Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information From Histogram of LBP, Intensity, and Shape

Juan E. Tapia, *Graduate Student Member, IEEE*, and Claudio A. Perez, *Senior Member, IEEE*

***Abstract*—In this paper, we report our extension of the use of feature selection based on mutual information and feature fusion to improve gender classification of face images. We compare the results of fusing three groups of features, three spatial scales, and four different mutual information measures to select features. We also showed improved results by fusion of LBP features with different radii and spatial scales, and the selection of features using mutual information. As measures of mutual information we use minimum redundancy and maximal relevance (mRMR), normalized mutual information feature selection (NMIFS), conditional mutual information feature selection (CMIFS), and conditional mutual information maximization (CMIM). We tested the results on four databases: FERET and UND, under controlled conditions, the LFW database under unconstrained scenarios, and AR for occlusions. It is shown that selection of features together with fusion of LBP features significantly improved gender classification accuracy compared to previously published results. We also show a significant reduction in processing time because of the feature selection, which makes real-time applications of gender classification feasible.**

***Index Terms*—Feature fusion, feature selection, gender classification, local binary patterns, mutual information.**

I. Introduction

HUMAN faces provide crucial information regarding gender, age, and ethnicity, in addition to identity. Several important fields for applications of gender classification have been identified, such as biometric authentication, surveillance and security systems, demographic information collection, marketing research, real time electronic marketing, criminology, augmented reality, and lately, new applications in social networks using face recognition [1]–[4]. Gender classification based on facial images is currently one of the most challenging problems in image analysis research [5].

In image understanding, raw input data often has very high dimensionality and a limited number of samples. In this area, feature selection plays an important role in improving accuracy, efficiency and scalability of the object identification process. Since relevant features are often unknown *a priori* in the real world, irrelevant and redundant features may be introduced to represent the domain [6]. However, using more features implies increasing 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 [7].

As is the case in most image analysis problems, with a limited amount of sample data, irrelevant features may obscure the distributions of the small set of relevant features and confuse the classifier [8]. It has been shown both theoretically and empirically that reducing the number of irrelevant or redundant features significantly increases the learning efficiency of the classifier [6], [8]–[13].

The 2 most popular methods used to reduce the dimensionality in gender classification according to our literature review are: [5], [13] Principal component analysis (PCA) [14] and linear discriminate analysis (LDA) [15]. PCA seeks to find a set of mutually orthogonal basic functions that capture the directions of maximum variance in the data and therefore reduce noise in the data. LDA is used to derive a discriminative transformation which maximizes the between-class scatter while minimizing the within-class scatter [16]. The LBP [17] transformation is used to extract features from facial expression images because of its low computational cost and effective texture description ability [18]. However, as more features are extracted, some of them may become redundant or irrelevant for classification [18]–[20].

Several studies have used feature selection in the face and gesture recognition area. Frank *et al.* [21] proposed automatic pixel selection for optimal facial expression recognition based on PCA Eigenfaces. Choi *et al.* [22] proposed pixel selection for optimal face recognition based on LDA discriminative position (pixels) in face images using eigen-spaces.

Bekios *et al.* [5] revisited and compared various linear classification algorithms using LDA, PCA and ICA. These methods can be very sensitive to illumination variations because they use pixel intensity value directly. Moreover, both PCA and LDA methods inherently assume the second order statistics of Gaussian distributions. This assumption may not be met in the case of real face recognition tasks [13].

Bing Li *et al.* [23], proposed a gender classification framework, that utilizes 6 facial components: forehead, eyes, nose, mouth, hair and clothing. The overall accuracy using a 5 five-fold cross-validation method reached 88.5% and 91.9% on 682 and 2,185 images on 2 databases. In Xu *et al.* [24] a hybrid

Manuscript received August 06, 2012; revised November 22, 2012; accepted January 12, 2013. Date of publication January 23, 2013; date of current version February 14, 2013. This work was supported by FONDECYT under Grant 1120613, by FONDEF under Grant D08I-1060, and by the Department of Electrical Engineering, Universidad de Chile. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jaihie Kim.

The authors are with the Department of Electrical Engineering, and Advanced Mining Technology Center, Universidad de Chile, Santiago 8370451, Chile, (e-mail: jtapiafarias@ing.uchile.cl; clperez@ing.uchile.cl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2013.2242063

method using local features (10 features extracted with an Active Appearance Model) and global features extracted with Adaboost was proposed. The authors showed that better accuracy can be obtained by fusing these features before classification. The overall accuracy using 5 five-fold cross-validation on 1,000 images was 92.38%. In both methods the features had large variability and random selection. Recently in [25] we proposed a method for feature selection based on information theory using 3 different mutual information measures with good preliminary results in gender classification.

Battiti *et al.* [7] defined the feature reduction problem as the process of selecting the most relevant k features from an initial set of n features and proposed a greedy selection method to solve it. Ideally, the problem can be solved by maximizing $MI(C, S)$, the joint MI (Mutual Information) between the class C and the subset of the selected features S. However, computing Shannon's MI between high dimensional vectors is impractical because the number of samples and the processing time required become prohibitive. To overcome this problem Battiti adopted an heuristic criterion for approximating the ideal solution. Instead of calculating the joint MI between the selected feature set and the class variable, only $MI(C; f_i)$ and $MI(f_i; f_j)$ were computed, where f_i and f_j are individual features. Battiti's mutual information feature selector (MIFS) selects the feature that maximizes the information about the class, corrected by subtracting a quantity proportional to the MI with the previously selected features. Since feature synergy was not considered, MIFS and its variants estimated the feature redundancy without regard to the corresponding classification task. A complete literature review and comparison among best gender classification methods was reported in Makinen and Raisamo [26], [27].

Several classifiers have been used in gender classification after feature extraction and selection. The classifiers that have yielded highest gender classification accuracy were Adaboost, multilayer neural network (NN), RBF networks and Support vector machines (SVM) [28]. Moghaddam and Yang [29] first reported the SVM with the Radial Basic Function kernel $(\mathrm{SVM} + \mathrm{RBF})$ as the best gender classifier. More recently, Makinen and Raisamo [26] compared the performance of SVM with other classifiers including NN [30] and Adaboost [31]. According to their published results, SVM achieved the highest performance.

In [19], [20] Zhang *et al.* Applied principal geodesic analysis (PGA) on 2.5 facial images extracted from Max-Planck database. This data base contains 100 female and 100 male images obtained by laser scanned human head without hair. Each facial needle-map is represented by a parameter vector, referred to as PGA feature vector. It is not possible to compare ours result to those [19], [20] since in this work we use only 2-D images.

In this paper, we focus on fusion and feature selection methods based on mutual information MI as a measure of relevance and redundancy among features, applied to gender classification. We present 2 approaches for gender classification that improve previously published results on the FERET [32], UND [33] and LFW [34] dataset described in Makinen and Raisamo [27], Perez *et al.* [25], Alexandre [35] and Shan [36]. We also determine accuracy in gender classification comparing information fusion from different spatial scales, with information fusion from different feature types on a single scale.

In Experiment 1, we use 3 different types of face features to classify gender. We extract intensity, shape and texture features using 3 different spatial scales. For the spatial scales we used the same ones used in Alexandre [35] 20×20, 36×36 and 128×128 for the FERET database and the UND database. In Experiment 2, an approach to gender classification based on histograms of Uniform LBP features (LBPH) using a different radii (1 to 8), 3 fusion scales and feature selection with MI was proposed. We also tested our method using the AR [37] face database to test robustness to occlusions (sunglasses and scarves).

II. Information Theory Feature Selection

In this section we introduce briefly some basic concepts and notions from information theory that are used in the 4 feature selection methods used in our study. Information theory provides an intuitive tool for measuring the uncertainly of random variables and the information shared by them, in which entropy and mutual information are 2 critical concepts.

A. Mutual Information (MI)

Entropy H is a measure of the uncertainly of random variables. Let X be (or represent) a discrete random variable. The entropy of X is defined as:

$$H(X) = -\sum_{x \in X} p(x) log(p(x)). \tag{1}$$

The mutual information, 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:

$$MI(x, y) = \sum_{i,j} p(x_i, y_j) log \tfrac{p(x_i, y_j)}{p(x_i)p(y_j)}. \tag{2}$$

We use mutual information to measure the level of "similarity" between pixels. The concept of minimal redundancy, as in (3), allows selection of pixel pairs that are maximally dissimilar. When 2 features are highly dependent on each other, the respective class-discriminating power would not change much if one of them were to be removed. Therefore, the following minimum redundancy $(minW_I)$ condition can be added for selecting mutually exclusive features.

$$minW_I, W_I = \frac{1}{\mid S \mid^2} \sum_{f_i, f_s \in S} MI(f_i; f_s) \tag{3}$$

where S denotes the feature subset, $\mid S \mid$ is the number of features in S, and $MI(f_i; f_j)$ is used to represent the mutual information between f_i and f_j. $MI(C; f_i)$ represent the mutual information between features f_i and the class C. In this work the classes represent the gender for each image. Thus, $MI(C; f_i)$ quantifies the relevance of features f_{ith} for the classification task. Therefore, since the maximum relevance condition maximizes the total relevance of all features in S, maximal relevance $(maxV_I)$ is used to search for features that approximate the mean value of all mutual information values between individual features f_i and class C.

$$maxV_I, V_I = \frac{1}{|S|}\sum_{f_i \in S} MI(C; f_i). \tag{4}$$

The first feature is selected according to V_I, for example, the feature with the highest $MI(C; f_i)$. Subsequent features are selected incrementally in the feature set S. If m features are already selected from S, and an additional feature is selected from $\Omega_s = \Omega - S$, then the two conditions are optimized: the min operation (3) is interpreted as a minimum redundancy computation; the max operation is interpreted as the maximum relevance (4).

1) Minimum Redundancy and Maximal Relevance (mRMR): Two forms of combining relevance and redundancy operations are used in [8], mutual information difference (MID), and mutual information quotient (MIQ), thus the $mRMR$ feature set is obtained by optimizing MID and MIQ simultaneously. Optimization of both conditions requires combining them into a single criterion function [8], [12] as:

$$f^{mRMR}(X_i) = MI(C; f_i) - \frac{1}{|S|} MI(f_i; f_s) \tag{5}$$

where, $MI(C; f_i)$ measures the relevance of the feature to be added for the output class and the term $(1/|S|)\sum_{f_i \in S} MI(f_i; f_s)$ estimates the redundancy of the f_{ith} feature with respect to the subset of previously selected features S.

2) Normalized Mutual Information Feature Selection (NMIFS): Estevez *et al.* [9] proposed an improved version of mRMR based on the normalized feature of mutual information; the MI between 2 random variables is bounded above by the minimum of their entropies. As the entropy of a feature could vary greatly, this measure must be normalized before applying it to a global set of features as

$$f^{NMIFS}(X_i) = MI(C; f_i) - \frac{1}{|S|}\sum_{f_i \in S} MI_N(f_i; f_s) \tag{6}$$

where, MI_N is the normalized MI by the minimum entropy of both features, as defined in

$$MI_N(f_i; f_s) = \frac{MI(f_i; f_s)}{min(H(f_i), H(f_s))}. \tag{7}$$

3) Conditional Mutual Information Feature Selection (CMIFS): In CMIFS [38], the feature S subset is built up step by step, by adding one feature at a time. It will not waste time on unnecessary features by removing classification redundancy features beforehand, and can detect both cooperation and redundancy interaction of features (synergy). In addition, CMIFS allows determination of the feature redundancy and this information can be used to remove features improving classification tasks. It can decrease the probability of mistaking important features as redundant features in the searching process. Let S be the set of already-selected features, and Ω the set of candidate features, $S \cap \Omega = \emptyset$ and C is the class. The next feature in Ω to be selected is the one that makes $MI(C; f_i, X_s)$ maximum, where $f_i \in \Omega$ and

$$\begin{aligned} MI(C; f_i, X_s) = MI(C; f_i) \\ - [MI(f_i; X_s) - MI(f_i; X_s \mid C)]. \end{aligned} \tag{8}$$

4) Conditional Mutual Information Maximization (CMIM): The CMIM [10], [39] approximates the relevance criterion, by considering the MI between the candidate feature variable f_i and the class C given each one of the variables in the set S, separately. It allows preserving a certain trade-off between the power prediction of f_i with respect to the output and the independence of candidate features with each single variable previously selected. CMIM considers that feature f_i is relevant only if it provides large amount of information about class C and this information is not contained in any of the variables already selected.

One strategy to find an optimal subset $S \subset F$, is to evaluate all possible subsets in F of cardinality d. However, this process generates a combinatorial explosion of possible solution. To avoid an exhaustive search, a greedy selection begins with the empty set of selected features and successively adds feature one by one. For the first feature selection, set F represents the initial set of m features for S empty set $(S = \emptyset)$. After the first iteration the set will not be empty set $(S \neq \emptyset)$.

$$\begin{aligned} &CMIM \\ &= \begin{cases} argmax_{f_i \in F}\{MI(f_i; C)\} & \text{for } S = \emptyset \\ argmax_{f_i \in F/S}\{min_{f_j \in S} MI(f_i; C/f_j)\} & \text{for } S \neq \emptyset. \end{cases} \end{aligned} \tag{9}$$

III. Databases, Feature Extraction and Fusion

A. Dataset Experiment 1

Two internationally available face databases were used to train and test the fusion and the MI feature selection methods and to allow comparison of results with those previously published [35]. The FERET database [32] contains gray scale images of 1,199 individuals with uniform illumination but with different poses. As in Makinen and Raisamo [27], faces of one image per person from the Fa and Fb subsets were used and duplications were eliminated. Therefore, 199 female and 212 male images were used from the FERET database.

The second database was composed of UND images; more specifically a set of images from Collection B (see Fig. 1). The image filenames used for training and testing, and also the window crop around the subjects faces are available as text files on a web page as reported in [35]. It contains gray scale images of 487 frontal face images with 186 female and 301 male images, collected and annotated by the researchers. To compare our results with those in [35] 3 image sizes were used: 20 × 20, 36 × 36 and 128 × 128.

1) Feature Extraction and Fusion for Experiment 1: We used 3 different types of face features to classify gender. We extracted intensity, shape and texture features using 3 different spatial scales.

Fig. 1. Examples of face images under unconstrained scenarios from the LFW database (top two rows). Face images under controlled scenarios from the UND database (bottom two rows).

The intensity feature for each pixel is the gray level of each pixel. The shape feature is extracted from the edges histogram. Vertical and horizontal edge maps were computed using the masks $[-1, 0, 1]$ and $[-1, 0, 1]^T$. Consider v and h to be the vertical and horizontal edge values at any pixel, obtained by convolution of the edge mask with the original image, respectively. The edge map is found using $\theta = \tan^{-1}((v/h))$ and the edge magnitude is given by $m = \sqrt{v^2 + h^2}$. The edge map is discretized at 18 degree intervals. Each pixel adds its magnitude m to the bin that corresponds to its edge directions θ. For N image windows, an image is represented by 20 $\times N$ real values. Since there are 6 possible variants for the shape and texture features at 128 × 128 and 36 × 36, and 3 possible variants at 20 × 20, given the different types of windows used in [35], we chose only the best case for each image size. In all cases we chose to use the variants with 50% overlay. For the 128 × 128, images the window size is 16 × 16; for the 36 × 36 images the window size is the 6 × 6, and for the 20 × 20 images the windows size is 10 × 10.

For the texture feature we used the local binary patterns (LBP) transformation. LBP is a gray-scale texture operator which characterizes the spatial structure of the local image texture. Given a central pixel in the image, a binary pattern number is computed by comparing its value with those of its neighbors. The original operator used a 3 × 3 window size containing 9 values. Other LBP operators were generated by changing the window size. LBP features were computed from pixel intensities in a neighborhood.

$$LBP_{P,R}(x,y) = \bigcup_{(x',y') \in N(x,y)} h(I(x,y), I(x',y')) \qquad (10)$$

where $N(x,y)$ is the vicinity around (x,y), $\cup$ is the concatenation operator, P is the number of neighbors, and R is the radius of the neighborhood.

LBP was first introduced in [17] showing high discriminative power in distinguishing texture features, and is widely used for face analysis. Later, in [18], [40] the uniform local binary pattern (ULBP) was introduced, extending the original LBP operator to a circular neighborhood with a different radius size and a small subset of LBP patterns selected. In this work we use, ‘U2’ which refers to a uniform pattern. LBP is called uniform when it contains at most 2 transitions from 0 to 1 or 1 to 0, which is considered to be a circular code. Thus the number of patterns is reduced to 59 bins (see Fig. 2).

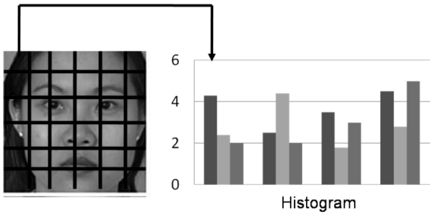

Fig. 2. Face image, divided into subregions with the corresponding concatenated LBP histogram.

We propose an effective feature selection method to obtain a reduced set of LBP features, using mutual information between features and class labels [25]. LBP features are extracted from all training and testing images. Then, the LBP features are organized in a matrix of $D \times N$ size, $F_{LBP} = \{f_1, f_2, \ldots f_N\}$ where f_i is a D dimensional LBP feature vector at the i_{th} pixel position. The mutual information $MI(C; f_i)$ is computed between the class C and the feature vector f_i for $i = 1, 2, \ldots N$ and obtain the selected feature index set $S_{LBP} = \{p_1, p_2, \ldots p_M\}$ by applying different feature selection methods (mRMR, NMIFS, CMIFS, CMIM), where M is the number of LBP features vectors, and p_i the denotes the index of the selected LBP feature vector at the i_{th} pixel position. Also, the LBP features with radii 1–8 may represent redundant patterns and therefore, feature selection by mutual information allows the selection of most relevant features.

