International Journal of Optomechatronics, 6: 92–119, 2012

ISSN: 1559-9612 print/1559-9620 online
DOI: 10.1080/15599612.2012.663463

Taylor & Francis
Taylor & Francis Group

GENDER CLASSIFICATION FROM FACE IMAGES USING MUTUAL INFORMATION AND FEATURE FUSION

Claudio Perez, Juan Tapia, Pablo Estévez, and Claudio Held
Department of Electrical Engineering and Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile

In this article we report a new method for gender classification from frontal face images using feature selection based on mutual information and fusion of features extracted from intensity, shape, texture, and from three different spatial scales. We compare the results of three different mutual information measures: minimum redundancy and maximal relevance (mRMR), normalized mutual information feature selection (NMIFS), and conditional mutual information feature selection (CMIFS). We also show that by fusing features extracted from six different methods we significantly improve the gender classification results relative to those previously published, yielding 99.13% of the gender classification rate on the FERET database.

Keywords: Feature fusion, feature selection, gender classification, mutual information, real-time gender classification

1. INTRODUCTION

During the 90's, one of the main issues addressed in the area of computer vision was face detection. Many methods and applications were developed including the face detection used in many digital cameras nowadays. Gender classification is important in many possible applications including electronic marketing. Displays at retail stores could show products and offers according to the person gender as the person passes in front of a camera at the store. This is not a simple task since faces are not rigid and depend on illumination, pose, gestures, facial expressions, occlusions (glasses), and other facial features (makeup, beard). The high variability in the appearance of the face directly affects their detection and classification. Automatic classification of gender from face images has a wide range of possible applications, ranging from human-computer interaction to applications in real-time electronic marketing in retail stores (Shan 2012; Bekios-Calfa et al. 2011; Chu et al. 2010; Perez et al. 2010a).

Automatic gender classification has a wide range of possible applications for improving human-machine interaction and face identification methods (Irick et al.

Address correspondence to Claudio Perez, Image Processing Laboratory, Department of Electrical Engineering, Universidad de Chile Casilla 412-3, Av. Tupper 2007, Santiago, Chile. E-mail: clperez@ing.uchile.cl

NOMENCLATURE

c	classes	$P(x, y)$	joint probabilistic distribution
CMIFS	conditional mutual information feature selection	$P(x)$	marginal probabilities x
		$P(y)$	marginal probabilities y
fi	feature (i)	S	subset features
$I(x, y)$	pixel position	$\|S\|$	cardinality of S
$I(xc, yc)$	central pixel position	SVM	support vector machine
mRMR	minimum redundancy and maximal relevance	$\cup$	concatenation operator
		V_I	relevance
MI	mutual information	W_i	redundancy
MI_N	normalized MI		
MID	mutual information difference	**Greek Letters**	
MIQ	mutual information quotient	Ω_s	set features
$N(x, y)$	vicinity around (x, y)		
$NMIFS$	normalized mutual information feature selection		

2007; Lu, Xu, and Shi 2009; Makinen and Raisamo 2008b). Gender classification may be used to partition a face recognition database building clusters to reduce the number of comparisons to identify a face (Wang et al. 2004; Wu, Ai, and Huang 2003; Alexandre 2010). Other novel applications include demographic information collection, consumer behavior assessment, and selective electronic marketing in retail stores.

Gender classification research is an emerging topic compared with other biometric identification methods such as fingerprinting; face recognition, or iris identification (Perez et al. 2010b; Jun et al. 2011). A complete literature review and comparison among best gender classification methods was performed in Makinen and Raisamo (2008a; 2008b). It concluded that a relatively small number of papers had been published with proven results on large and internationally available databases. Most other articles comparing different gender recognition approaches report results on non-identified subsets of larger databases, or on homemade examples which are not possible to replicate for comparison purposes (Brunelli and Poggio 1995; Shakhnarovich, Viola, and Moghaddam 2002; Moghaddam and Yang 2002; Wu, Smith, and Hancock 2010). Consequently, our literature review focuses only on those methods capable of being compared on standard datasets.

