An Evolutionary Wrapper for Feature Selection in Face Recognition Applications

Leandro Vignolo,
Diego Milone, *Member, IEEE*
Research Center for Signals, Systems and
Computational Intelligence
Universidad Nacional del Litoral
Santa Fe, Argentina
ldvignolo@fich.unl.edu.ar

Carlos Behaine,
Jacob Scharcanski, *Senior Member, IEEE*
Electrical Engineering Graduate Programme
Informatics Institute
Universidade Federal do Rio Grande do Sul
Porto Alegre, RS 91501-970, Brazil
jacobs@inf.ufrgs.br

***Abstract*—Active shape models is an adaptive shape-matching technique that has been used for locating facial features in images. However, when a number of features is extracted for each landmark point, distortions caused by noise or illumination, and the dimensionality of the final representation, have a negative impact in the performance of a classifier. In this paper, an evolutionary wrapper for selection of the most relevant set of features for face recognition is presented. The proposed strategy explores the space of multiple feasible selections using genetic algorithms. Experimental results show that the proposed approach allows to improve the classification performance in comparison with another enhanced method and a state of the art face recognition approach.**

***Index Terms*—wrappers, evolutionary algorithms, feature selection, face recognition.**

I. Introduction

The purpose of a face recognition system is to automatically identify or verify an individual identity using a digital image [1]. Face recognition has received significant attention in biometrics, motivating important developments in several research areas such as image processing and artificial intelligence. In face recognition we classify a given face in K different face classes. This is usually done by comparing the features extracted from the image with those extracted from a database of face images [2]. This task remains a challenge because the unavoidable changes in expression, pose, and illumination, which introduce variability in the extracted features with respect to the training data [3].

In face modeling with Active Shape Models (ASM), a number of facial points are selected from an input image, but only some of these points are useful for characterizing the face, while the others have small contributions or are noisy. As the training of ASM converges towards salient edges, if these edges are distorted by noise or illumination, erroneous feature matchings might arise [4]. Despite recent improvements made to the ASM techniques, the matching errors are usually high at some face locations [5], [6]. Even after some new implementations that improve the landmark location accuracy, the detection of facial features with varying pose and illumination is still a challenge [7], [3]. Moreover, once a set of face points is selected by the ASM method, a number of features describing each face location need to be extracted. Then, the resulting feature vectors representing the faces are usually of high dimensionality, which makes the task of the classifier more difficult [8]. Also, large feature sets are prone to overfitting and, hence, to achieve poor generalization performance [9]. In [4], the performance of ASM was improved by weighting the features according to a method based on adjusted mutual information. As the authors shown, this criterion allowed the selection of the more relevant landmark points, in order to improve the face classification results. However, the flexibility provided by the full set of features obtained with ASM has not yet been fully exploited by means of feature selection techniques, in order to reduce the dimensionality of the representation while improving classification results. On the other hand, significant progress has 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 [10], [11], and for the search of optimal representations [12], [13], [14].

This work presents a wrapper for selecting the most relevant features for face recognition. In order to explore the space of feasible solutions we use a genetic algorithm, which is guided by the accuracy obtained in a face classification task. 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 by an independent data set.

The organization of this paper is as follows. First a brief description of the ASM is given in Section II-1, and next our wrapper method for the selection of relevant features for face images is presented in Section II-A. Section III describes the experiments and discuss the results on face classification. Finally, the general conclusions and future work proposal are presented.

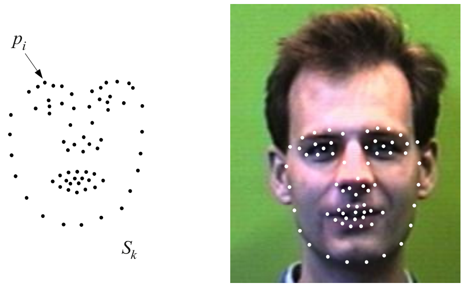

Fig. 1. Illustration of the landmark points used to model a face (a) and their location on an image (b) [4].

