.eswa.2011.09.119.

Young, G. A. (1990). Alternative smoothed bootstraps. *Journal of the Royal Statistical Society Series B (Methodological), 52*(3), 477–484.

Youssef, H., Sait, S. M., & Adiche, H. (2001). Evolutionary algorithms simulated annealing and tabu search: A comparative study. *Engineering Applications of Artificial Intelligence, 14*(2), 167–181. http://dx.doi.org/10.1016/S0952-1976(00)00065-8.

Zheng, Z., Jiong, J., Chunjiang, D., Liu, X., & Yang, J. (2008). Facial feature localization based on an improved active shape model. *Information Sciences, 178*(9), 2215–2223.