Four different methods for gender classification were compared (Makinen and Raisamo 2008b) in two large and internationally available faces databases; the FERET (Phillips et al. 1997) and WWW (Makinen and Raisamo 2008b). The methods were: a multi-layer Perceptron neural network (NN) and a support vector machine (SVM) with pixel-based inputs, an Adaboost, and an SVM with Local Binary Pattern (LBP) features as inputs. The best result was Adaboost with 93.33% for image sizes of 32×40 pixels. In Makinen and Raisamo (2008a), these methods were compared for three different image face sizes: 24×24, 36×36, and 48×48 pixels. The best classification rate was achieved with SVM with 36×36 pixel images reaching an accuracy of 86.54% on the FERET database. An approach to gender recognition based on shape, texture, and intensity features using different scales was proposed (Alexandre 2010). Different methods for gender classification

were compared in the same group of images on the FERET database. The best gender classification accuracy based on pixel intensities was 87.85% for a 36×36 pixel image. Results using shape features yielded 91.59% correct classification for 128×128 size, and 93.46% using LBP texture features, also on 128×128 image sizes. Fusing the three types of features (intensity, shape and texture) yielded the best score of 95.33% on the FERET database (Alexandre 2010). Using three spatial scales of 20×20, 36×36, and 128×128 and the three types of features (intensity, shape, and texture), a score of 99.07% was reached on the FERET database (Alexandre 2010). The total number of inputs was increased nearly nine-fold, 46,845 inputs, by using three types of features and three different scales. Computational time depends on the number of inputs to the classifier and it is an important factor in most real-time applications involving face processing (Perez et al. 2007, Perez et al. 2005) and therefore a feature selection process is desirable. Also, in Mayo (2008), a method was proposed for gender classification by expanding the training data set with examples of faces that had been deliberately misaligned. The overall best result for gender classification on the FERET database was 99.07%.

Starting from a large number of features extracted from input data, feature selection is the process of selecting a subset of relevant features which contain useful information for distinguishing one class from the others (Vinh, Thang, and Lee 2010). One of the main goals of feature selection is to represent the data in a lower dimensional space (Sun, Bebis, and Miller 2004; Bekios-Calfa et al. 2011). The lack of an effective method for selecting an appropriate set of features has been compensated in part, by classification algorithms capable of dealing at least partially with redundant and irrelevant features (Sun, Bebis, and Miller 2004). The goal of the feature selection process is to choose the smallest subset of features that carry as much information about the class as possible. In Perez, Cament, and Castillo (2011) a method for feature selection based on an entropy measure for face classification was proposed for a local matching Gabor-based approach. Besides significantly reducing the number of computations, the recognition rate was also improved.

Mutual information (MI) has been used as a feature selection criterion because of its good representation of relevance and redundancy between random variables and its robustness in noisy environments and data transformations. Moreover, MI can provide an optimal feature set regardless of the classifiers (Chow and Huang 2005). Mutual information has been used with success in other applications to select features (Huang, Cai, and Xu 2006; Peng, Long, and Ding 2005).

In this article we report the use of feature selection based on MI and fusion of three features for gender classification. Starting from a large number of features, such as all pixels in the image, a selection is performed from the input data. The objective is to find the best subset of relevant features that contains useful information to distinguish one class from the other. In this way, the data is represented in a lower dimensional space allowing higher classification performance because noisy inputs are discarded and processing time can be reduced significantly since fewer features are computed. We also show that fusion among selected features improve classification results. We reached significantly better results than all those previously published in gender classification. A reduction in computational time is essential in many real time applications, and therefore feature selection methods and fusion of features is highly desirable.

We also show that fusion of features extracted from six different methods improves significantly classification performance. We compare the results of three different mutual information measures: minimum redundancy and maximal relevance (mRMR) (Ding and Peng 2003), normalized mutual information feature selection (NMIFS) (Estévez et al. 2009), and conditional mutual information feature selection (CMIFS) (Cheng et al. 2008), and show that the fusion of features from different methods significantly improves the classification results relative to those previously published. We compare our results to those of the best gender classification methods published based on standard face databases (FERET and WWW face databases) in the literature. Using different image sizes and database partition we obtained significant gender classification improvements that ranged from 1.2% to 12.7% on the FERET database and from 4.1% to 8.9% on the WWW face database. This is significantly better than all previously published results that reduced the number of classifier inputs on the same databases. Based on our literature review, this is the first time feature selection based on mutual information has been applied to gender classification.

2. CLASSIFICATIONS ALGORITHMS

In this article we report the use of feature selection based on MI and fusion of three features for gender classification. Starting from a large number of features (all pixels in the image), we select from the input data the best subset of relevant features which contains only the useful information to distinguish one class from the other. Selected features are used as inputs to the classifiers. Fusion among selected features should improve classification results. Also, a reduction in computational time is essential in many real time applications. The diagram in Figure 1 shows the main steps of the proposed method for both experiments (1 and 2); face detection, alignment, feature selection/fusion, and classification.

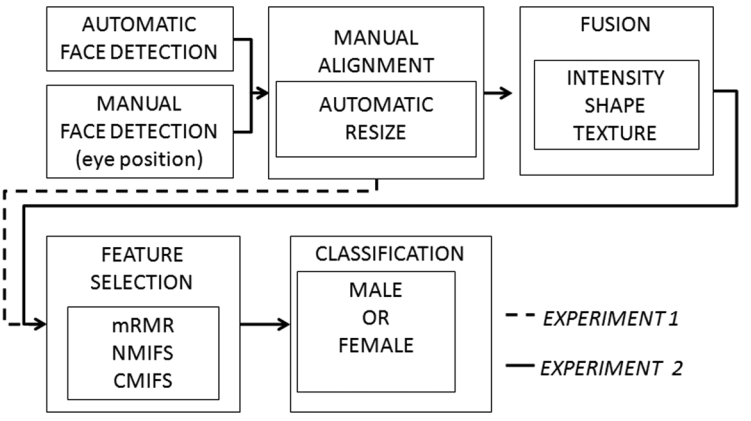

Figure 1. Flow chart representing the experiments performed. The dashed arrow indicates that this stage was used in experiment 1. The black arrow indicates the fusion of features used in experiment 2.

2.1. Feature Selection Criteria

The goal is to select a feature subset that best characterizes the statistical property of a target classification variable, subject to the constraint that these features are mutually as dissimilar to each other as possible, but marginally as similar to the classification variable as possible. Different forms of mRMR, where ''relevance'' and ''redundancy'' were defined. MID and MIQ represent the Mutual Information Difference and Quotient, respectively, to combine the relevance and redundancy that are defined using MI. These are the two most used forms of mRMR. NMIFS is an improved version of mRMR based on the normalized feature of mutual information; the MI between two random variables is bounded above by the minimum of their entropies. As the entropy of a feature could vary greatly, this measure should be normalized before applying it to a global set of features. CMIFS, is a greedy algorithm, that detects both cooperation and redundancy interaction of features. These characteristics can be obtained from images of intensity, texture, color, shape, and others.

Traditionally, information theory is used to quantify concepts of relevance and redundancy and can be used in feature selection methods (Ding et al. 2003). We formalized these concepts and applied them to feature selection in gender classification. Given an input set F (feature) and output class C (gender), the first step is to find which features have more information to describe C. The decision of which features should be chosen is usually associated to the degree dependency of each single feature when used to describe C. However, a group of features may be more relevant than the same features acting independently. This means that many different levels of relevance can be defined. The concept of redundancy is associated with the level of dependency between two or more features in F. This can be quantified by the common information shared by features (Peng et al. 2005). A critical issue in discriminant analysis is feature selection. Instead of using all available variables (features or attributes) in the data, a subset of features is selected to be used to discriminate classes. Some advantages of feature selection are the dimension reduction to reduce the computational cost, reduction of noise to improve the classification accuracy and more interpretable features or characteristics that can help identify and monitor the operation of a classifier system. If a feature or pixels have expressions randomly or uniformly distributed in different classes, its mutual information with these classes is zero (Ding and Peng 2003). In this article we report use of feature selection based on mutual information (MI) for gender classification to show that it provides a general and powerful framework for reducing the number of features and improving gender classifications rates. We compare the results of three different mutual information measures: mRMR, NMIFS, and CMIFS. We also show that the union and features selection significantly improves the classification results relative to those previously published. In the classifications task, redundant features may act as noisy inputs to the classifier and should be removed. The mutual information measures mRMR, NMIFS, and CMIFS recognize a redundant candidate features based on its dependency with the selected features. The goal of mRMR is to select a feature subset that best characterizes the statistical property of a target classification variable, subject to the constraint that these features are mutually as dissimilar to each other as possible, but marginally as similar to the classification variable as possible. We showed several different forms of mRMR, where ''relevance'' and ''redundancy'' were defined using

mutual information (MID, MIQ). MID and MIQ represent the Mutual Information Difference and Quotient, respectively, to combine the relevance and redundancy that are defined using MI. They are the two most used mRMR schemes (Ding and Peng 2003; Akadi et al. 2009). NMIFS is based on the normalized feature of mutual information; the MI between two random variables is bounded above by the minimum of their entropies. As the entropy of a feature could vary greatly, this measure should be normalized before applying it to a global set of features (Peng et al. 2005). CMIFS is a greedy algorithm; it will remove classification redundancy features beforehand. CMIFS can be used to detect both cooperation and redundancy interaction of features (Ding and Peng 2003). These MI measures can be applied to images of intensity, texture, color, shape, and others.

The *MI* (Ding and Peng 2003) 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

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

We use mutual information to measure the level of "similarity" between pixels. The concept of minimal redundancy, as in equation (2), allows selection of pixel pairs that are maximally dissimilar (Ding and Peng 2003). When two features highly depend on each other, the respective class-discriminative power would not change much if one of them were removed. Therefore, the following minimum redundancy (Min Red) condition can be added to select mutually exclusive features.

$$\text{MinRed, Re}d = \frac{1}{|S|^2} \sum_{fi,fj \in S} MI(fi; fj) \tag{2}$$

where S denotes the feature subset, $|S|$ is the number of features in S, and $MI(fi; fj)$ is used to represent the mutual information between features. *fi* y *fj*. $MI(c; fi_i)$ quantifies the discriminant power for different target classes that can be obtained by the mutual information among target classes $c = \{c_1, c_2, c_3 \ldots . c_K\}$. Thus, $MI(c; fi)$ quantifies the relevance of fi_{th} feature for the classification task.

Therefore, the maximum relevance condition shown in equation (3) maximizes the total relevance of all pixels in s (Ding and Peng 2003). Maximal relevance (Max Rel) is to search features that approximate with the mean value of all mutual information values between individual features *fi* and class c

$$\text{Max Rel, Re}l = \frac{1}{|S|} \sum_{fi \in S} MI(c; fi) \tag{3}$$

The first feature is selected according to Re*l*, i.e., the feature with the highest-$MI(c; fi)$. 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 $\Omega = \Omega - S$ then the two conditions are optimized: the min operation from equation (2) is interpreted as minimum redundancy computation; the max operation is interpreted as

maximum relevance as shown in equation (3). The criterion combining the above two constraints is called "minimum-redundancy-maximal-relevance" (mRMR).

2.1.1. Minimum redundancy and maximal relevance (mRMR). Two forms of combining relevance and redundancy operations are used here (Peng, Long, and Ding 2005): mutual information difference (MID) equation (4), and mutual information quotient MIQ equation (5) as

$$MID = \max(\mathrm{Re}l - \mathrm{Re}d), \tag{4}$$

$$MIQ = \max(\mathrm{Re}l/\mathrm{Re}d). \tag{5}$$

The mRMR feature set is obtained by optimizing the conditions in equation (4) and equation (5) simultaneously. Optimization of both conditions requires combining them into a single criterion function (Ding and Peng 2003) as shown in the following

$$f^{mRMR}(X_i) = MI(c;fi) - \frac{1}{|S|}\sum_{fi\in S} MI(fi;fj) \tag{6}$$

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

2.1.2. Normalized mutual information feature selection (NMIFS). In Estévez and colleagues (2009), we proposed an improved version of mRMR based on the normalized feature of mutual information; the MI between two random variables is bounded above by the minimum of their entropies. As the entropy of a feature could vary greatly, this measure should be normalized before applying it to a global set of features (Estévez et al. 2009) as

$$f^{NMIFS}(X_i) = MI(c;fi) - \frac{1}{|S|}\sum_{fi\in S} MI_N(fi;fj), \tag{7}$$

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

$$MI_N(fi;fj) = \frac{MI(fi;fj)}{\min(H(fi),H(fj))}. \tag{8}$$

2.1.3. Conditional mutual information feature selection. Let S be the set of already-selected features, and the set of candidate features Ω, $S\ \Omega = \emptyset$ and c is the output class set. The next feature in Ω to be selected is the one that makes $MI(c;fi,S)$ maximum, where $f_i \in \Omega$ and

$$MI(c;fi,S) = MI(c;fi) - [MI(fi;S) - MI(fi;S|c)] \tag{9}$$

Then c is the output class, and S is the selected feature subset, Ω is the candidate feature subset, and $f_i \in \Omega S$. The following pseudo code describes feature selection using CMIFS (Cheng et al. 2008) where f_i is the ith feature selected.

(i) Initialization: set $\Omega = \{fi/i = 1, \ldots, (N)\}$, initial set of N features, and $S = \{\emptyset\}$, empty set.
(ii) Compute the MI with respect to the classes: compute $MI(fi; c)$, for each $f_i \in \Omega$.
(iii) Select the first feature: find $fi = \max_{i=1,\ldots,N}\{MI(fi; c)\}$. Set $\Omega_s \leftarrow \Omega/\{fi\}$; set $S \leftarrow \{fi\}$.
(iv) Greedy selection: repeat until $|S| = k$.
  a. Compute the MI between features; compute $MI(fi; fj)$ for all pairs $(fi; fj)$, with $f_i \in \Omega$ and $f_j \in S$, if it is not available.
  b. Select the next feature $f_i \in \Omega$ that maximizes the measure. Set $\Omega_s \leftarrow \Omega/\{fi\}$; set $S \leftarrow \{fi\}$.

The theoretical complexity for the feature selection stage is:

$|S|$ is the number of feature in S, N is the number of data samples available. MID and MIQ requires $O^*(^*N)$, mRMR, requires: $O^*(N)$, NMIFS, requires: $O^*(N^*\log^*N)$, CMIFS, requires $O^*(S\text{-}N)$.

2.2. