IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 17, 2019, accepted June 12, 2019, date of publication June 18, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923626

Clusters of Features Using Complementary Information Applied to Gender Classification From Face Images

JUAN E. TAPIA[1], (Graduate Student Member, IEEE), AND CLAUDIO A. PEREZ[2], (Senior Member, IEEE)

[1]R&D Department, Universidad Tecnologica de Chile -INACAP, Santiago 720, Chile
[2]Department of Electrical Engineering, and Advanced Mining Technology Center, Universidad de Chile Av. Tupper 2007, Santiago 720, Chile

Corresponding author: Juan E. Tapia (j_tapiaf@inacap.cl)

This work was supported in part by CONICYT, through grant FONDECYT INICIACION 11170189, Universidad Tecnologica de Chile-INACAP, and in part by the Department of Electrical Engineering and Advanced Mining Technology Center, Universidad de Chile, under FONDECYT 1161034.

ABSTRACT Face recognition performance by computers has been shown to be more accurate than that of humans. However, a bias with soft-biometrics features has been detected. This bias reduces recognition performance when gender is used. Feature selection for gender classification from face images is a difficult problem since faces contain symmetrical and redundant features. We argue that traditional methods, based on mutual information using pairs of features to estimate the relevance and redundancy among features, fail to select the right set of features in cases where there are strong spatial correlations among features, which is the case with facial images. In this paper, a new method is proposed fusing a filter and a wrapper to measure the relationships among image features, and to select feature clusters based on mutual information for gender classification. We applied this method on nine face datasets using an SVM classifier. We were able to achieve 98.2% correct gender classification in the testing partition using the UND, 95.56% with the Morph II, 98.33% on the LFW, and 98.66% on celebA databases. We validated the results using a cross-test with three different datasets: COFW, Adience, and Image of Groups, that were not used to define the parameters of our method. Additionally, the method was tested with a Random Forest. All the results achieved are better than those previously published on the same databases, and with a significantly smaller number of total features.

INDEX TERMS Gender classification, gender from faces, cluster of features, mutual information, feature selection.

I. INTRODUCTION

Face recognition (FR) has grown to become a prominent biometric technique for identity authentication and has been widely applied in many areas, such as airports, public security, and daily life [51].

The National Institute of Standards and Technology (NIST) has run multiple facial recognition tests since 1993. Every successive test has shown performance improvements, sometimes by an order of magnitude.[1] Automated facial recognition accuracy is now far superior to that reached by humans [51]. Several methods have been proposed for FR including transformations, Gabor, other hard biometric features, and the combination of hard and soft biometric features [12], [13], [40], [61]. Currently, face recognition systems have surprisingly better accuracy, but we have to deal with demographic bias in biometric systems [9], [51].

Face recognition data can be prone to error, which can implicate people for major crimes or illegal entrances they have not committed. Facial recognition has been shown to be inaccurate and not good particularly for recognizing African-Americans and other ethnic minorities, females, and young people, often misidentifying or failing to identify them, thus impacting certain groups disparately.[2]

The SC-37 ISO-Committee has been developing an ISO/IEC Technical Report # 22116 called 'Identifying and mitigating the differential impact of demographic factors in biometric systems, with focus on the impact of gender, age, and ethnicity on the iris, face, and fingerprints. Even one

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Khurram Khan.

[1]https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

[2]https://www.bbc.com/news/technology-33347866

of the biggest gender projects, 'Gender Shades', presented problems with face detection and gender classification using the systems developed by IBM, Microsoft, and Face++ [9]. These systems work very well for white males but not well for African-American females.

The previous examples show that gender classification is an open problem whose resolution could help to reduce the bias in FR systems. Most of the FR systems are 'black boxes' and are unable to identify the relevant face features that allow discriminating gender.

Gender classification from facial images is one of the most challenging problems in image analysis research because of the symmetry and complementarity of faces [1], [6], [13], [30], [33], [44], [49], [50], [56]–[58], [60], [65], [66], [73]. In a biometric recognition framework, gender information may lead to searching only half of the database. Most gender classification methods reported in the literature use all of the features extracted for classification purposes [1], [3], [37], [44], [71], [73]. As a result, gender-irrelevant information could be fed into the classifier, and the classifier generalization capacity could be reduced, especially when the training set is small. It has been shown both theoretically and empirically that reducing the number of irrelevant or redundant features increases the learning efficiency of the classifier significantly [4], [18], [26], [42], [48].

The problem of adequate feature selection has not been solved for complex problems such as gender classification from faces because of the redundancy, relevance, symmetry of the face, feature complementarity, and feature location, among other factors [8], [18], [22], [40], [63]. Using more features implies increasingly higher computational cost in the feature extraction process, slowing down the classification process. Exhaustive evaluation of possible feature (2^N) subsets is usually computationally prohibitive. *MI* has solved the problem of feature selection in complex scenarios partially with methods that use only the information between pairs of features such as, *mRMR*, *DISR*, and *CMIM* [8], [18], [22]. These traditional approaches work very well in simple scenarios but do not capture the relationship among three or more features.

There are limitations to the computed ranking of *MI* between pairs, and the trade-off between relevance and redundancy. The main limitation is derived from the fact that possible redundancies among variables are not taken into account. Indeed, two redundant yet highly relevant variables, when taken individually, will both be well-ranked. Contrarily, two variables can be complementary to the output (i.e., highly relevant together) while each of them appears to be irrelevant when taken individually. As a results, the features could be ranked erroneously, or removed by the ranking filter process.

Ranking does not consider previous information for each feature; before computing the ranking for all features, they are all given the same weights. Features that are not individually relevant may therefore become relevant in the context of others, and/or features that are individually relevant may not be useful at all because of possible redundancies.

II. STATE OF THE ART

Alexander et al. [1], the best gender classification was based on shape features on the UND database [23]. The best result was reached when fusing 3 types of features (intensity, shape, and texture) and 3 sizes of images (20×20, 36×36, 128×128). However, many more inputs were used by fusing the 3 scales and the 3 types of features. The total number of inputs was increased nearly ninefold reaching a gender classification rate of 91.19% for the UND database using 46,845 features.

Han et al. [30], presented a generic framework for automatic demographic estimation from a single face image. They extracted biologically inspired features from a face image, and selected demographic informative features using a boosting algorithm. They then used a hierarchical estimator consisting of between-group classification and within-group regression to predict age, gender, and race. For gender classification they used the MORPH II database, the PCSO dataset, the FERET dataset, and the LFW dataset, reaching 97.6%, 97.1%, 96.8%, and 94%, respectively.

Jia et al. [35], presented a simple and effective method for classifying gender by training a linear classification algorithm on a massive dataset assembled and labeled entirely by automated means. Four million images and more than 60,000 features were used in total to train online classifiers, and the method, tested with the LFW dataset, reached 96.86% accuracy.

Moeini et al. [47], proposed an automatic feature extraction method with two types of features. Then, two separate dictionaries for male and female genders were defined for representing the gender in facial images. Also, two dictionary learning methods were proposed to learn the defined dictionaries in the training process. The Sparse Representation Classification (SRC) is adopted for classification in the testing process. Finally, a probability decision making approach is proposed to classify the gender from estimated values by SRC and proposed gender formulation. They used three public available databases including the FERET, LFW and Groups databases to compare their results to the state-of-the-art. The results reached were: 91.9%, 94.9% and 84.4%, respectively.

Castrillana et al. [14], the authors analyze the state-of-the-art gender classification accuracy on three large datasets: MORPH, LFW and Groups. They discussed the difficulties and bias on the databases, concluding that the most challenging and with the wildest complexity was the Groups database. The achieved results were compared with those of a Convolutional Neural Network, achieving an accuracy of 94.2% reducing the gap with other simpler datasets.

Only a few methods have been published using Deep Learning for gender classification with faces [19], [31], [36], [41], [43], [70].

Levi et al. [41] proposed a method to classify age and gender automatically using a simple convolutional neural network architecture that can be used with limited data. The method reached an accuracy of 84.7% using the Adience dataset.

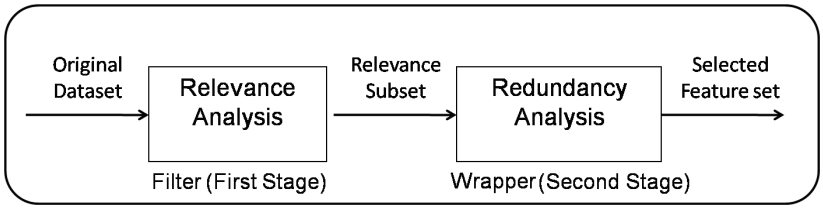

FIGURE 1. Block diagram of the Filtrapper method, a new framework for feature selection composed of two stages: Filter and Wrapper, decoupling relevance and redundancy analysis.

Liu et al. [43] proposed a novel deep learning framework for face attribute prediction in the wild using CNNs, LNet, and Anet, with carefully designed pre-training strategies. The method is robust to background clutter and face variation using the LFWA and CelebA datasets with 30 or more attributes. This approach reached 98% accuracy for the CelebA and 94% for the LFWA datasets.

Hand et al. [31] used a multi-task CNN and 40 attributes with the LFWA and CelebA datasets. Gender is among these attributes. This approach reached 98.17% accuracy for the CelebA dataset and 94.1% for the LFWA dataset in gender classification. They demonstrated through experiments that a multi-task framework for attribute prediction outperforms independent classifiers. Taking advantage of implicit and explicit relationships among attributes promotes improved attribute prediction, which will lead to improved gender classification.

In our previous work, [50], [57], [58], we showed that feature selection and fusion methods using faces improved gender classification accuracy. We reported an extension of the use of feature selection based on computed Mutual Information (MI) between pairs of features, or between features and classes, that reached the highest classification performance published at that time on the FERET database with 99.13% accuracy using 18,900 features, and on the UND database with 94.01% accuracy using 14,200 features. The approaches present in this paper outperform the previous results with a less number of features.

An interesting method to achieve feature selection is to compute MI by ranking each individual features by their relevance using the filter approach. [5]. Feature ranking methods are generally considered to be fast and effective, especially when there is a large number of features and the number of training examples are relatively small (For example 10K features and 100 examples). However, the ranking reduces the selection power of the criterion because it does not compute any form of complementarity among variables [64], making it ill-suited for feature selection on datasets that have high levels of complementarity. Using more features implies increasingly higher computational cost in the feature extraction process, slowing down the classification process, and also increasing the time needed for training and validation, which may lead to classification over-fitting [29].

III. CONTRIBUTION

In this paper, we propose a fusion of hybrid filter/wrapper method based on a relief method [38], [39] that employs weighted MI using the relationships among neighbor images applied to gender classification. This method has two stages, the first one called Filter, and the second one called Wrapper. See Figure 1. Both stages represent a Filtrapper approach and will be explained in detail in Section 3.

An automatic feature selection method based on a gender classification scheme is proposed. Feature selection is an important stage in classifying images under controlled and uncontrolled scenarios. We demonstrate that by adding complementary information from clusters of images the error rate can be reduced and a robust classifier developed [59].

In our method we modified the following feature selection methods: minimum Redundancy Maximal Relevance (mRMR) [48], Conditional Mutual information Maximization ($CMIM$) [22], and Double Input Symmetrical Relevance ($DISR$) [45] by using weighted features and renaming them as: $W-mRMR$, $W-CMIM$, and $W-DISR$.

We argue that our proposed weighted methods, $W-CMIM$, $W-mRMR$, and $W-DISR$, are more general purpose MI filters because they determine the relevance of three or more features in a cluster and may have complementarity with the class variable or the candidate feature. This is especially relevant in pattern recognition in image problems such as gender classification. The main contributions of our paper are the following:

In our proposed method we use clusters of images to estimate the quality of groups of features, thereby solving the problem of redundancy and relevance in gender classification from faces. One of the main contribution of the proposed weighted methods, $W-mRMR$, $W-DISR$ and $W-CMIM$, is that they share information among features, and reduce the quantity of redundant and irrelevant features. Therefore, selected features are more discriminative than in the current approaches, which use only the information between pairs of features, such as in traditional MI, or those that use the raw data. These traditional approaches do not capture the relationship among three or more features. In our proposed method, we use clusters of images to estimate the quality of groups of features. We can therefore obtain better results with a smaller number of features making feasible, for example, gender classification applications on mobile devices.

A second contribution is that the results reported in the manuscript are significantly better than those previously reported in the literature on the problem of identifying gender from faces. Additionally, we used nine different datasets to support our results, which is a significant increase relative to previously reported work. The datasets are: UND, LFW-a,

LFW-b, Morph II, CelebA, LFWA, Adience, Groups, and COFW datasets. In total we used more than 300,000 images.

To validate the results a cross-test was performed using three databases: COFW, Adience, and Image of Groups (Groups). As a result of our proposed method, the total number of selected features for gender classification was significantly reduced maintaining or improving the accuracy. The reduction in the number of features implies a significant decrease in processing time, making real-time applications of gender classification feasible.

IV. BACKGROUND

In this section, we briefly introduce some basic concepts and notions from information theory that are used in the proposed feature selection method. Information theory provides an intuitive tool for measuring the uncertainty of random variables and the information shared by them. Feature Selection [22], *MI* [17], and Complementary Information [63], [64] are critical concepts.

A. FEATURE SELECTION

Feature selection involves selecting features from a dataset in order to improve classification accuracy and decrease computation complexity [29]. It is similar to feature extraction which involves the creation of feature vectors from the original dataset via manipulating data space. The latter technique may be considered a ''superset'' of feature selection [17], [29].

Feature selection can be classified into three main groups [29]: Filters, Wrappers, and Embedded. Feature selection methods can also be categorized on the basis of the search strategies used [46]. The following search strategies are commonly used: Forward selection that starts with an empty set and adds features greedily one at a time; Backward elimination that starts with a feature set containing all the features and removes features greedily one at a time.

The fusion of Filter/Wrapper approach [32], [54], [62] is computationally less expensive than the original wrapper approach because the evaluation of the predictor performance by cross-validation is achieved for only a few preselected feature sets.

Feature selection is a broad field in continuous evolution, since selection of the most relevant and non-redundant features have not been achieved for complex problems such as those in gender classification [29], [32], [63], [64]. A new strategy to select the most important, relevant, and non-redundant features, using neighboring image information based on mutual information was presented. Our results show that significant improvements in gender classification can be reached with our proposed feature selection method. The method could be extended to other areas of feature selection as well. Additionally, by reducing the number of selected features, the proposed method reduces the computational time of the feature extraction process.

B. MUTUAL INFORMATION

MI is defined as a measure of how much information is contained jointly in two variables [17], or how much information of one variable determines the other variable. *MI* is the foundation for information theoretic feature selection since it provides a function for computing the relevance of a variable with respect to the target class [21]. The *MI* between two variables, x and y, is defined based on their joint probabilistic distribution $p(x, y)$ and the respective marginal probabilities $p(x)$ and $p(y)$ as:

$$MI(x, y) = \sum_{i,j} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \quad (1)$$

We use categorical *MI* in this paper, which can be estimated by tallying the samples of categorical variables in the data building adaptive histograms to compute the joint probability distribution $p(x, y)$ and the marginal probabilities $p(x)$ and $p(y)$ based on the Fraser algorithm [24]. The concept of minimal redundancy allows selection of feature pairs that are maximally dissimilar. If two features are highly dependent on each other, the respective class-discriminating power does not change much if one of them is removed.

C. COMPLEMENTARY INFORMATION FOR FEATURE SELECTION

For multiple variables, Interaction Information (*II*) [8], [64], [72] was proposed as a measure of the amount of information collected in a set of variables that is greater than the information present in any pair of those variables. Thus, Entropy [17] and *MI* [17] correspond to the first and second order measures of *II*, respectively, together with their third, fourth, and higher-order variants. *II* provides a way of characterizing the structure of multivariate information [15], [64]. Multivariate methods take feature reliancy into account and achieve better results because no simplifying supposition are made about feature independence [29]. The complementary information [63], [64], also known as synergy (interaction information), and measures the degree of relation between an individual feature *fi* and a group of features S, and belongs to the class C through the expression $II(f_i; S/C)$.

Conditional forms of the *MI* can be stated, but, unlike entropies, they condition the reduction of the *MI*. This is because the knowledge of a third variable can make the two original variables dependent upon each other:

$$II(X; Y/Z) = \sum_{z \epsilon Z} p(z) * MI(X; Y/Z = z). \quad (2)$$

The *II* measures the influence of variable Z on the amount of information shared between variables (X, Y) as:

$$II(\{X_1, \ldots, X_n\}) = MI(\{X_1, \ldots, X_{n-1}\}/X_n) - MI(\{X_1, \ldots, X_{n-1}\}). \quad (3)$$

Suppose we are given a random variable, S, and a random vector, $X = \{X_1, X_2, \ldots, X_{n-1}\}$. Our goal, then, is to decompose the information that X provides about S in terms of the

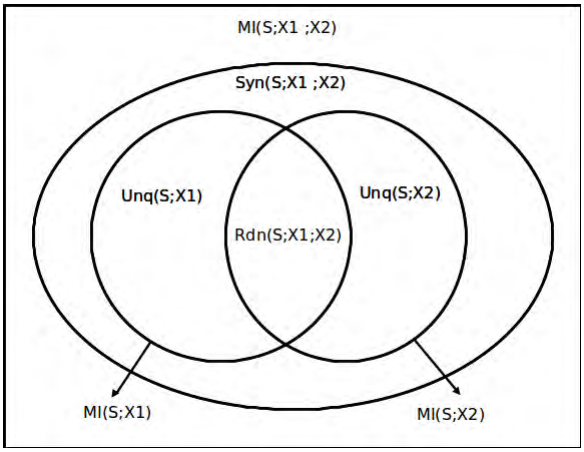

FIGURE 2. Venn diagram showing the multivariate information structure for three variables [64]. Labeled regions correspond to Unique information (Unq), Redundancy (Rdn) and Synergy (Syn).

partial information contributed either individually or jointly by various clusters of X.

Consider the simplest case of a system with three variables [64]. The total information that $X = \{X_1, X_2\}$ provides about S is given by the $II(S; X_1, X_2)$, with which X_1, X_2 we can identify three distinct possibilities.

First, X_1 may provide information that X_2 does not, or vice versa (unique information). For example, if X_1 is a copy of S, and X_2 is a degenerate random variable, then the total information from X is reduced to the unique information from X_1. Second, X_1 and X_2 contribute to the total information, providing the same or overlapping information (redundancy). For example, if X_1 and X_2 are both copies of S, then they provide complete information redundantly. Third, the combination of X_1 and X_2 may provide information that is not available from either one alone. This is called "Synergy" [63], [64]. (See Figure 2).

V. METHODS

In this paper we report the development of a fusion of hybrid feature selection method, composed of two stages, a Filter and a Wrapper, to classify gender from face images, obtaining which is called a Filtrapper [11], [54], [62]. Our method is based on *MI* and uses clusters of images to estimate the ranking, rather than pairs of features as do traditional methods based on *MI*.

The first stage, Filter, is used to estimate the relevance of the images based on the Relief-F algorithm reported in [38], [39], [53], [67], [69]. The goal of the Relief-F algorithm is to estimate the quality of features in relation to how well their values differentiate among images that are close to each other.

We transform each image in a vector of a matrix, A. Each row of A represents one image and each column in A represents one feature. Therefore we have a matrix A of size (M_{images} x $N_{features}$). The Filter stage has the follows steps:

a) From a Matrix A, we selected a random image (R_i).

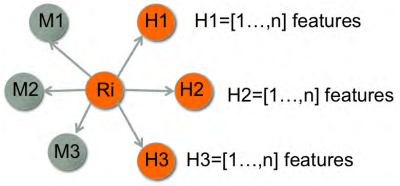

FIGURE 3. Representation of the one cluster created (*Cp*) in the first stage, based on Relief-F with three nearest images (p=3). *Ri* is a random image; *H*1, *H*2, and *H*3 represent the nearest neighbor images of class 1; and *M*1, *M*2, and *M*3 represent the nearest neighbor images of class 2. The neighbors are localized using Euclidean distance. In Figure 3, the number of features in one image is represented by the size of each image with the letter *n*.

b) Measure the Euclidean Distances (*ED*) among R_i and all the images belonging to class 1(Female) and all the images belonging to class 2 (Male).

c) Select the p images with the minimum distance for each class, (*mED*). Thus we created cluster of images of size p called Cluster p, (*Cp*). See Figure 3.

d) Estimate the average *MI* for all the images that belong to each clusters *Cp* called *MI* average cluster, (MI_{ac}):

$$MI_{ac} = \sum MI(R_i, Cp_i)/(p) \tag{4}$$

e) Rank the clusters in relation to the values of MI_{ac}.

f) Those previous steps (a-e) are repeated t times. (Parameter selected by the user.)

g) Finally, the best clusters from t times were selected using the Hausdorff distances [34] among clusters. The Hausdorff distance among clusters from two different patterns is the one that maximizes the minimal distance between different images of two clusters: $diff_{w-max}$:

$$diff_{w-max}(A, Cp_i, Cp_j) = diff_{features}(A, Cp_i^*, Cp_j^*), \tag{5}$$

where Cp_i^*, Cp_j^* are cluster images that satisfy this condition. See Figure 4. This relationship was used as a measure of information, or as a measure of class separability based on that feature cluster.

We used the information that belongs to the best cluster to weight the value of each n feature into the matrix A. See Figure 4.

These weights are normalized between 0-1 before computing the feature selection method. Then, we applied this information into three different feature selection methods based on *MI*: *mRMR*, *CMIM*, and *DISR*, considering the weight for all the features. The later computation defines the three new improved measures proposed in this paper: $W - mRMR$, $W - CMIM$, and $W - DISR$. Features depend on a wrapper using SVM [29].

The second stage of the fusion, Wrapper, is used to compute the best parameters of the algorithm. To assess the relevance of each feature, we define 3 parameters: t represents the number of times the process is repeated, p the number of images selected, and n the number of features selected from A. In this work, we changed t in steps of 5 up to

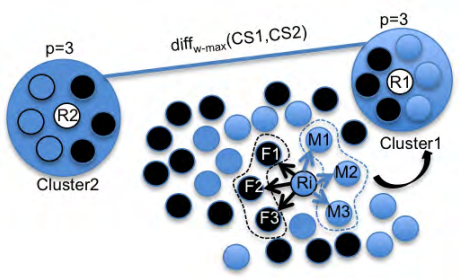

FIGURE 4. Illustration of two iterations of our proposed method using ($p = 3$). Three "male" nearest neighbors ($M1, M2, M3$), and three "female" nearest neighbors ($F1, F2, F3$) are used. The random images in this iteration are called Ri. The information distance of the nearest image to the Ri is used to update the weight vector and create a cluster of features based on the information among images. Cluster1 and Cluster2 represent two clusters that were created previously. The final clusters are selected using the Max-Haussdorff distance ($diff_{W-max}$).

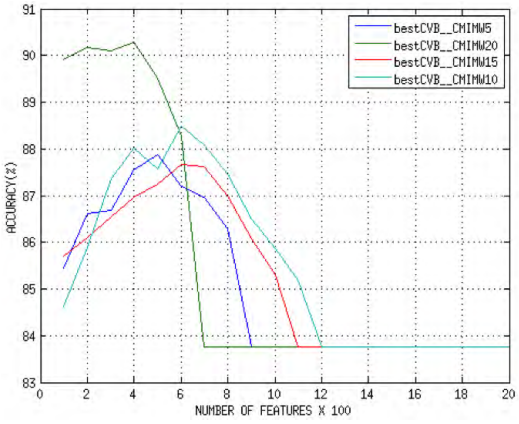

FIGURE 5. Shows the classification rate as a function of the number of features, from 100 to 2,000 selected features. The best result was reached with the $W - CMIM$ method using 400 features; $p = 20$ and 20 clusters ($t = 20$).

10, and n in steps of 100 features up to the total image size. We also explored values of p from 5 up to 50 images classes, searching for the value that would produce the best classification rate using a forward and backward selection from the Wrapper method. An SVM with Gaussian kernel was used as a classifier.

This filtrapper fusion approach is computationally less expensive than the original wrapper approach because the evaluation of the predictor performance, for example by a cross-validation test, is performed for only a few preselected feature sets.

For example, a combination of these 3 parameters can produce the best classification rate. When using a $p = 5$, we selected 10 nearest neighbors, 5 from class 1 (female), and 5 from class 2 (male). The process is repeated t times to create the t clusters. One example of the best parameters for the LFW database (set-a) is presented in Figure 5. The accuracy can be observed by changing the number of p from 5 to 20, on 300 images (150 males and 150 females) from the LFW database. Each image has 2,944 total features (64×46).

A. FEATURE SELECTION METHODS

1) WEIGHTED MINIMUM REDUNDANCY AND MAXIMAL RELEVANCE ($W - MRMR$)

Two forms of combining relevance and redundancy operations are reported in [18]; mutual information difference (MID), and mutual information quotient (MIQ). Thus, the $mRMR$ feature set is obtained by optimizing MID and MIQ simultaneously. The trade-off of both conditions requires to integrate them into a single criterion function [18], [48] as follows:

$$f^{mRMR}(X_i) = MI(c; fi) - \frac{1}{S}\sum MI(fi; fs), \tag{6}$$

where, $MI(c; fi)$ measures the relevance of the feature to be added for the class, and the term $\frac{1}{S}\sum_{fi\epsilon S} MI(fi; fs)$ estimates the redundancy of the fi_{th} feature with respect to the cluster of previously selected features, S.

In our approach, we compute the relevance and redundancy, adding the weight information from filter stage, therefore, $w_i(X_i)$ is used directly instead of the feature X_i, while computing the MI score to make the relevance ranking:

$$f^{W-mRMR}(X_i) = MI(c; fi) * w(i)^2 - \frac{1}{S}\sum MI(fi; fs) * w(i) * w(s), \tag{7}$$

where $w(i)^2$ and $w(s)$, are used to weight the importance of the relation between features.

For the computation of redundancy of each feature candidate f_i with the already selected features f_S, we employed the weight function instead of using the features themselves. We computed redundancy over all pairs according to the weights.

2) WEIGHTED CONDITIONAL MUTUAL INFORMATION MAXIMIZATION ($W - CMIM$)

The $CMIM$ criterion is a tri-variate measure of the information associated with a single feature about the class, conditioned upon an already selected feature [22]. It loops over the selected features and assigns each candidate feature a score based upon the lowest Conditional Mutual Information (CMI) between the selected features, the candidate feature, and the class [22], [29]. Then, the selected feature is the one with the maximum score.

$$CMIM = \begin{cases} arg\ max_{fi\in F}\{MI(fi; c)\} \\ for\ S = \emptyset arg\ max_{fi\in F/S}\left\{min_{fj\in S} MI(fi; c/fj)\right\} \\ for\ S \neq \emptyset. \end{cases} \tag{8}$$

The $CMIM$ criterion selects relevant variables, avoids redundancy and, unlike previous methods, does not ignore variable complementarity. However, it does not necessarily select a variable that is complementary to the already selected variables. In fact, a variable that has high complementarity information to the already selected variable will be has by a high (CMI). The $CMIM$ takes the minimum value as the

score for that feature because, at each iteration, the score for that feature can only decrease and therefore, scores that are already below the current best score are not updated, since they do not affect the computation. The *CMIM* algorithm only considers *CMI* and it does not improve the scores of highly complementary variables. This means it may not be well suited for datasets with high complementarity such as gene expression data. Our proposed $W-CMIM$ considers that the feature *fi* is relevant only if it provides high information about C, considering the synergy and complementarity detected by the modified Relief-F. In our proposed method, a product between *MI* and a weight function $w(i)^2$, is used to weight the importance of the candidate feature. This feature brings the information of its neighbors detected in the filter stage,

$$W-CMIM = \begin{cases} arg\ max_{fi\in F}\left\{MI(fi;c)*w(i)^2\right\} \\ for\ S=\emptyset \\ arg\ max_{fi\in F/S}\left\{min_{fj\in S}\,MI(fi;c/fj)*w(i)^2\right\} \\ for\ S\neq\emptyset. \end{cases} \tag{9}$$

3) WEIGHTED DOUBLE INPUT SYMMETRICAL RELEVANCE (W-DISR)

The DISR criterion [45] combines two properties of feature selection methods. First, a combination of variables can return more information from the output class than the sum of the information returned by each of the variables taken individually, considering the complementarity among variables [8], [29], [45]. Second, in the absence of additional information on how cluster d variables should be combined, a combination of the best performing clusters of $d-1$ variables is intuitively assumed to be the most promising set.

In DISR, it is assumed that the i_{th} node represents the variable X_i, the binary variable $w_i, i = 1,\ldots,n$ takes the value 1 if the i_{th} variable is selected and 0 otherwise, and the weight is computed by (10),

$$w_{i,j} = \frac{MI(X_{i,j};Y)}{H(X_{i,j})}, \tag{10}$$

W-DISR (11) is an improvement over DISR, it uses the weight function $w(i)^2$ computed by the W-Rank before computing *MI*, weighing the importance of the candidate feature. This weighted feature brings the information of its n neighbors detected in the first stage, guiding the selection method in search of the most relevant cluster of features,

$$w_{i,j} = \frac{MI(X_{i,j};Y)}{H(X_{i,j})} * w(i)^2. \tag{11}$$

VI. EXPERIMENTS, DATABASES AND FEATURE SELECTION

Gender Classification was assessed using two different feature selection experiments. The first used the Labeled Face in the Wild (LFW) [55], MORPH II [52], UND [57], LFWA and CelebA [43] face databases for gender classification. The second experiment validated the results using a cross-test with three different datasets: COFW [10], Adience [20], and Image of Groups (Groups) [25]. In all the experiments we followed the same protocol used by the authors of previous papers, especially with regard to the partition of the datasets for training and testing.

The databases LFW, LFW-A, and CelebA, have the coordinates available for cropping the face images. This data is available from the creators of the databases. For the 'UND' and 'MORPH II' databases, the OpenCV library was used to detect and crop the frontal images. A pre-trained Open-Face implementation [2] was also used to detect and crop the faces for the COFW, Image of Groups, and Adience Databases.

A. EXPERIMENTS

Experiment 1: A gender classification experiment was performed using using three sets of the LFW (set-a, set-b and set A), MORPH II, UND and CelebA face databases which are standard and challenging databases available for this purpose. For example, LFW contains images called Labeled Faces in the Wild. We selected features in steps of 100 features up to the 2,946th for image sizes of 64×46. The p nearest neighbors were computed in steps of 5. This allows comparison of our results with those published previously [1], [14], [27], [28], [30], [47].

The traditional dimensional reduction algorithms, such as Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) were applied [6]. PCA is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The objective is to select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component. The basic idea of the LDA is to use the well-known Fisher criterion to determine a number of discriminant vectors, and exploit them as transform axes by which samples are transformed into a new space. These vectors maximize the ratio of the between-class distance to within-class distance in the new space. It seems that the LDA outperforms the PCA in classification accuracy in many cases because of label information. However, Guo et al., [27] found that unsupervised dimensionality reduction methods, such as PCA, are not able to project face images to sufficiently discriminative subspaces.

In Experiment 1, the databases were partitioned to have 80% training data and 20% testing data. Results in Experiments 2 were obtained with a fivefold cross-validation, and an SVM classifier with a Gaussian Kernel. These datasets allow person-disjoint results to be computed; that is, no person has an image in both the training images and the testing images. For experiment with CelebA and LFWA we follows the same protocols suggested by the authors of [31], [43] for the train, test and validation datasets. The files are available in.[3]

[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

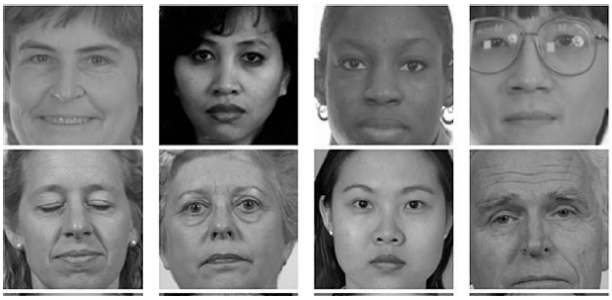

FIGURE 6. Examples of face images from the UND database.

Experiment 2: We validated our results using a cross-test with three different face databases: COFW [10], Adience [20] and Image of Groups (Groups) [25]. These databases show a diverse range of operational conditions and are representative of the challenges to be addressed in real-world situations. We resized all the images to 64×46 pixels. Further, we used a Random Forest classifier [7] with the model that reached the best results for Experiment 2 to show that the results are not dependent on the SVM classifier. The random forest model was trained with 500 Trees and Gini Index.

B. DATABASES AND CLASSIFIER

The images in the UND face database were taken in a controlled scenario and it contains a set of images from Collection B (See Figure 6). The image filenames used for training and testing, as well as the window crop around the subjects' faces, are available as text files on a web page as reported in Alexandre, 2010. It contains gray scale images of 487 frontal faces with 186 female and 301 male images, collected and annotated by the researchers. The images were not aligned.

Labeled Faces in the Wild contains 13,233 face color photographs of 5,749 subjects collected from the web to investigate gender classification of real world face images under unconstrained scenarios [55]. LFW is composed of real life faces, with varying facial expressions, illumination changes, head pose variations, occlusions, and use of make-up, and includes some of poor image quality. For LFW we used three sets of images. The first one, LFW set-a used 7,443 face images, (2,943 females and 4,500 males) and resized the images to 64 $\times$ 46 pixels, manually labeling the ground truth for the gender of each face. All the images were previously aligned with commercial software [55] See Figure 7 (top). The second one, LFW set-b consists of 13,010 images (6,505 females and 6,505 males) which cover images with different poses, and the third in according with the protocols suggested by [31], [43].

The MORPH-II database contains 55,134 images of more than 13,000 individuals. It was collected in the wild [52]. We chose 7,500 random face images (3,500 females and 4,000 males) and resized them to 64×46 pixels. See Figure 7 (bottom).

The COFW database is composed of 1,007 face images obtained under real-world conditions. The face images show

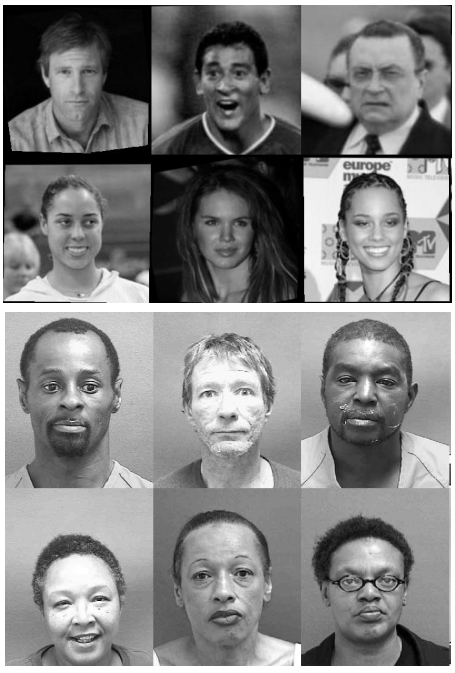

FIGURE 7. Example of face images from LFW (top row), Example of face images from MORPH II (bottom row).

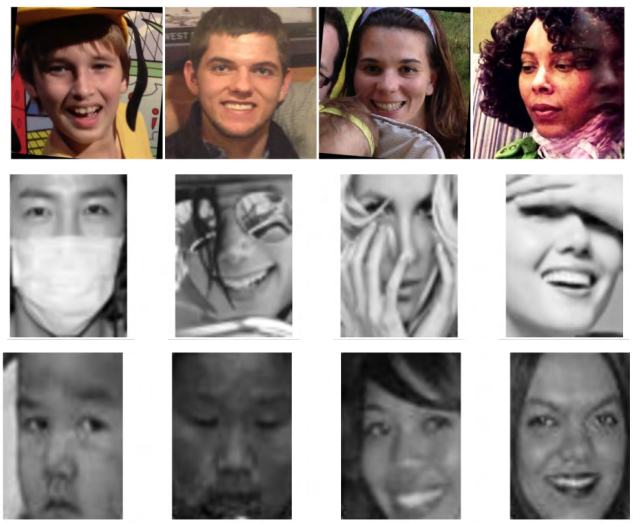

FIGURE 8. Examples of face images from Adience (top row), COFW (middle row) and Groups (bottom row) databases.

large variations in shape due to differences in pose and expression. Occlusions are also present in different degrees due to the use of accessories such as sunglasses, hats, and interactions with objects. This database is divided into two datasets: a training set of the original non-augmented 845 LFPW faces + 500 COFW faces (1,345 total), and a testing dataset with 507 COFW faces. We tested on the entire database (COFW-All). Gender labels were manually annotated. See Figure 8.

The Adience database is a collection of face images for testing age and gender classification methods. This database attempts to capture faces in a diverse variety of appearances, noise, pose, and lighting conditions in uncontrolled environments. Images are extracted from Flickr

FIGURE 9. Example of face images CelebA (top row) and LFWA datasets (bottom row).

albums created from automatic uploading of iPhone5 (or later) smartphone images. The total number of images is 26,580 corresponding to 2,284 different subjects. Gender, age, and subject labels are provided. This database defines five partitions for cross validation testing. We tested on each of these five partitions independently, only considering the frontal set as defined by the authors of the database. The OpenCV face detector was applied to these five partitions yielding a total of 10,632 face images (93.2% of all frontal images). The results reported are the average of the five partitions used for testing (Adience-All). See Figure 8.

The Groups dataset is composed of group shots and was built from 5,080 Flickr images containing 28,231 faces, labeled with age and gender. The images have people lying, standing or sitting, on elevated surfaces. People often have glasses, face occlusions, or unusual expressions. We tested on the whole database (Groups-All). See Figure 8.

In order to explore the performance of our method, we tested it using two additional large datasets: CelebA and LFWA [43]. See Figure 9. CelebA contains 10,000 identities, each of which has twenty images. There are 200,000 images in total. LFWA has 13,233 images of 5,749 identities. Each image in CelebA and LFWA was annotated with 40 face attributes and five key points by a professional labeling company. CelebA and LFWA have over 40 and 30 attribute labels, respectively. To compare the results, we used the same protocol indicated by the authors [43].

Figure 9, shows sample face images from the CelebA and LFWA datasets. The two datasets are available for research.

C. FEATURE SELECTION FOR EXPERIMENTS 1 AND 2

In Experiment 2, the proposed selection methods $W-mRMR$, $W-CMIM$ and $W-DISR$ were applied to the LFW, MORPH II and UND databases with no fusion to be able to compare our results with those of [1], [14], [30], [47], [57]. All the images were resized to 64×46 pixels.

The proposed method was tested on 9 different datasets with 300.000 images. The UND dataset has 400 images; LFW-a has 7,443; LFW-b has 13,010; Morph II has 55,134; Adience has 10,632; Groups has 28,231; COFW has 1,007; CelebA has 220,000; and LFWA has 13,223 images. The features were selected in a standalone process. We measured the time invested in each feature selection process on a dataset of 10,000 images. The $W-mRMR$ takes 7.86 hours; W-DISR takes 8.95 hours; and W-$CMIM$ takes 9.97 hours. Feature selection times were measured for a group of 10.000 images in the filter stage and SVM in the wrapper stage. The implementation of the method included a parallel computation for SVM and Random forest classifiers. Computations were measured on an Intel I7 with 8 cores, 64 GB of RAM, and Ubuntu 16.04 S.O.

VII. RESULTS

A. RESULTS OF EXPERIMENT 1

Table 1 shows the results of gender classification on the UND, LFW (set-a, set-b, set-A), MORPH II and CelebA. In Experiment 2, the results are compared to those published previously with various data sets. The results of gender classification without feature selection and with the SVM classifier are shown in the second column of Table 1. Results generally improve with feature selection. The best classification result on UND was reached using the $W-CMIM$ feature selection method; it was 93.25% with 1,300 features selected and 15 nearest images. The total number of features was reduced to 44% of the original vector size. On the LFW the best result reached was 97.00% with $W-CMIM$. The number of selected features was 900 with 25 nearest images. The number of features was reduced to 30.57% of the original vector size. In the case of MORPH II the best result reached was 95.56% with $W-CMIM$. The number of selected features was 1,300 with 25 nearest images. The number of selected features was reduced to 44% of the original vector size.

We also compared our results with those approaches that used deep learning methods to classify gender. See Table 1 and 2. Our approach reached competitive results but with less number of features. LFWA dataset reached an accuracy of 98,33 % and with celebA dataset we reached 98,66% of accuracy, both with only 900 features.

Our method yielded better accuracy in gender classification than the results reported in [6], [14], [30], [35], [47], [55], using LFW database and better results than [14], [68] with MORPH II. Also we improve the accuracy of [31], [43] with celebA and LFWA dataset. Our proposed method requires only 900 features with 25 nearest neighbor images in comparison with 46,845 features from Alexander et al. in the UND

TABLE 1. Gender classification results for the UND, LFW (set-a, set-b), MORPH II, LFWA and CelebA face databases. The first column shows the database. The second column shows the results with no feature selection. Columns 2 and 3 show the results with PCA and LDA dimensionality reduction methods. Columns 5, 6, and 7 show our results with weighted feature selection ($W-mRMR$, $W-CMIM$ and $W-DISR$, respectively). The best number of selected features and p, shown in parenthesis, indicate the best number of neighboring images for the classification rate. In column 8 we show the best results published previously. The best result for each feature selection method is highlighted in bold. PC represents the best number of Principal Components. F represents the best number of Fisherfaces.

Database	Raw Data (%)	PCA(%) PC=500	LDA(%) F=300	W-mRMR (%)	W-DISR (%)	W-CMIM (%)	Previous Results (%)
UND	85.46 +/- 0.89	80.35 +/- 0.65	85.35 +/- 0.55	90.25 +/- 0.90 (1,110)(p=15)	91.25 +/- 0.83 (1,000)(p=10)	93.25 +/-0.74 (1,300)(p=15)	86.78 [1]
LFW set-a	81.27 +/- 1.67	75.20 +/- 0.50	83.20 +/- 0.10	87.79 +/- 0.79 (800) (p=15)	90.93 +/- 0.86 (900)(p=10)	97.00 +/-0.81 (900)(p=25)	95.40 [30] 96.86 [35] 92.60 [55] 98.60 [47]
LFW set-b	78.97 +/- 1.67	73.45 +/- 0.65	81.45 +/- 0.65	85.79 +/- 0.79 (800)(p=15)	88.93 +/- 0.86 (900)(p=10)	95.00 +/- 0.81 (900)(p=25)	95.40 [30] 96.86 [35] 92.60 [55] 93.60 [47]
Morph II	89.07 +/- 0.90	81.45 +/- 0.75	88.45 +/- 0.25	93.15 +/- 0.81 (900) (p=10)	94.25 +/- 0.90 (1,100)(p=15)	95.56 +/-0.89 (1,300)(p=25)	95.20 [28] 88.00 [16]
Celeb A	80.00 +/- 0.90	75.45 +/- 0.55	81.45 +/- 0.15	92.00 +/- 0.90 (850)(p=40)	95.33 +/- 0.86 (1,000)(p=50)	98.66 +/- 0.75 (900)(p=60)	98.00 [43] 98.17 [31]
LFWA	78.00 +/- 0.90	73.25 +/- 0.35	75.25 +/-025	88.33 +/- 0.66 (800)(p=15)	96.66 +/- 0.76 (900)(p=10)	98.33 +/-0.33 (900)(p=25)	94.00 [43] 94.20 [31]

TABLE 2. Results reported in the literature on Deep Learning approaches to gender classification from faces compared to our results (in the last row). Our results show the number of features in parenthesis.

Author	Database	Classifier	Accuracy (%)	Notes
[41]	Adience	CNN	86.40	26K Images 2,284 People Gender and Attributes
[31]	CelebA	MCNN	98.17	220K Images
	LFWA	AUX	94.20	2,284 People Gender and Attributes
[43]	CelebA	CNN	98.00	220K Images
	LFWA		98.17	2,284 People Gender and Attributes
[36]	LFW	CNN	98.90	7,189 Images
			91.34	28,231 Images
[70]	Feret	CNN	95.70	1,351Images 193 people
	Caspel		95.23	1,040 images 208 Images
[19]	MORPH II	CNN	88.20	26,580 Image 2,284 People
	Adience	ELM	87.30	55,000 Images Gender and Age
Our Proposal	UND	W-mRMR W-DISR W-CMIM + SVM and RF With Filtrapper	93.25 - 1,300	Gender
	LFW-a		97.00 - 900	
	LFW-b		95.00 - 900	
	MORPH II		95.56 - 1,300	
	CelebA		**98.66 - 900**	
	LFWA		**98.33 - 900**	
	Adience		92.25 - 900	
	Groups		95.50 - 900	
	COFW		96.50 - 900	

database; 60,000 features used in Jia et al., 57,600 features to build dictionaries of patches from the LFW database. A concatenation of several features from Castrillon et al., called C1, C2, C3, C4 and C5 from images of 59×65 pixels, were used in Groups and Morph datasets and 51,529 features from the images of size 227×227 using a CNN with the Adience Dataset. In summary, our results reached the highest classification performance with a significantly lower number of features.

Table 1 and Table 2 presents a summary of the best results on gender classification reported in the literature for comparisons with the performance of our method with traditional and deep learning methods. Our approach improved the results, reaching at least the same, or even greater accuracy, but with a significantly smaller number of features compared to previously published results, and with nine different datasets. The last two rows of Table 1 and the last row of Table 2 shows the accuracy and the number of features selected by our approach (in parenthesis). The selected number of features by our method represents, on average, only 10% of the total number of features.

We compared our results to those of Alexander et.al. [1], on the UND database. The authors used only a Wrapper method with an SVM classifier and reached 86.78 % accuracy using 7,100 features. The best result was 91.19 % using fusion of intensity, shape, and texture with 46,845 features. In our previous work we improved these results using traditional feature selection methods by reaching 94.01% with 14,200 features. Now our results are better than those reported in Alexandre, 2010 and Tapia et al., 2013 with fewer features. We achieved 98.25% with only 2,100 features using 30 nearest neighbors using the same fusion [57]. And we reached 93.25% using $W-CMIM$ with 1,300 features and 15 nearest neighbors without fusion of features.

We compared our results for the LFW and Morph II databases with Shan et al. [55], Han et al. [30], and Jia et al. [35], Chu et al. [16], Moeni et al. [47] and Castrillon et al. [14]. In these publications the authors used embedded methods, such as regular Adaboost, and Multiclass Adaboost, to select features, and an SVM classifier with 5 fold cross-validation. They used different variations of SVM as classifiers [30], [35], [55]. Moeini et al. use a gender dictionary with 10 fold CV. Castrillon et al. [14] use a LBP and HOG features with SVM. Our results are better than those reported in [16], [30], [35], [55] and very competitive with the state of the art in [14], [47] with a less quantity of features selected. The differences between traditional and

FIGURE 10. Examples of face images from LFW database with feature selected on white pixels. (Male and Female). **Left:** Raw Images without Feature selection. **Middle:** Feature selection used traditional (Pairs) ***CMIM*** **(Best 700 features). Right:** Feature selection used proposal (Groups) $W - CMIM$.

TABLE 3. Gender classification accuracy results for the cross-test using COFW, Adience, and Groups databases. The models were trained with LFW (set-a, set-b) and Morph II datasets. In column 5 we show the best results previously published.

Train Set/Method	COFW All (%)	Adience All (%)	Groups All (%)	Previous Results (%)
LFW (set-a)/ W-mRMR	84.25	81.50	83.00	
LFW (set-a)/ W-DISR	87.25	88.90	91.50	85.92 [14]
LFW (set-a)/ W-CMIM	**95.50**	**90.25**	**94.50**	83.3 [14]
LFW (set-b)/ W-mRMR	86.25	83.50	85.00	84.4 [47]
LFW (set-b)/ W-DISR	90.25	90.90	93.50	80.22 [47]
LFW (set-b)/ W-CMIM	**96.50**	**92.25**	**95.50**	
Morph II / W-mRMR	91.25	84.50	87.90	
Morph II / W-DISR	93.25	**89.00**	91.50	62.04 [14] 72.32 [14]
Morph II / W-CMIM	**94.00**	88.50	**93.25**	67.40 [14]

new proposal features selected over a particular male and female images are presented in Figure 10 and 11.

We also added Table 2 with a summary of the results reported with deep learning applications in gender classification.

B. RESULTS OF EXPERIMENT 2

An additional experiment was performed to validate our best results on three different validation sets with face images from different subjects than those used in Experiment 2. In the first validation test, we used the feature selection method with the best classification rate from Table 1, to test cross-database performance using gender classifiers previously trained with the training set of LFW (set-a, set-b, set-A), Morph II and CelebA. We tested on three different face databases: COFW [10], Adience [20] and Image of Groups (Groups) [25]. These databases show a diverse range of operational conditions and are representative of the challenges to be addressed in real-world situations.

The analysis of the best results from Table 1, using four classifiers trained with the LFW and Morph II databases respectively, are presented in Table 3, to show that better results can be obtained with the features selected by our proposed method. These results are better than those previously published in the state of the art using cross-evaluation datasets allowing to test the generalization performance with other databases.

TABLE 4. Comparison of the best results from Table 1, using SVM and Random Forest classifiers (R-Forest). We also report the parameters of the SVM classifier for the best results.

Methods	W-mRMR		W-DISR		W-CMIM	
	SVM	R-Forest	SVM	R-Forest	SVM	R-Forest
Best Method Table 1	93.15% 900 p=10 C=70.24 $\gamma = 0.027$	94.00% Tree= 500	94.25 1,100 p=15 C=70.51 $\gamma = 0.029$	93.75% Tree=500	**97.00%** 900 p=25 C=71.99 $\gamma = 0.0311$	**96.75%** Tree=500

The localization of features for gender classification depends largely on the most frequent type of face images used during the feature selection process which uses hundreds of thousands of images from the feature selection databases. The selected clusters represent the best features that allow separation and maximize the distance between both classes (male and female).

In the second validation test, the analysis of the best results from Table 1, using a Random Forest classifier [7] with 500 trees is presented in Table 4, to show that similar results can be obtained by a different classifier with the features selected by our proposed method. In Table 4, we compare the best gender classification results using the SVM with feature selection and Random Forest.

VIII. TIME COMPLEXITY

It is important to emphasize that the feature selection process is performed offline only one time, and the selected features are used to classify gender online. Table 5, shows the results in seconds spent by the four feature selection methods used in this research. The reported time is an average of all of the extracted features. The time was computed using an Intel I7-6500U of the 6th generation 2.7 GHz, with 32 GB RAM and an Ubuntu 16.04 operating system. The feature selection methods were implemented in Python. A C++ implementation in a parallel processing architecture which should reduce the computational time significantly.

The computational time of the $mRMR - W$, $DISR - W$ and $CMIM - W$ algorithm depends on the estimation of MI. The estimation of the MI is $O(N \times logN)$. Therefore, the time complexity is $O(2 \times M \times N \times logN)$, where 2 is the number of classes of the vector (Male, Female); M is the number of features in the set; and N is the number of image samples available.

FIGURE 11. Examples of groups of feature selected from face images from Adience database. *W − DISR* (top row) and *W − CMIM* (bottom row) from 100, 200, 300, 500 and 700 groups of features.

TABLE 5. Shows the average time spent for each feature selection and reduction method to select the best 300 features, 300 principal components (PCA) and Fisherfaces (LDA) in the standalone process from 10.000 images, using a Python implementation. The time was computed using an Intel I7-6500U of 6th generation,2.7 GHz, with 32 GB RAM and an Ubuntu 16.04 operating system. Feature selection is performed offline only one time.

Feature Selection and Reduction Method	Time (hrs.)	Processing by	Classification Time per image (ms)
PCA	1	CPU	0.0130
LDA	1.5	CPU	0.0145
mRMR.W	7.86	CPU	0.090
DISR-W	8.95	CPU	0.090
CMIM-W	9.97	CPU	0.0100
CNN-VGG16	8.0 500 Epochs	GPU Average Training Time	0.0115

IX. CONCLUSIONS

In this paper, a new framework for gender classification using an efficient fusion of a filter/wrapper and feature selection strategy based on clustering of images was proposed. Our method decouples relevance and redundancy and uses complementary information in the classification task. The proposed method eliminates those features that are not relevant for the classification problem in the filter stage, thus improving classification accuracy, reducing the required computational time, and making gender classification feasible in real time.

The advantages of this approach are that features that become redundant because of the current content of the selected cluster will go to the end of the new ranking, while features that become relevant because of interactions with the content of the currently selected cluster will be placed at the beginning of the new ranking.

The results of Experiments 1 and 2 show the importance of using the complementarity feature selection on high dimensional data to improve classification performance. In all of the experiments, the proposed method showed its efficiency and effectiveness in feature selection in supervised learning in domains in which data has many redundant and/or irrelevant features. We validated the results using three different databases with very challenging conditions of the face images taken under real conditions using cell-phone photos with occlusion and several variations of pose and changing illumination. Similar results were reached using a pre-trained SVM and RF classifiers. This shows the general purpose of our proposed methods.

This proposal has an advantage over the number of parameters required by Deep Learning (DL) [71]. DL has been shown to be a powerful tool for identifying and classifying gender from face images, reaching results very close to the best state of the art. Nevertheless, DL cannot explain the results in terms of selected features and the relevance or redundancy of each one. Also, DL cannot explore the complementary properties among features because DL does not estimate any measure of the quality of the information, but mutual information does. One of the positive aspects of DL is that powerful, previously pre-trained networks with a huge number of images can be adjusted to a new problem within relatively short training periods. Another positive aspect is that the pre-trained network performs feature selection with DL automatically, and is adjusted by the fine-tuning process, saving time in this manner.

X. FUTURE WORK

As future work, we are developing a new mesh method to align the best features selected with different face poses in real time. This method will allow us to evaluate the efficiency of our proposal in the video. Also, a deeper analysis of African-Americans will be developed in order to measure the influence of the selected features on face images of people of various races or ethnicities.

REFERENCES

[1] L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1422–1427, 2010.

[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, CMU, 2016.

[3] Y. Andreu, P. García-Sevilla, and R. A. Mollineda, "Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes," *Image Vis. Comput.*, vol. 32, no. 1, pp. 27–36, 2014.
[4] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, "Fusion of domain-specific and trainable features for gender recognition from face images," *IEEE Access*, vol. 6, pp. 24171–24183, 2018.
[5] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[6] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Revisiting linear discriminant techniques in gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 858–864, Apr. 2011.
[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[8] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 27–66, Jan. 2012.
[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, vol. 81, S. A. Friedler and C. Wilson, Eds., New York, NY, USA, Feb. 2018, pp. 77–91.
[10] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA, Dec. 2013, pp. 1513–1520.
[11] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection Filter–Wrapper based on low quality data," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6241–6252, 2013.
[12] L. A. Cament, L. E. Castillo, J. P. Perez, F. J. Galdames, and C. A. Perez, "Fusion of local normalization and Gabor entropy weighted features for face identification," *Pattern Recognit.*, vol. 47, no. 2, pp. 568–577, 2014.
[13] L. A. Cament, F. J. Galdames, K. W. Bowyer, and C. A. Perez, "Face recognition under pose variation with local Gabor features enhanced by active shape and statistical models," *Pattern Recognit.*, vol. 48, no. 11, pp. 3371–3384, 2015.
[14] M. Castrillón-Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild," *Image Vis. Comput.*, vol. 57, pp. 15–24, Jan. 2016.
[15] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, Mar. 2012.
[16] W.-S. Chu, C.-R. Huang, and C.-S. Chen, "Identifying gender from unaligned facial images by set classification," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 2636–2639.
[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proc. IEEE Bioinformatics Conf. (CSB)*, Aug. 2003, pp. 523–528.
[19] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning CNN–ELM for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, Jan. 2018.
[20] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.
[21] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
[22] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.
[23] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Proc. Int. Conf. Audio-Video-Based Biometric Person Authentication*, 2003, pp. 44–51.
[24] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, no. 2, pp. 1134–1140, Feb. 1986.
[25] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. CVPR*, 2009, pp. 256–263.
[26] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, "Group sparse representation based classification for multi-feature multimodal biometrics," *Inf. Fusion*, vol. 32, pp. 3–12, Nov. 2016.
[27] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.
[28] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 761–770, 2014.
[29] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction, Foundations and Applications* (Studies in Fuzziness and Soft Computing). New York, NY, USA: Springer-Verlag, 2006.
[30] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. Machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.
[31] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4068–4074.
[32] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8144–8150, 2011.
[33] D. Huang, H. Ding, C. Wang, Y. Wang, G. Zhang, and L. Chen, "Local circular patterns for multi-modal facial gender and ethnicity classification," *Image Vis. Comput.*, vol. 32, no. 12, pp. 1181–1193, 2014.
[34] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[35] S. Jia and N. Cristianini, "Learning to classify gender from four million images," *Pattern Recognit. Lett.*, vol. 58, p. 35– 41, Jun. 2015.
[36] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Gender classification by deep learning on millions of weakly labelled images," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 462–467.
[37] S. A. Khan, M. Nazir, S. Akram, and N. Riaz, "Gender classification using image processing techniques: A survey," in *Proc. IEEE 14th Int. Multitopic Conf. (INMIC)*, Dec. 2011, pp. 25–30.
[38] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," *Machine Learning*. Berlin, Germany: Springer-Verlag, 1994, pp. 171–182.
[39] I. Kononenko and I. Bratko, "Information-based evaluation criterion for classifier's performance," *Mach. Learn.*, vol. 6, no. 1, pp. 67–80, Jan. 1991.
[40] L. Leng, J. Zhang, M. K. Khan, X. Chen, and K. Alghathbar, "Dynamic weighted discrimination power analysis: A novel approach for face and palmprint recognition in DCT domain," *Int. J. Phys. Sci.*, vol. 5, no. 17, pp. 2543–2554, 2010.
[41] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) workshops*, Jun. 2015, pp. 34–42.
[42] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Trans. on*, vol. 45, no. 6, pp. 1209–1221, Jun. 2015.
[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738. doi: 10.1109/ICCV.2015.425.
[44] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 541–547, Mar. 2008.
[45] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 261–274, Jun. 2008.
[46] D. Mladenić, "Feature selection for dimensionality reduction," in *Subspace, Latent Structure and Feature Selection*, (Lecture Notes in Computer Science), vol. 3940, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer, 2006, pp. 84–102.
[47] H. Moeini and S. Mozaffari, "Gender dictionary learning for gender classification," *J. Vis. Commun. Image Represent.*, vol. 42, pp. 1–13, Jan. 2017.
[48] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[49] C. A. Perez, L. A. Cament, and L. E. Castillo, "Methodological improvement on local Gabor face recognition based on feature selection and enhanced Borda count," *Pattern Recognit.*, vol. 44, no. 4, pp. 951–963, Apr. 2011.
[50] C. Perez, J. Tapia, P. Estevez, and C. Held, "Gender classification from face images using mutual information and feature fusion," *Int. J. Optomechatronics*, vol. 6, no. 1, pp. 92–119, 2012.

[51] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, R. Chellappa, D. White, and A. J. O'Toole, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proc. Nat. Academic Sci. USA*, vol. 115, no. 24, pp. 6171–6176, 2018.
[52] K. Ricanek, Jr., and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Southampton, U.K., 2006, pp. 341–345.
[53] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.
[54] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognit.*, vol. 35, no. 4, p. 835– 846, 2002.
[55] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognit. Lett.*, vol. 33, pp. 431–437, Mar. 2012.
[56] X. B. Song, Y. Abu-Mostafa, J. Sill, H. Kasdan, and M. Pavel, "Robust image recognition by fusion of contextual information," *Inf. Fusion*, vol. 3, no. 4, pp. 277–287, 2002.
[57] J. E. Tapia and C. A. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of LBP, intensity, and shape," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 488–499, Mar. 2013.
[58] J. E. Tapia and C. A. Perez, "Gender classification using one half face and feature selection based on mutual information," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Manchester, U.K., Oct. 2013, pp. 3282–3287.
[59] J. Tapia, C. Perez, and K. Bowyer, "Gender classification from the same iris code used for recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1760–1770, Aug. 2016.
[60] Q. Tian and S. Chen, "Joint gender classification and age estimation by nearly orthogonalizing their semantic spaces," *Image Vis. Comput.*, vol. 69, pp. 9–21, Jan. 2018.
[61] P.-W. Tsai, M. K. Khan, J.-S. Pan, and B.-Y. Liao, "Interactive artificial bee colony supported passive continuous authentication system," *IEEE Syst. J.*, vol. 8, no. 2, pp. 395–405, Jun. 2014.
[62] Ö. Uncu and I. Türkşen, "A novel feature selection approach: Combining feature wrappers and filters," *Inf. Sci.*, vol. 177, no. 2, pp. 449–466, 2007.
[63] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.
[64] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," Apr. 2010, *arXiv:1004.2515*. [Online]. Available: https://arxiv.org/abs/1004.2515
[65] J. Wu, W. A. P. Smith, and E. R. Hancock, "Facial gender classification using shape-from-shading," *Image Vis. Comput.*, vol. 28, no. 6, pp. 1039–1048, Jun. 2010.
[66] M.-H. Yang and B. Moghaddam, "Gender classification using support vector machines," in *Proc. Intell. Image Process. Conf.*, vol. 2, 2000, pp. 471–474.
[67] K. Ye, K. A. Feenstra, J. Heringa, A. P. Ijzerman, and E. Marchiori, "Multi-RELIEF: A method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting," *Bioinformatics*, vol. 24, no. 1, pp. 18–25, Jan. 2008.
[68] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 808–812, Jun. 2018.
[69] A. Zafra, M. Pechenizkiy, and S. Ventura, "ReliefF-MI: An extension of ReliefF to multiple instance learning," *Neurocomputing*, vol. 75, no. 1, pp. 210–218, 2012.
[70] H. Zhang, Q. Zhu, and X. Jia, "An effective method for gender classification with convolutional neural networks," *Algorithms and Architectures for Parallel Processing*. Cham, Switzerland: Springer, 2015, pp. 78–91.
[71] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep RoR architecture," *IEEE Access*, vol. 5, pp. 22492–22503, 2017.
[72] Z. Zhang and E. R. Hancock, "Hypergraph based information-theoretic feature selection," *Pattern Recognit. Lett.*, vol. 33, no. 15, pp. 1991–1999, 2012.
[73] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

JUAN E. TAPIA received the P.E. title in electronics engineering from Universidad Mayor, in 2004, and the M.S. degree in electrical engineering from Universidad de Chile, in 2012. He received the Ph.D. Degree from Department of Electrical Engineering, Universidad de Chile, in 2016. Also spent one year of Internship with the University of Notre Dame. His current interests include pattern recognition and machine learning applied to soft-biometrics, gender classification, feature fusion, and feature selection.

CLAUDIO A. PEREZ received the B.S. and P.E. title in electrical engineering, and the M.S. degree in biomedical engineering, all from Universidad de Chile, in 1980 and 1985, respectively. He was a Fulbright student with the Ohio State University, where he obtained a Presidential Fellow, in 1990, and received the Ph.D. degree, in 1991. He was a Visiting Scholar with UC, Berkeley, in 2002 through the Alumni Initiatives Award Program from Fulbright Foundation. He is a Professor with the Department of Electrical Engineering, Universidad de Chile. He was the Department Chairman from 2003 to 2006, and Director of the Office of Academic and Research Affairs with the School of Engineering, Universidad de Chile, from 2014 to 2018. His research interests include biometrics, image processing applications, and pattern recognition. He is a Senior Member of the IEEE, Systems, Man and Cybernetics and the IEEE-CIS societies.