II. Wrapper for the Selection of Features in Face Images

1) Facial features based on Active Shape Models: ASMs are point distribution models which iteratively deform to fit the shape of an example of the object [5]. The shapes are constrained by these point distribution models (PDM) to vary only according to a training set of examples. The shape of an object is represented by a set of points and the algorithm matches the model to a new image. In the case of face recognition, the ASM is trained on a set of training face images and N PDM points are used to represent the shape of each face within a face class (Fig. 1(a)). However, the location of the points of a PDM, the *landmark* points, on face images can have location or matching errors (Fig. 1(b)) [4]. Then, the faces of each class $k = 1, ..., K$ will be represented by N landmark points $S_{k,\epsilon} = p_i(x_i + \epsilon_{x_i}, y_i + \epsilon_{y_i})^k$, where $i = 1, ..., N$, (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. Each facial characteristic that is considered relevant (e.g. eye centers, mouth contours, etc.) is represented by a landmark p_i, and the particularities of that point in the image are described by Q features (e.g. chrominance, texture, etc.). The features for landmark p_i will be denoted $\{F_{j,i}\}$, with $j = 1, ..., Q$.

Each one of the N landmark points p_i, is described by the mean and the variance of the measurements of every feature j, $\mu_{F_{j,i}}$ and $\sigma^2_{F_{j,i}}$, respectively. These are computed for all features $F^m_{j,i}$, with $m = 1, \ldots, M$, where M is the number of training image samples:

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

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

Here $\mu_{j,i}(r,q) = \frac{1}{M}\sum_{m=1}^{M} F^m_{j,i}(r,q)$, $\sigma^2_{j,i}(r,q) = \frac{1}{M}\sum_{m=1}^{M} \left(F^m_{j,i}(r,q) - \mu_{j,i}(r,q) \right)^2$ and (r,q) are the pixel coordinates within the window $W = w \times w$, centered at the landmark point p_i, and calculated for the jth image feature [4]. In order to consider the feature variability within the $w \times w$ vicinity of landmark p_i, the window variance was used in Eq. 2. We assume that the probability density of the location errors is approximately Gaussian [15]. Here $w = 2\max\{\sigma_\epsilon\}$ is adopted, where σ_ϵ is the standard deviation of the landmark location errors, measured during ASM training.

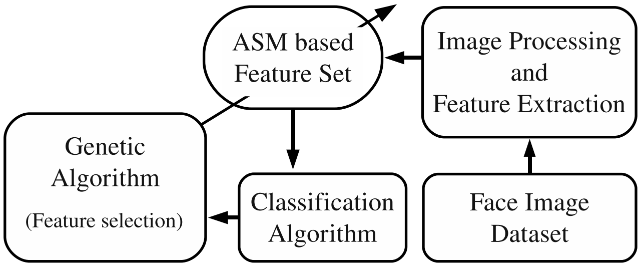

Fig. 2. General scheme of the proposed wrapper method.

Note that the face detector by Demirel et al. [16] is used in this work. Further details about the location of landmark points in face images and the set of features used can be found in [4].

A. Evolutionary Wrapper

Genetic algorithms (GA) are meta-heuristic optimization methods inspired by the process of natural evolution [17]. A typical GA consists of three different operators: selection, variation and replacement [18]. Selection simulates the advantage of the fittest individuals in nature, giving them more chance to reproduce. The variation operators combine information from different individuals and also maintain population diversity. The chromosome of an individual within the population, stores the information of a possible solution, and its fitness is measured by a problem specific objective function. The selected parents are mated to generate the offspring, which replaces the population, and the cycle is repeated until a desired termination criterion is reached [19].

The proposed wrapper consists on a classical GA with binary chromosomes, in which every individual represents a different selection of the facial features extracted from an input image by means of ASM. In this GA, the objective function needs to evaluate the image feature set suggested by a given chromosome, providing a measure of the separability obtained for the face classes. Therefore, a classifier algorithm is used as fitness function, so that the success classification rate is assigned as fitness value for each evaluated individual. The scheme of the proposed wrapper method is shown in Fig. 2.

The selection of individuals is done considering the set of coefficients represented by each chromosome, using the roulette wheel selection scheme. The chromosomes which allow better classification results are assigned higher probability. When a particular individual is evaluated, a set of images is classified in order to estimate the discrimination capability of the given configuration. In order to do this, each feature vector is first reduced to the subset of coefficients indicated by the chromosome.

For guiding the search of the GA, maintaining a low computational cost, a simple classifier algorithm was considered as objective function. This classifier assigns the class to which its mean is closer to the feature vector of test image. The mean is first computed from the feature vectors of the images in the tranning set, and the distance used is based on the Euclidean norm. Then, once an optimized solution was found, a classifier based on the k-nearest neighbors (KNN) rule [8] was used in order to evaluate its classification performance on the test set.

The GA uses the classic mutation and one-point crossover, and an elitist replacement strategy was applied in order to maintain the best individual to the next generation. As the stopping criteria for the optimization, we considered that the GA had converged after 100 generations without fitness improvement, or when the maximum of 500 generations was reached.

Based on [4], 68 ASM landmark points were considered, and the two color chrominance channels C_r and C_b from the YCbCr color space were used as the features for describing each landmark point. This means, that the dimensionality of a complete feature vector is $N \times Q = 136$. Regarding the codification of every individual within the GA, we considered two different alternatives, involving search spaces of significantly different sizes. In the first case, each gene represents a particular feature, independently from the landmark associated to it. Thus, in this approach the chromosome size is 136, and the two features associated to a given landmark point can be selected independently of one another. In the second alternative, each gene in a chromosome represents one of the ASM landmark points, so its value indicates whether the corresponding pair of coefficients is selected or not (the chromosome size in this case is reduced to 68). As no restriction applies for the combinations of coefficients, in both alternatives the GA initialization consists on a random settling of the genes in the chromosomes.

III. Results and Discussion

For the experiments, a set of images from the Essex Face Database was used [20], containing significant individual diversity and considerable expression changes. In order to compare results with previous works, 100 face classes were used, 5 images per class were randomly selected for training and other 15 images per class were separated for the test set. As mentioned in Section II-A, the ASM algorithm was applied to obtain 68 landmark points, using a window of 11 pixels.

In order to estimate the generalization performance of the optimized feature subsets, the data from the test set was not used for the evaluation of the fitness during the optimization. In this manner, after the optimized feature subsets were obtained, their classification performance was evaluated with a different data set that was not involved in the feature selection process. This final evaluation was performed with the described test data set and employing a KNN classifier (with $k = 1$). We ran many optimization experiments considering different alternatives and combination of parameters, and here we discuss those that we found most relevant.

In the first experiment, referred to as GA-136+MEAN, we used a chromosome size 136 (as described before). We used the classifier based on the distances to the mean of each class, which was evaluated on the training data set in order to compute the fitness of every candidate solution. For the GA, the population size was set on 30 individuals while crossover and mutation probabilities were set to 0.8 and 0.025, respectively. The optimization converged to a set of 62 features, which achieved a performance of 97.2% with the KNN classifier on the test data set. Fig. 3 shows a plot of the fitness value against the number of generations for this experiment, allowing to appreciate the fast convergence of the GA.

For further analysis, in order to obtain better generalization capability, we decided to enlarge the training data set using the Smoothed Bootstrap Resampling (SBR) method [21]. This method is used when data is not enough to ensure statistically significance, and the new samples are created by adding noise to their feature values. Specifically, we used zero mean Gaussian noise with $\sigma = 0.1$, since this value allowed to preserve the variance of the original tranning data. Therefore, in the second experiment (GA-136+MEAN+SBR), we used 20 examples for each class in order to evaluate the classification performance during the optimization. The GA converged to a solution consisting of 68 features, which achieved 97.4% success rate on the test data. As it can be noticed, the resampling of the training data allowed to achieve better generalization. Despite that, in this case, a larger number of features were selected. In Fig. 4 the fitness value against the number of generations is shown, where it can be noticed that the GA took much more generations (220) to converge in this experiment.

As stated in Section II-A, in the third experiment we reduced the length of the chromosomes to the number of landmark points (68). So that, within the chromosome, the selection of a given landmark point implies that the two corresponding features are used. In this experiment, labeled GA-68+MEAN+SBR, the GA converged to a more reduced feature set of length 56, which allowed to obtain a classification accuracy of 98.0% on the test data. This result suggests that the reduction of the chromosome size produced a simpler search space, allowing the GA to find a better solution. Fig. 5 shows a plot of the fitness value against the number of generations for this experiment, where it can be appreciated that the best solution was found after only 63 generations. Also, by comparing to Fig. 4, it can be noticed that the codification strategy with reduced chromosome size allowed a faster convergence when using a resampled training data set.

Table I summarizes the results of the experiments described above and compares the performances obtained by the optimized subsets of features, with two different state of the art approaches. The second column exhibit the classification accuracy achieved by the different methods on the test data set, the third column shows the number of features involved, and the last column displays the error reduction percent with respect to the Enhanced ASM [4]. As it can be seen, all the optimized representations obtained by means of the evolutionary wrapper allowed better classification performance. It should be noticed that they provided larger feature sets when compared to the Enhanced ASM. However, the feature set provided by GA-

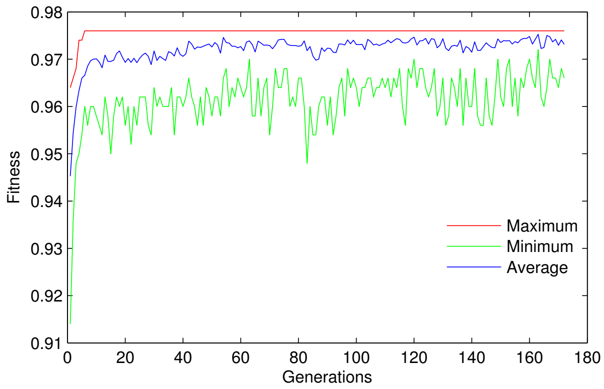

Fig. 3. Convergence of experiment GA-136+MEAN.

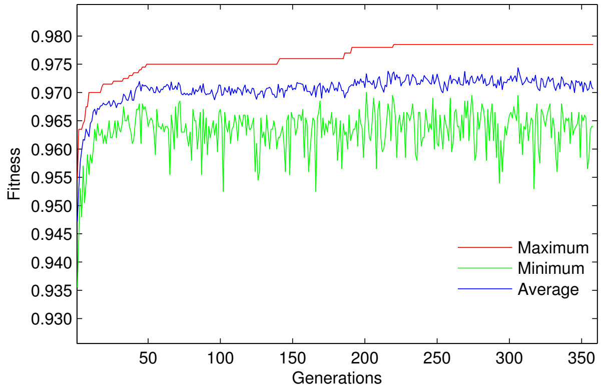

Fig. 4. Convergence of experiment GA-136+MEAN+SBR.

68+MEAN+SBR improves the accuracy of the enhanced ASM in more than 4%, with only two more features.

In order to perform further performance analyses we changed the 100-classes task into a binary classification task and computed the ROC curve, following the methodology presented in [22]. For this binary classification task we took the 15 test patterns of a given class and labeled them as the *registered user* class, and other 15 test patterns were taken randomly from the other 99 classes in order to compose the *unregistered user* class. This was repeated for each of the 100 classes and the classification results obtained were averaged. As the unregistered users are unknown, the tranning patterns corresponding to this class were not used in the classification, but only the patterns corresponding to the *registered user* class. Then, the rule used to classify the test samples was based on the euclidean distance to the tranning samples of the *registered user* class. The rule was as follows: if the distance from the test pattern to the tranning patterns was less than the threshold, δ, it was labeled as *registered*; otherwise, the sample was classified as *unregistered*. Fig. 6 shows the ROC curves constructed with the true positive rate (TPR) and false positive rate (FPR) indexes obtained from the average of the results for the 100 binary classifications. The classification performance obtained with feature set GA-68+MEAN+SBR (solid line), and the complete feature set (dashed line), for different values of threshold δ (varying from 0 to 120) is shown. High values of TPR indicate that most of the test samples that belong to the *registered* class are classified as such. Besides, high values of FPR occur when *unregistered* samples are labeled as *registered*. As can be seen from the figure, the highest TPR implies that a FPR different from zero needs to be tolerated. It is important to note that our optimized feature set allows to improve the results of the complete feature set, obtaining higher TPR values without increase of the FPR.

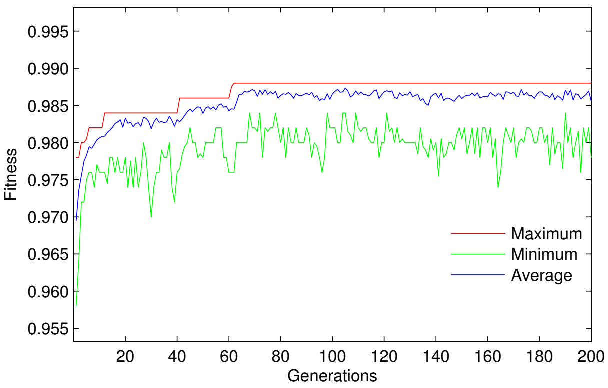

Fig. 5. Convergence of experiment GA-68+MEAN+SBR.

TABLE I
CLASSIFICATION RESULTS OBTAINED FOR THE TEST DATA

Method	Accuracy	# of features	Error reduction
DFBFR [16]	93.73%	2×100^2	-
Enhanced ASM [4]	95.33%	54	(reference)
GA-136+MEAN	96.93%	62	34.26%
GA-136+MEAN+SBR	97.40%	68	44.33%
GA-68+MEAN+SBR	**98.00%**	56	57.17%

IV. CONCLUSIONS AND FUTURE WORK

In this work, two different wrapper optimization strategies have been proposed, exploiting the benefits provided by evolutionary computation techniques, in order to search for an optimal feature set with application to face recognition. The results, obtained in a classification task using a well known data set, shown that the optimized feature set provides improved accuracy in comparison to other state of the art approaches. This means that the task of a classifier is simplified when using the optimized representation, due to a better class separation in the space of selected features. Therefore, the proposed strategy provides a valid alternative for the selection of relevant features for face recognition.

In future work we would consider the use of a multi-objective GA [23], in order to minimize the number of relevant features while maximizing the classification rate. Also, we would consider the use of other heuristic search methods, such as particle swarm [24], [25] and scatter search [26]. Besides, the use of a larger data set with increased variability would be considered for future experiments in order to prove the robustness of the method.

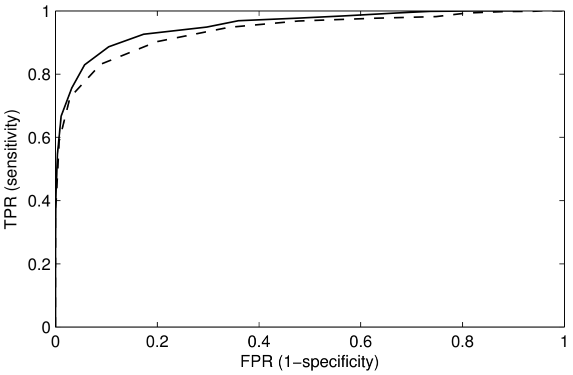

Fig. 6. ROC curve generated by varying the threshold in the binary classification task. The solid line corresponds to the optimized feature set and the dashed line to the complete feature set.

Acknowledgment

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 facial 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

[1] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*. Springer, Aug. 2011.

[2] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 2567 –2573.

[3] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Computer Vision - ECCV 2008*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin / Heidelberg, 2008, vol. 5305, pp. 504–513, 10.1007/978-3-540-88693-8-37.

[4] C. A. R. Behaine and J. Scharcanski, "Enhancing the performance of active shape models in face recognition applications," *TO APPEAR in IEEE Transactions on Instrumentations and Measurement*, 2012.

[5] A. Hill, T. Cootes, and C. Taylor, "Active Shape Models and the shape approximation problem," *Image and Vision Computing*, vol. 14, no. 8, pp. 601–607, 1996, 6th British Machine Vision Conference.

[6] J. Kim, M. Çetin, and A. S. Willsky, "Nonparametric shape priors for active contour-based image segmentation," *Signal Processing*, vol. 87, no. 12, pp. 3021 – 3044, 2007, special Section: Information Processing and Data Management in Wireless Sensor Networks.

[7] Z. Zheng, J. Jiong, D. Chunjiang, X. Liu, and J. Yang, "Facial feature localization based on an improved active shape model," *Information Sciences*, vol. 178, no. 9, pp. 2215–2223, May 2008.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, Oct 2007.

[9] J. Handl and J. Knowles, "Feature subset selection in unsupervised learning via multiobjective optimization," *International Journal of Computational Intelligence Research*, vol. 2, no. 3, pp. 217–238, 2006.

[10] S. Chatterjee and A. Bhattacherjee, "Genetic algorithms for feature selection of image analysis-based quality monitoring model: An application to an iron mine," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 5, pp. 786 – 795, 2011.

[11] Y.-X. Li, S. Kwong, Q.-H. He, J. He, and J.-C. Yang, "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*, vol. 13, pp. 61–73, 2010, 10.1007/s10772-010-9070-4.

[12] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification," *EURASIP Journal on Advances in Signal Processing*, vol. Volume 2011, 2011, doi:10.1155/2011/284791.

[13] ——, "Evolutionary Cepstral Coefficients," *Applied Soft Computing*, vol. 11, no. 4, pp. 3419 – 3428, 2011.

[14] C. Charbuillet, B. Gas, M. Chetouani, and J. Zarader, "Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification," *Speech Communication*, vol. 51, no. 9, pp. 724 – 731, 2009, special issue on non-linear and conventional speech processing.

[15] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Computer Vision and Image Understanding*, vol. 102, no. 2, pp. 117 – 133, 2006.

[16] H. Demirel and G. Anbarjafari, "Data fusion boosted face recognition based on probability distribution functions in different colour channels," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 25:1–25:10, january 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/482585

[17] H. Youssef, S. M. Sait, and H. Adiche, "Evolutionary algorithms, simulated annealing and tabu search: a comparative study," *Engineering Applications of Artificial Intelligence*, vol. 14, no. 2, pp. 167 – 181, 2001.

[18] T. Bäck, U. Hammel, and H.-F. Schewfel, "Evolutionary computation: Comments on history and current state," *IEEE Trans. on Evolutionary Computation*, vol. 1, no. 1, pp. 3–17, 1997.

[19] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, 1992.

[20] Vision Group, University of Essex (UK), "Face recognition data." [Online]. Available: http://cswww.essex.ac.uk/mv/allfaces/faces94.html

[21] G. A. Young, "Alternative smoothed bootstraps," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 52, no. 3, pp. pp. 477–484, 1990. [Online]. Available: http://www.jstor.org/stable/2345671

[22] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior, "The relation between the roc curve and the cmc," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, oct. 2005, pp. 15 – 20.

[23] C. Coello Coello, G. Lamont, and D. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed., ser. Genetic and Evolutionary Computation. Berlin, Heidelberg: Springer, 2007.

[24] D. Chen and C. Zhao, "Particle swarm optimization with adaptive population size and its application," *Applied Soft Computing*, vol. 9, no. 1, pp. 39 – 48, 2009.

[25] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE International Conference on Neural Networks*, vol. 4, August 1995, pp. 1942–1948.

[26] R. Mart, M. Laguna, and F. Glover, "Principles of scatter search," *European Journal of Operational Research*, vol. 169, no. 2, pp. 359 – 372, 2006, feature Cluster on Scatter Search Methods for Optimization.