For Experiment 1, as in [35], the face image was divided into N overlapping blocks, and the LBP operator was applied to each block using 8-connected neighbors and a radius of one. Then, a histogram with 59 bins was created for each block. The histograms were concatenated and the best features were selected using mRMR, NMIFS, CMIFS and CMIM in the ranges 50–400 for image size 20 × 20, 50–1,296 for size 36 × 36, and 50–16,384 for size 128 × 128. For N image windows, an image is represented by 20 $\times N$ real values. After feature extraction, we fused that information at the feature level by concatenating the feature vectors from different sources into a single feature vector that becomes the input to the feature selection methods, and then the selected features become the inputs to the classifier. The classifiers are trained both with the selected features for each feature extraction method, and with the fused selected features.

Fig. 3 shows the 7 combinations of features and spatial scales we tested in Experiment 1. Fig. 3 shows that L1, L2 and L3 were obtained from vertical fusion of features at different spatial scales (but with the same type of feature) while L4, L5 and L6 show the horizontal fusion of features for different feature types, but on the same spatial scale. Combination L7 includes all scales and all features, and the features were selected with mutual information methods. For each case we chose windows with 50% overlap.

The best gender classification accuracy based on shape features was published in [35], showing 96.26% accuracy in the FERET database and 86.78% in the UND database. The best results using the same size but different features yielded 95.33% correct classification on the FERET database for 128 × 128 size images, and 80.62% using 36 × 36 size images with the UND database. Fusing the 3 types of features (intensity, shape and

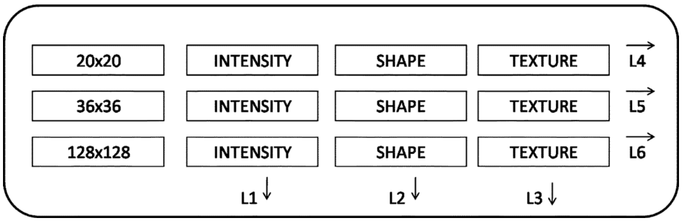

Fig. 3. Representation of the possible combinations of the three feature types (intensities, shape, and texture) and the three different spatial scales (20 × 20, 36 × 36, and 128 × 128) for Experiment 1.

texture) and 3 sizes of images (20 × 20, 36 × 36, 128 × 128) yielded the best score 99.07% on the FERET database and 91.19% for the UND database. However, many more inputs for the 20 × 20, 36 × 36 and 128 × 128 scales were used. By fusing the 3 scales and the 3 types of features, the total number of inputs was increased nearly ninefold. The databases were partitioned to have 80% training data and 20% testing data. All results were obtained with fivefold cross-validation, simulations using an SVM classifier with a Gaussian Kernel.

B. Dataset Experiment 2

In this experiment, we used the recently built public database Labeled Faces, in the Wild (LFW) [34] to investigate gender classification of real world face images under unconstrained scenarios. This public database enables future benchmarks and assessment (http://vis-www.cs.umass.edu/lfw/). LFW, a database for studying the problem of unconstrained face recognition, contains 13,233 face color photographs of 5,749 subjects collected from the web. LFW is composed of real life faces, with varying facial expressions, illumination changes, head pose variations, occlusions and use of make-up, including poor image quality. Thus, gender recognition in real life is much more challenging than gender recognition of faces captured in constrained environments. All the images were aligned with commercial software [36], see Fig. 4. Examples are shown in Fig. 1. As in [36], we chose 7,443 face images, 2,943 females and 4,500 males and manually labeled the ground truth for gender of each face. The images in the FERET database and the UND database are of good quality, under controlled conditions, while in the LFW, quality varies significantly. We also compared our results with those reported in [36] where they used 7,443 images of 64 × 46 pixels.

1) Feature Extraction and Fusion For Experiment 2: An approach to gender recognition based on histograms of LBP features (LBPH) with different radii and 3 scales had been proposed [36]. Results using SVM raw pixels with the dimension of 2,944 and standard LBP with the dimension of 2,478 reached 91.27% and 93.38%, respectively. The best gender classification accuracy achieved, applying SVM with boosted multiscale LBP features with 500 selected LBPH bins in LFW databases was 94.81%. However, the total number of inputs increased nearly a hundredfold by using 8 different radii, 3 scales, shifting and scaling steps of 12, 18, and 24 pixels vertically and 10, 15, and 20 horizontally. In this way, additional subwindows, could be obtained from each image $(725 \text{ regions} \times 59 \text{ bins} \times 8 \text{ radii} = 342{,}200)$. The histograms were concatenated and the best features were selected using Adaboost for image size 64 × 46 as in [36]. Each face image can be seen as a composition of micro-patterns which can be described effectively by LBP. The LBP histograms are always extracted from local regions. However, not all bins in the LBP histograms are discriminative for facial representation [25].

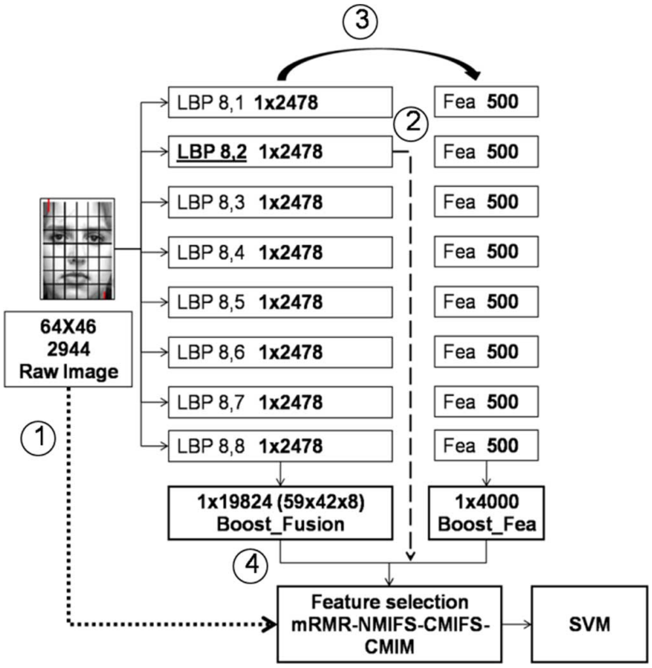

Fig. 4. Diagram showing the fusion of selected LBPH features selection in Experiment 2. The number in circles indicate the test number.

We assessed and investigated the gender recognition performance using LBPH fusion and feature selection methods for different subwindows shifted and scaled separately in steps of 12 pixels vertically and 10 pixels horizontally for the first scale, 18 vertical and 15 horizontal pixels for the second scale and 24 × 20 for the last scale. Finally, fusion was performed among the best results of each feature selection method for the 3 scales.

In experiment 2, we performed four tests for each of the three spatial scales (see Fig. 4): 12 × 10 18 × 15 and 24 × 20. For Test (1) the inputs (raw pixel intensity) to the SVM classifier were selected by the four feature selection methods (mRMR, NMIFS, CMIFS and CMIM). For Test (2) the inputs to the SVM classifier were histograms of LBPH(8, 2) were also selected by the four feature selection methods. For test (3) the inputs to the SVM classifier were a concatenation of 500 features extracted from each of the 8 radii LBP (LBP (8, 1) to LBP (8, 8)). Therefore, the resulting concatenated vector was 4,000 $(500 \text{ features} \times 8 \text{ radii} = 4{,}000 \text{ bins})$ for each face image. Then, feature selection was applied using mRMR, NMIFS, CMIFS and CMIM. For Test (4) the inputs to the SVM classifier were a concatenation of features 8 radii LBP (LBP(8, 1) to LBP (8, 8)) which resulted in a total of 19, 824 features $(42 \text{ regions} \times 59 \text{ bins} \times 8 \text{ radii})$ for each face image. Then, feature selection was applied using mRMR, NMIFS, CMIFS and CMIM.

Because computational time that depends directly on the number of inputs to the classifier is an important factor in most real-time applications involving face processing, we computed the time to compare different methods.

TABLE I
GENDER CLASSIFICATION EXPERIMENTAL RESULTS ON THE FERET DATABASE. THE FIRST COLUMN SHOWS THE TYPE OF FEATURE USED INCLUDING FUSION OF DIFFERENT FEATURE TYPES AND SPATIAL SCALES. THE SECOND COLUMN SHOWS PREVIOUS RESULTS WITH NO FEATURE SELECTION. COLUMNS 3–6 SHOW OUR RESULTS WITH FEATURE SELECTION AND FEATURE FUSION

FUSION	FERET(%) [35]	mRMR(%)	NMIFS(%)	CMIFS(%)	CMIM(%)
Intensity (L1)	95.33 +/- 0.89 (18,000)	92.59 +/- 0.76 (8,200)	93.82 +/- 0.81 (14,700)	93.82 +/- 0.91 (14,800)	95.82 +/- 0.87 (11,700)
Shape (L2)	96.26 +/- 0.88 (7,100)	86.41 +/- 0.78 (800)	81.48 +/- 0.75 (2,050)	89.95 +/- 0.90 (450)	82.71 +/- 0.88 (3,750)
Texture (L3)	93.46 +/- 0.91 (20,945)	**98.76 +/- 0.74** **(1,200)**	**95.06 +/- 0.77** **(6,700)**	**97.53 +/- 0.88** **(750)**	97.53 +/- 0.71 (550)
20x20 (L4)	85.98 +/- 0.77 (1,831)	91.35 +/- 0.81 (850)	91.35 +/- 0.91 (500)	92.59 +/- 0.87 (800)	93.82 +/- 0.86 (750)
36x36 (L5)	91.59 +/- 0.77 (10,855)	93.82 +/- 0.90 (7,650)	93.82 +/- 0.91 (1,300)	95.06 +/- 0.90 (5,650)	95.06 +/- 0.88 (3,800)
128x128(L6)	95.33 +/- 0.81 (34,159)	95.06 +/- 0.66 (1,000)	91.35 +/- 0.79 (9,950)	97.53 +/- 0.70 (11,000)	**97.83 +/- 0.75** **(1,000)**
All (L7)	99.07 +/- 0.95 (46,845)	96.30 +/- 0.73 (19,700)	95.06 +/- 0.77 (35,200)	97.53 +/- 0.75 (33,450)	97.53 +/- 0.71 (21,550)
best fea	**99.13 +/- 0.66** **(18,900)**				

TABLE II
GENDER CLASSIFICATION EXPERIMENTAL RESULTS ON THE UND DATABASE. THE FIRST COLUMN SHOWS THE TYPE OF FEATURE USED INCLUDING FUSION OF DIFFERENT FEATURE TYPES AND SPATIAL SCALES. THE SECOND COLUMN SHOWS PREVIOUS RESULTS WITH NO FEATURE SELECTION. COLUMNS 3–6 SHOW OUR RESULTS WITH FEATURE SELECTION AND FEATURE FUSION

FUSION	UND(%) [35]	mRMR(%)	NMIFS(%)	CMIFS(%)	CMIM(%)
Intensity (L1)	85.46 +/- 0.89 (18,000)	83.77 +/- 0.77 (1,800)	87.28 +/- 0.81 (1,800)	86.48 +/- 0.91 (1,050)	87.82 +/- 0.90 (1,200)
Shape (L2)	84.58 +/- 0.78 (20,945)	72.36 +/- 0.76 (200)	77.19 +/- 0.82 (1,150)	76.56 +/- 0.71 (300)	75.59 +/- 0.88 (2,650)
Texture (L3)	86.78 +/- 0.93 (7,100)	84.21 +/- 0.81 (1,800)	87.28 +/- 0.77 (1,800)	91.01 +/- 0.76 (5,500)	**92.05 +/- 0.81** **(2,200)**
20x20 (L4)	73.57 +/- 1.01 (1,831)	**84.21 +/- 0.81** **(1,500)**	84.21 +/- 0.81 (300)	86.40 +/- 0.77 (300)	86.85 +/- 0.65 (500)
36x36 (L5)	80.62 +/- 0.86 (10,855)	82.07 +/- 0.90 (9,800)	**88.15 +/- 0.78** **(2,150)**	89.47 +/- 0.81 (1,600)	90.35 +/- 0.70 (1,250)
128x128 (L6)	79.30 +/- 0.77 (34,159)	71.05 +/- 0.77 (3,100)	88.15 +/- 0.79 (7,950)	**92.10 +/- 0.84** **(7,900)**	89.47 +/- 0.83 (9,900)
All (L7)	91.19 +/- 0.54 (46,845)	81.14 +/- 0.78 (18,400)	87.72 0.91 (26,250)	92.54 +/- 0.81 (13,000)	88.16 +/- 0.80 (15,300)
best fea	**94.01 +/- 0.54** **(14,200)**				

IV. EXPERIMENTS AND RESULTS

A. Results—Experiment 1

Tables I and II compare our results with those previously published [35] for different image sizes on the FERET and UND databases. Results represent the average of the gender classification performance of 5 cross-validations with a random partition of the database. The first column of Table I shows the method used; the second column shows the results of the best classification rates published in [35] for the SVM classifier using 3 image sizes: 20 × 20, 36 × 36 and 128 × 128; and, in parenthesis, the feature vector size. Columns 3–6 show the results using our proposed feature selection methods: mRMR, NMIFS, CMIFS and CMIM, respectively. Each row shows the average classification rate for 5 simulations, standard deviation and, in parenthesis, the number of selected features for each method.

The results obtained on the FERET and UND databases with our methods are better than those previously published [35]. It should be emphasized that the gender classification results improved significantly, and the number of input features was reduced drastically, which has important implications for real time implementation of the proposed method. This topic is discussed in the computational time section of the paper. The method with the best result (L3) was tested with the LFW database, and the results showed that the fused features reached better result than those previously published [36] (see Table III).

TABLE III
RESULT TO TEST CROSS DATABASE PERFORMANCE FOR GENDER CLASSIFICATION ON THE LFW DATABASE FOR THE BEST RESULTS FROM THE FERET AND UND DATABASES. COLUMNS SHOW THE BEST RESULTS OBTAINED FOR L3 TEST WITH 4 MI MEASURES AND IN PARENTHESIS IS SHOWN THE NUMBER OF FEATURES. THE BEST RESULT WAS REACHED WITH mRMR AND 600 FEATURES

FUSION	LFW [36]	mRMR(%)	NMIFS(%)	CMIFS(%)	CMIM(%)
Best_fea LFW (L3)	94.81 +/- 1.10 (500)	**95.60 +/- 0.45** (600)	92.01 +/- 0.49 (650)	94.84 +/- 0.81 (500)	95.01 +/- 0.54 (400)

As expected, our results show lower classification performance on the UND database compared to those on the FERET because images in the UND database vary in quality, pose, illumination and partial occlusion.

We performed 7 tests, named L1 to L7. L1 is represented by a vector with the fusion of features from pixel intensities from different spatial scales. L2 is represented by a vector with the fusion of features from the shapes from different spatial scales. L3 is represented by a vector with the fusion of textures features (LBP) from different spatial scales. The 3 spatial scales were 20 × 20, 36 × 36 and 128 × 128. L4 is represented by a vector with the fusion of 3 features (intensity, shape and texture) for size 20 × 20. L5, L6 are represented by vectors with the fusion of the same 3 previous features but for sizes 36 × 36 and 128 × 128, respectively. L7 is represented by a vector with the fusion of 3 scales (20 × 20, 36 × 36, and 128 × 128) and 3 types of features (intensity, shape and texture). 2 tests were performed for L7. The first one, considers the concatenation of features for each of the selection methods from L1 to L6 (the concatenation of mRMR for L1 to L6, the concatenation of NMIFS for L1 to L6, concatenation of CMIFS L1 to L6 and the concatenation of CMIM L1 to L6), named "All-L7". The second test, was named "Best Fea", where the fusion of the best methods from L1 to L6 (i.e., those methods that reached the highest scores). In this test, L1 reached 95.8% with 11,700 features, L2 89.9% with 450 features, L3 98.7% with 1,200 features, L4 93.8% with 750 features, L5 95.1% with 3,800 features, and L6 reached 97.8% with 1,000 features. The total of selected features adds 18,900 features.

It can be observed in Table I that for L1 the CMIM feature selection method reached the best classification performance of 95.82% with only 11,700 features which is 60% of the vector size required with no selection. In the case of L2, the best feature selection method was CMIFS achieving 89.95% with only 450 features, which is 50% of the vector size with no selection. For L2 the highest classification rate published in [35] was 96.26%, however, in our simulation of this method we reached 89.95% ± 0.90. In the case of L3 the best method was mRMR achieving 98.76% with only 1,200 features which is only 1.7% of the original vector size. In the case of L4 and L5, the best classification performance was 93.82% with 750 features, and 95.06% with 3,800 features, both with the CMIM feature selection method. The number of selected features corresponded to 41% and 35% of the vector sizes with no selection, respectively. In the case of L6 the best result was 97.83% which was reached with the

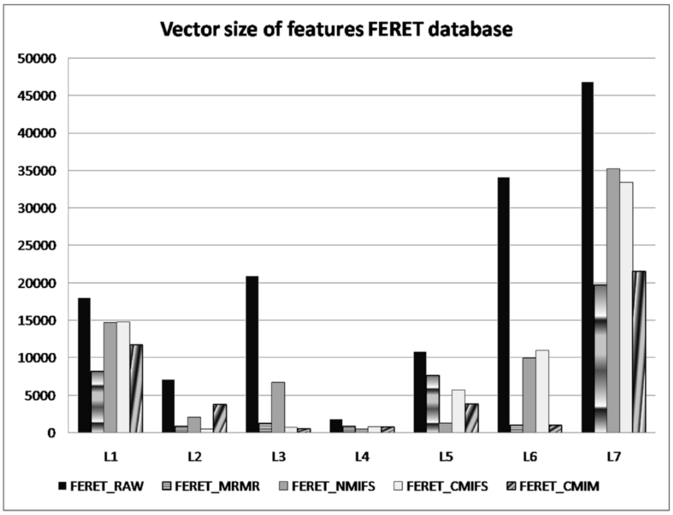

Fig. 5. Comparison of feature vector size for each of the methods from L1 to L7 and the complete set of features used in [35] for the FERET database.

CMIM feature selection method with 1,000 features, 3% of the vector size with no feature selection. The best gender classification rate reached 97.53% with 21,550 features which is only 46% of the vector size with all. The maximum classification rate was 99.13% on the FERET database. The total number of selected features was 18,900 which is 42% of the total number of features.

Fig. 5 shows a histogram comparing the vector sizes for all cases of feature selection and feature fusion (L1–L7), and the case of no feature selection for the FERET database.

Table II shows the results of gender classification with the UND database. In Experiment 2, the best result for L1 using the CMIM feature selection method reached 87.82% with 1,200 features equivalent to 6% of the original vector size with no selection. In the case of L2, L3, L4, L5, and L6 the best results were reached with the NMIFS feature selection, (77.19%), CMIM (92.05%), CMIM (86.85%), CMIM (90.35%), and CMIFS (92.10%), respectively. The number of selected features was: 1,150, 2,200, 500, 2,150, and 7,900, respectively, which were 6%, 29%, 27%, 2% and 23% of the original vector sizes. In "All-L7" the best result was achieved with CMIFS reaching a classification performance of 92.1% with 13,000 selected features which is 27.75% of the vector size with no selection. The classification rate for "Best fea" was 94.01% with 14,200 selected features, 30% of the vector size with no selection.

Table III shows the result of best fusion (L3) for gender classification with LFW database. The result for L3 using mRMR feature selection method reached 95.60% with 600 features equivalent to 2.9% of the 20,950. This result is better than the best one obtained in [36], where gender classification reached only 94.81% with 500 features.

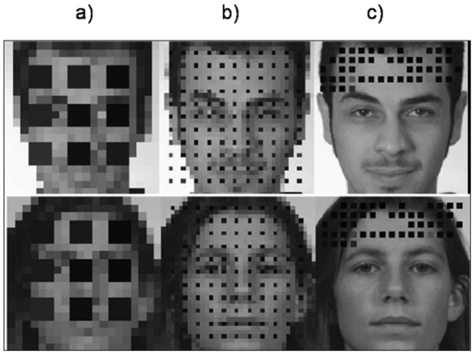

Fig. 6. Two images, male and female, from the FERET database. The squares represent 1,200 selected features from L3 using mRMR which reached the best results for Experiment 1. The fusion considers three scales for image sizes: 20 × 20, 36 × 36, and 128 × 128. The squares' intensities moving towards black represent an increasing number of bins selected in that area.

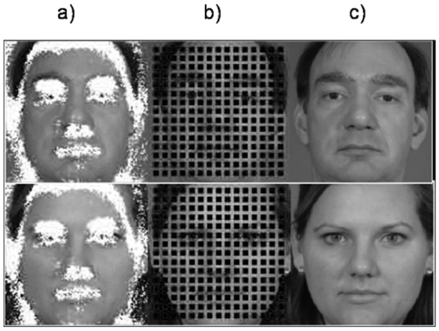

Fig. 7. Two images, male and female, from UND database are shown. Feature fusion L6 with 7,900 features selected using CMIFS achieved the best results for Experiment 1. The fusion considers three features: intensity, shape, and texture for image size of 128 × 128. The figure shows the selected features from (a) intensity images, (b) shape image, and (c) texture image. In this example for L6, none of the texture features were selected.

Fig. 6 shows examples of selected features for the best results obtained for Experiment 1, with the feature selection method mRMR. Two images (male and female) are shown from the FERET database for the L3 feature fusion with 1,200 selected features. The features were selected from the LBP histogram using 3 different image sizes (20 × 20, 36 × 36 and 128 × 128). Each square shows the selected area and the increasing intensity towards black represents the number of bins that were selected in this area. If the area was not selected, no square is shown.

Fig. 7 shows examples of selected features for the best results obtained in Experiment 1 with the feature selection method CMIFS. Fig. 7 shows 2 images (male and female) from the UND database for the L6 feature fusion with 7,900 selected features. The method fused selected features from the intensity histogram, shape and texture (LBP) in 3 images of size 128 × 128. The white pixels in (a) represent the selected pixels from the intensity features; the squares in (b) represent the selected areas from shape features where darker indicates a higher number of selected bins. If the area was not selected, no square appears. Fig. 7(c) shows the best result of feature fusion for the UND database with the CMIFS method using only 7,900 features from the total of 34,159. For this best case, non of the features from Fig. 7(c) were selected. If we increase the number of selected features, features from Fig. 7(c) will be selected, but this increment will not improve the gender classification rate for method L6.

Feature selection is performed using the detected face within a rectangle in sizes 20 × 20, 36 × 36 and 128 × 128. These images have diverse backgrounds and in particular when comparing faces with long hair and short hair these features contribute to differentiate between men and women. The same may occur with different hairstyles that may help to differentiate gender. Also in the feature selection it is important the synergy among different features.

TABLE IV
RESULT OF GENDER CLASSIFICATION RATES ON THE LFW DATABASE, FOR CASES L1–L7 FOR DIFFERENT LBP FEATURES AND SPATIAL SCALE. THE FIRST FOUR ROWS SHOW RESULTS PREVIOUSLY PUBLISHED. COLUMNS 4–7 SHOW RESULTS OF GENDER CLASSIFICATIONS WITH FEATURE SELECTION mRMR, NMIFS, CMIFS, AND CMIM, AND, IN PARENTHESIS, THE TOTAL NUMBER OF SELECTED FEATURES IS SHOWN

Fusion	Vector	LFW (%) [36]	mRMR (%)	NMIFS (%)	CMIFS (%)	CMIM (%)
Raw Pixel	2,944	91.27 +/- 1.67	91.07 +/- 1.01 (600)	92.01 +/- 1.01 (850)	91.70 +/- 0.89 (350)	92.02 +/- 1.01 (400)
LBP (8,2)	2,478	93.38 +/- 1.50	N/A	N/A	N/A	N/A
Boosted Adaboost	342,000	94.40 +/- 0.86 (500)	N/A	N/A	N/A	N/A
Boosted SVM	342,000	94.81 +/- 1.10 (500)	N/A	N/A	N/A	N/A
Raw_Fea 12x10	2,944	N/A	91.62 +/- 1.05 (2,350)	91.55 +/- 1.01 (2,250)	91.09 +/- 0.89 (2,500)	91.84 +/- 0.87 (1,650)
LBP(8,2) 12x10	2,478	N/A	92.65 +/- 0.95 (1,950)	92.25 +/- 0.95 (1,950)	92.58 +/- 1.01 (1,950)	92.68 +/- 0.95 (1,900)
Boost_Fea SVM 12x10	4,000	N/A	92.05 +/- 1.02 (500)	91.07 +/- 0.91 (500)	92.07 +/-1.05 (500)	92.31 +/- 0.92 (500)
Boosted Fusion SVM 12x10	19,824	N/A	92.40 +/-0.89 (1,750)	93.96 +/- 1.01 (5,150)	94.76 +/- 1.05 (3,900)	**96.73 +/- 1.02** **(3,050)**
Raw_Fea 18x15	2,944	N/A	90.42 +/- 0.90 (1,900)	92.36 +/- 1.01 (1,500)	92.24 +/- 0.78 (1,150)	92.36 +/- 1.01 (1,150)
LBP(8,2) 18x15	2,478	N/A	93.90 +/-0.75 (2,000)	94.20+/- 0.96 (1,650)	94.96 +/- 0.90 (800)	95.94 +/- 0.81 (600)
Boost_Fea 18x15	4,000	N/A	92.90 +/- 0.78 (500)	92.20 +/- 0.86 (500)	92.04 +/- 0.96 (450)	91.94 +/- 0.83 (450)
Boosted fusion SVM 18x15	19,824	N/A	94.90 +/- 0.81 (1,850)	93.20 +/- 0.93 (5,650)	**95.94 +/- 0.96** **(3,200)**	95.94 +/- 0.80 (3,600)
Raw Fea 24x20	2,944	N/A	91.71 +/- 1.01 (1,750)	92.16 +/- 0.98 (1,900)	92.48 +/- 0.90 (1,500)	92.85 +/- 0.79 (1,450)
LBP (8,2) 24x20	2,478	N/A	93.96 +/- 0.96 (1,900)	94.57 +/- 0.89 (1,700)	94.89 +/- 1.01 (600)	95.89 +/- 0.98 (1,150)
Boosted Fea SMV 24x20	4,000	N/A	93.06 +/- 0.76 (500)	93.67 +/- 0.83 (500)	92.19 +/- 1.01 (500)	94.79 +/- 0.88 (500)
Boosted Fusion SVM 24x20	19,824	N/A	94.96 +/- 0.96 (1,900)	95.57 +/- 0.79 (6,700)	95.89 +/- 0.81 (3,600)	**96.89 +/- 0.98** **(4,150)**
Fusion Best	19,824	N/A	**98.01 +/- 0.95** **(10,400)**			

B. Results—Experiment 2

In the first 4 rows Table IV shows the gender classification results previously published for different image sizes on the LFW database. In rows 5–17 Table IV shows our results with feature selection based on MI and SVM classifiers (Experiment 2). Results represent the average of 5 cross-validations with a random partition maintaining the database ratio between male and female. The first column shows the method used, and the second column shows the vector size. The third column shows the best classification rates published in [36] for SVM classifiers using LBP features with an image size of 64 × 46 pixels, and, in parenthesis, the number of selected features. Columns 4–7 show the results of the same classifiers but using our proposed feature selection methods mRMR, NMIFS, CMIFS and CMIM for 3 different spatial resolutions 12 × 10, 18 × 15 and 24 × 20. Each row shows the average classification rate for 5 simulations, the standard deviation, and in parenthesis, the number of selected features for each model.

We summarize the results of SVM with raw pixels in the first row, Standard LBP (8, 2) features in the second row, and the Union of multiresolution of the LBP feature for different radii in the third and fourth rows. The row Raw Fea in Table IV represents the feature selection of the raw image (1 × 2,944), LBP 8, 2 represents the feature selection using LBP over the raw image (1 × 2,478). Boost Fea represents the 500 best features from

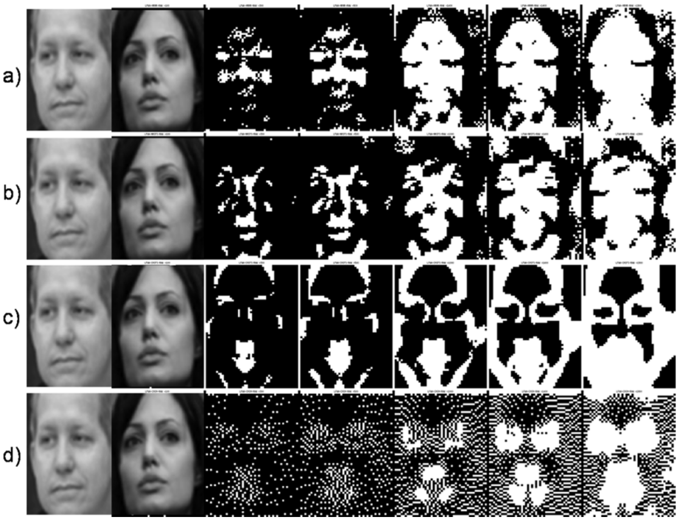

Fig. 8. Two original images, male and female, from the LFW database. The selected pixels for 300, 500, 1,000, 1,400, and 1,900 pixels are shown in white using: (a) mRMR, (b) NMIFS, (c) CMIFS, and (d) CMIM for Experiment 2.

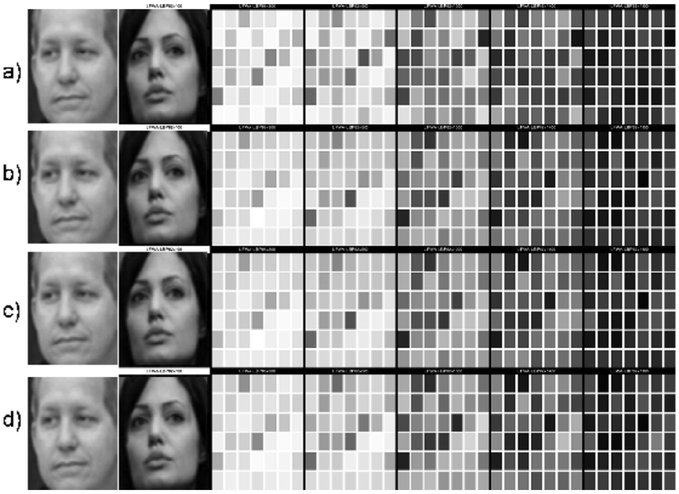

Fig. 9. Two original images, male and female, from the LFW database. Selected features from the LBPH(8, 2) histogram with 300, 500, 1,000, 1,400, and 1,900 features are shown for (a) mRMR, (b) NMIFS, (c) CMIFS, and (d) CMIM for Experiment 2. The darker the square, the larger the number of bins selected for the histogram.

TABLE V
RESULTS OF THE COMPUTATIONAL TIME REQUIRED TO CLASSIFY ONE IMAGE FROM THE SPECIFIED DATABASE, FOR THE METHODS WITH BEST GENDER CLASSIFICATION PERFORMANCE. TIME WAS MEASURED USING MATLAB ON A 2.5-GHz I7 PC WITH 8 GB OF MEMORY

Method	Without Selection	Features Selected	Raw Data (Sec)	mRMR (Sec)	NMIFS (Sec)	CMIFS (Sec)	CMIM (Sec)
L3 FERET	20,945	1,200	0.107	0.0016	0.0055	0.0061	0.007
L6 UND	34,159	7,900	0.027	0.0102	0.035	0.059	0.068
Boosted Fusion LFW 12x10	342,000	3,050	0.612	0.0037	0.0051	0,0055	0.0078
Best Fea FERET	139,735	18,900	0.0142				
Best Fea UND	139,735	14,200	0.0129				
Fusion Best LFW	59,742	10,400	0.0110				

each LBP vector with different radii from 1–8 $(1 \times 8 \times 500 = 4{,}000)$ and finally Boost Fusion represents the best N features for each LBPH vector with different radii $(1 \times 8 \times 2{,}478 = 19{,}824)$.

The best result for scale 12 × 10 was 96.73% for CMIM feature selection with 3,050 features; this represents only 0.89% of the original vector size. The best result for scale 18 × 15 was 95.94% for the CMIFS feature selection with 3,200 features, which represents only 0.93% of the vector size with no feature selection. The best result for scale 24 × 20 was 96.89% for the CMIM feature selection method with 4,150 features which represents only 1.21% of the vector size with no selection. Results from each of these scales separately (12 × 10, 18 × 15, 24 × 20) achieved better results than those published in [35]. The fusion of features from these 3 scales, improved even more the classification performance to 98.01% with only 10,400 features which is only 3.04% of the vector size with no selection.

Fig. 8 shows examples of selected pixels obtained in Experiment 2 with the feature selection methods, (a) mRMR, (b) NMIFS, (c) CMIFS and (d) CMIM. Fig. 8 shows 2 images one male and one female from the LFW database for 300, 500, 1,000 and 1,900 selected pixels are shown in white for the same size, 64 × 46.

Fig. 9 shows examples of selected features obtained in Experiment 2, with feature selection methods, (a) mRMR, (b) NMIFS, (c) CMIFS and (d) CMIM. Fig. 9 shows 2 images, one male and one female from the LFW database, with 300, 500, 1,000 and 1,900 selected features on size 64 × 46 images. The square shows the selected area, and the black intensity increase in the number of bins selected in that area. If the area was not selected, the square is white.

C. Computational Time

Table V, shows the computational time required to classify one image for the best results in Experiments 1 and 2 considering different fusion strategies, L1–L7, and using the feature selection methods, mRMR, NMIFS, CMIFS, and CMIM. The first column shows the best methods; the second column shows the feature vector size without selection; the third column shows the feature vector size with selection; column 4 shows the computation time required for all features without feature selection; and columns 5–8 show the computational time required for the different feature selection methods.

For Experiment 1, the best result was reached with 1,200 features (of 20,945) and the feature selection method, mRMR that required 1.6 ms for L3 on the FERET database. Also in Experiment 1 the best result was reached with 7,900 features of 34,159 for the feature selection method, CMIFS that required 5.9 ms for L6 on the UND database. For Experiment 2, the best result was reached with 3,050 features (of 19,824) and the feature selection method CMIM required 7 ms on the LFW database for Boosted fusion with subwindow sizes of 12 × 10 pixels. Rows 5, 6 and 7, show the computational time required for the best fusion method considering all features with the FERET, UND and LFW databases.

The computational time employed for "Best fea" (Fusion of all features L1 to L6) for FERET and UND was 14 ms with 18,900 features and 12 ms with 14,200 features. The computational time needed for "Fusion Best" (Fusion of the best for LFW) was 11 ms with 10,400 features.

TABLE VI
GENDER CLASSIFICATION EXPERIMENTAL RESULTS ON THE AR FACE DATABASE INCLUDING OCCLUSIONS (SUNGLASSES AND SCARVES). THE FIRST COLUMN SHOWS THE FOUR BEST FUSION SOLUTIONS FOUND FOR THE LFW DATABASE AND NOW TESTED FOR THE AR DATABASE. THE SECOND COLUMN SHOWS THE VECTOR SIZE. THE THIRD–SIXTH COLUMNS SHOW THE RESULTS OF GENDER CLASSIFICATION FOR THE FOUR FEATURE SELECTION METHODS (mRMR, NMIFS, CMIFS, AND CMIM). IN PARENTHESIS IS SHOWN THE SELECTED NUMBER OF FEATURES. THE HIGHEST SCORES ARE HIGHLIGHTED IN BOLD

Fusion	Vector	mRMR(%)	NMIFS(%)	CMIFS(%)	CMIM(%)
Raw_Fea 24x20	2,944	82.78 +/-0.87 (150)	88.72+/-0.88 (350)	90.97+/-0.76 (150)	**92.48+/-0.65** **(150)**
LBP(8,2) 24x20	2,478	87.21+/- 0.98 (500)	91.72+/-0.80 (300)	94.48+/-0.55 (200)	**95.24+/-0.76** **(200)**
Boosted SVM 24x20 Boosted	4,000	91.02+/- 0.80 (500)	90.67+/-0.78 (500)	**93.15+/- 0.90** **(500)**	91.79+/-0.56 (500)
Fusion SVM 24x20	19,824	92.30+/-0.77 (2100)	93.27+/-0.44 (5900)	93.55+/-0.33 (3300)	**96.43+/-0.75** **(4500)**

In this work we used the database Labeled Faces in the Wild which contains labeled face photographs with a wide range of conditions typically encountered in everyday life. The database exhibits "natural" variability in factors such as pose, lighting, race, accessories, occlusions, and background. The described experimental method was designed to make our research consistent and comparable with previously published results. We added an experiment with the AR face database [37] which includes occlusions with sunglasses and scarves. It contains images of 126 individuals (70 men and 56 women). Images include frontal faces with different facial expressions, illumination conditions and occlusion (sunglasses and scarves). In our test we used session 1, with 667 images (333 male and 334 female) including 200 face images with sunglasses and 200 images with scarves and 267 frontal images. For our model we used the same parameters selected for the LFW database to test our methods in the AR database using a scale 24 × 20. The 4 tests performed were: Raw_fea, LBP (8, 2), Boosted_Fea and Boosted_Fusion for the 4 feature selection methods (mRMR, NMIFS, CMIFS and CMIM).

Table VI shows the gender classification results obtained for the AR face database including occlusions (sunglasses and scarves). The tested fusion methods were the same that resulted with the highest scores in the LFW database and are shown in the first column of Table VI. In columns 3–6 of Table VI, it can be observed that the best results were obtained with the CMIM feature selection method reaching the highest score of 96.4% correct gender classification rate in this database including occlusions with sunglasses and scarves.

The computational time required for the best gender classification methods was measured using Matlab on a 2.5 GHz I7 PC with 8 GB of memory. Computational time shown in Table V can be further reduced by implementing the methods in C and using parallel computation.

D. Statistical Analysis

We used the ANOVA (analysis of variance) multicomparison test [41] to determine whether or not differences among results were statistically significant. We compared the results of the different methods with fusion and feature selection using MI versus the results without feature selection (Raw data).

TABLE VII
BEST RESULT OF GENDER CLASSIFICATION RATES PUBLISHED ON THE FERET, UND, AND LFW DATABASE, COMPARED WITH THE OUR PROPOSED METHODS. THE FIRST FOUR ROWS SHOW PREVIOUSLY PUBLISHED RESULTS. COLUMN 2 SHOWS THE GENDER CLASSIFICATIONS AND COLUMN 3 SHOWS THE TOTAL NUMBER OF SELECTED FEATURES. ROWS 5–11 SHOW THE RESULTS OF OUR PROPOSED METHODS

Methods	Classification Rates [%]	Number of Features
FERET [35]	99.07	46,845
UND [35]	91.19	46,485
LFW [36]	94.81	342,200
FERET [25]	99.13	33,800
Best_fea_FERET	99.13	18,900
Best_fea_UND	94.01	14,200
Best_fea_LFW(L3)	95.60	600
1)Boosted Fusion_LFW(12x10)	96.73	3,050
2)Boosted Fusion_LFW (18x15)	95.94	3,200
3)Boosted Fusion_LFW (24x20)	96.89	4,150
Fusion_Best_LFW (1+2+3)	98.01	10,400

The results in [35] were deterministic because only 1 partition of the database was employed. We replicated the result of [35] and use a fivefold cross-validation method with the same group of images. By using cross-validation method, we can compare the statistical significance of the results using the ANOVA test.

In Table I, the ANOVA showed that L1, L3, L4, L5, L6 have means that are significantly different for the FERET database considering L1 to L7. In all cases p was smaller than 1.51e-06 ($p < 0.001$) which is highly statistically significant. The best result was obtained with mRMR L3 fusion with 1,200 features followed by L7, L6, L5, L1, L4. Only for L2 our results were lower than those published in [35] (96.26%). In this other 6 methods (L1–L7), our results were significantly better than those previously published. Result indicate that when combining and selecting only shape features, feature selection does not improve gender classification results.

In Table II, the ANOVA showed that L1, L3, L4, L5, L6, L7 have means that are significantly different for the UND database considering L1 to L7. In all cases p was lower than 0.001 which is highly statistically significant. The best result was obtained with L6 fusion with 7,900 features followed by L3, L5, L1, and L4. Again the L2 (shape features) yielded the lowest classification result.

In Table IV, regarding the LFW database, the ANOVA showed that Boosted fusion 12×10, LBP(8, 2) 18×15, Boosted fusion 18×15, LBP(8, 2) 24×20 and Boosted fusion 24×20 have means that are significantly different compared to Raw data (without selection). In all cases p was lower than $p < 0.001$ which is highly statistically significant. The best result was obtained with Boosted fusion 24×20 with 96.89% and 4,150 feature selection, followed by Boosted fusion 18×15, LBP(8, 2) 24×20, LBP(8, 2) 18×15, and Boosted Fusion 12×10. The Raw Fea 12×10 had the lowest classification result.

After analyzing the results, it was concluded that feature selection and fusion improved the performance of gender classification significantly in the 3 databases FERET, UND and LFW, see Table VII. Also, comparing results among the 3 databases, it can be stated that results on the FERET are, in general, better than those obtained in the UND and LFW databases, because the face quality is better in the FERET database compared to the other 2 databases. The fusion using LBP features yielded the best results on both experiments allowing the representation of the data in a lower dimensional space and in shorter computational time. These results are significantly better than all those previously published (see, Table VII).

V. Conclusion

A new method for gender classification of faces is proposed using feature selection based on mutual information and fusion of intensity, shape and texture features, as well as different spatial scales. Four different measures of MI: mRMR, NMIFS, CMIFS, and CMIM, were employed to select features. The method was assessed using unconstrained face images from the LFW database and on face images taken under controlled conditions such as those in the FERET and UND databases.

In Experiment 1, the best performance was obtained with the fusion of 18,900 selected features (Best_Fea) reaching a classification rate of 99.13% on the FERET database. This is the best result reported to date for gender classification on the FERET database.

For the UND database, the best gender classification performance was obtained with the fusion of 14,200 selected features (Best_Fea) reaching a classification rate of 94.01%. This is the best result reported to the present for gender classification on the UND database.

In Experiment 2, the best performance was obtained with the fusion of 10,400 features from 3 different spatial scales (Best Fusion) obtaining a classification rate of 98.01%. This is the also the best result reported so far for gender classification on the LFW database.

The 4 selection methods used in this work, mRMR, NMIFS, CMIFS and CMIM quantify the features relevance and redundancy which is used in feature selection. These methods provide tools to select features with low redundancy and high relevance for the classification task (male, female). In this form, the problem dimensionality can be reduced improving classification rate and shortening the computational time required for feature extraction/classification as our results show. In the presence of a very large number of features (tens of thousands), it is common to find a large number of features that do not contribute to the classification process because they are irrelevant or redundant with respect to a particular the class [42], [43]. The feature selection methods used in this paper act as filters eliminating most of the features with low relevance or high redundancy and provide an efficient approach in terms of the computational time required for gender classification [19], [25]. These methods are considered effective for feature selection, especially when a large number of features are processed [8]. In our model, the process of training the SVM classifier was achieved more efficiently and effectively eliminating a significant number of features with low relevance and high redundancy. The selected features were independent of the classifier training method. Our approach does not remove all redundant features because they usually have similar rankings. In particular, since faces are relatively symmetrical several redundant features may arise with similar scores and therefore not all redundant features will be eliminated. Fusion of different type of features at different scales (intensity, shape and texture), provide a complementary form of considering a group of features more relevant than the same features acting independently. Levels of relevance can be defined in terms of those features that provide the highest information with respect to class C (male, female) and this information does not exist in other pairs of features. Fusion allows replacing features with weak relevance by other features (Intensity, Shape or texture) without loss of information.

Our results show that gender classification can be significantly improved by feature selection using different spatial scales, and by fusion of the selected intensity, shape and texture features. We performed experiments for different spatial scales and feature types in order to compare our results to those previously published. Our results show that for each spatial scale and for each feature type, feature selection improves results. Our results also show that feature fusion at the feature level, i.e., concatenating selected features at the classifier input, also improves gender classification compared to cases with no feature fusion. Combination of our results including feature selection and fusion for different spatial scales and feature types, yielded the highest performances published to date on the 3 standard databases.

These results also show that improvements were greater by fusing features from different scales, even when using a single type of feature, than those obtained by fusing different features on a single scale. Nevertheless, the highest gender classification performance was obtained by fusing features from different scales and types previously selected by the MI methods.

Another important result of the proposed feature selection method based on MI is that, depending on the image size, the total number of features was reduced 70% on the FERET database, 73% on the UND and 90% on the LFW database. Therefore, computational time is significantly reduced which makes real time applications of gender classification feasible.

References

[1] V. Axelrod and G. Yovel, “External facial features modify the representation of internal facial features in the fusiform face area,” *NeuroImage*, vol. 52, no. 2, pp. 720–725, 2010.

[2] C. A. Perez, C. M. Aravena, J. I. Vallejos, P. A. Estevez, and C. M. Held, “Face and iris localization using templates designed by particle swarm optimization,” *Pattern Recognit. Lett.*, vol. 31, no. 9, pp. 857–868, 2010.

[3] 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, 2011.

[4] C. Perez, V. Lazcano, P. Estevez, and C. Estevez, “Real-time iris detection on faces with coronal axis rotation,” *IEEE Trans. Syst., Man, Cybern., C, Applicat. Rev.*, vol. 37, no. 5, pp. 971–978, Sep. 2007.

[5] J. Bekios-Calfa, J. 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.

[6] N. Kwak and C.-H. Choi, “Input feature selection for classification problems,” *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

[7] 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.

[8] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” in *Proc. IEEE Bioinformatics Conf. CSB 2003*, pp. 523–528.

[9] P. A. Estevez, 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.

[10] F. Fleuret and I. Guyon, “Fast binary feature selection with conditional mutual information,” *J. Mach. Learning Res.*, vol. 5, pp. 1531–1555, 2004.

[11] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognit.*, vol. 42, no. 7, pp. 1330–1339, 2009.

[12] 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.

[13] B. Jun, T. Kim, and D. Kim, "A compact local binary pattern using maximization of mutual information for face analysis," *Pattern Recognit.*, vol. 44, no. 3, pp. 532–543, 2011.

[14] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 1991 (CVPR '91),* , Jun. 1991, pp. 586–591.

[15] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[16] K. Fukunaga and J. M. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, Jun. 1983.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[18] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010.

[19] Z. Zhang and E. Hancock, "Feature selection for gender classification," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, J. Vitrià, J. Sanches, and M. Hernández, Eds. Berlin/ Heidelberg, Germany: Springer, 2011, vol. 6669, pp. 76–83.

[20] Z. Zhang, E. Hancock, and J. Wu, "An information theoretic approach to gender feature selection," in *Proc. IEEE Int. Conf. Computer Vision Workshops (ICCV Workshops)*, Nov. 2011, pp. 1425–1431.

[21] C. Frank and E. Noth, "Automatic pixel selection for optimizing facial expression recognition using eigenfaces," *Pattern Recognit.*, vol. 2781, pp. 378–385, 2003.

[22] S.-I. Choi, C.-H. Choi, and G.-M. Jeong, "Pixel selection in a face image based on discriminant features for face recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2008, pp. 1–6.

[23] B. Li, X.-C. Lian, and B.-L. Lu, "Gender classification by combining clothing, hair and facial component classifiers," *Neurocomputing*, vol. 76, no. 1, pp. 18–27, 2012, Seventh International Symposium on Neural Networks Advances in Web Intelligence.

[24] Z. Xu, L. Lu, and P. Shi, "A hybrid approach to gender classification from face images," in *Proc. 19th Int. Conf. Pattern Recognition, 2008 (ICPR 2008)*, Dec. 2008, pp. 1–4.

[25] C. Perez, J. Tapia, P. Estévez, 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.

[26] E. Makinen and R. Raisamo, "An experimental comparison of gender classifications methods," *Pattern Recognit. Lett.*, vol. 29, pp. 1544–1556, 2008b.

[27] 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.

[28] M.-H. Yang and B. Moghaddam, "Support vector machines for visual gender classification," in *Proc. 15th Int. Pattern Recognition Conf.*, 2000, vol. 1, pp. 1115–1118.

[29] M.-H. Yang and B. Moghaddam, "Gender classification using support vector machines," in *Proc. Int. Image Processing Conf.*, 2000, vol. 2, pp. 471–474.

[30] Z. Sun, X. Yuan, G. Bebis, and S. J. Louis, "Neural-network-based gender classification using genetic search for eigen-feature selection," in *Proc. Int. Joint Conf. Neural Networks (IJCNN '02)*, 2002, vol. 3, pp. 2433–2438.

[31] B. Wu, H. Ai, and C. Huang, "Lut-based adaboost for gender classification," in *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2003, pp. 104–110.

[32] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1997, pp. 137–143.

[33] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," *Assessment of Time Dependency in Face Recognition: An Initial Study*, pp. 44–51, 2003.

[34] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments University of Massachusetts, Amherst, MA, USA, Tech. Rep., Oct. 2, 2007, pp. 7–49.

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

[36] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 431–437, 2012.

[37] A. Martinez and R. Benavente, The AR Face Database Tech. Rep. 24, CVC, 1998.

[38] H. Cheng, Z. Qin, W. Qian, and W. Liu, "Conditional mutual information based feature selection," in *Proc. Int. Symp. Knowledge Acquisition and Modeling (KAM '08)*, 2008, pp. 103–107.

[39] G. Wang and F. H. Lochovsky, "Feature selection with conditional mutual information maximin in text categorization," in *Proc. 13th ACM Int. Conf. Information and Knowledge Management*, 2004, pp. 342–349.

[40] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPv) with global matching," *Pattern Recognit.*, vol. 43, pp. 706–719, Mar. 2010.

[41] A. Cuevas, M. Febrero, and R. Fraiman, "An ANOVA test for functional data," *Computat. Statist. Data Anal.*, vol. 47, no. 1, pp. 111–122, 2004.

[42] A. R. Webb*, Statistical Pattern Recognition*. Hoboken, NJ, USA: Wiley, 2002, ISBN: 0-470-84513-9.

[43] T. M. Cover and J. A. Thomas*, Elements of Information Theory*, ser. Wiley Series in Telecommunications. Hoboken, NJ, USA: Wiley, 1991.

Juan E. Tapia (S'10) received the P.E. degree in electronics engineering from Universidad Mayor in December 2004 and the M.S. degree in electrical engineering from Universidad de Chile in 2012. He is working toward the Ph.D. degree at the Department of Electrical Engineering, Universidad de Chile.

His main interests are pattern recognition and machine learning applied to face recognition, gender classification, feature fusion, and feature selection.

Claudio A. Perez (M'90–SM'04) received the B.S. and P.E. degrees 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 at the Ohio State University where he received the Ph.D. degree in 1991.

In 1990, he was a Presidential Fellow and received a Graduate Student Alumni Research Award from O.S.U. In 1991, he received a Fellowship for Chilean scientists from Fundacion Andes. He was a visiting scholar at UC, Berkeley in 2002 through the Alumni Initiatives Award Program from Fulbright Foundation. He is a Professor at the Department of Electrical Engineering, Universidad de Chile. He was the Department Chairman from 2003 to 2006. His research interests include biometrics, image processing applications, and pattern recognition. He is member of the editorial board of the *International Journal of Optomechatronics*, Associate Editor of *BMC Neuroscience*, Senior Member of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETIC: SYSTEMS, and IEEE Computational Intelligence Society, and member of Sigma-Xi and OSU Alumni Association.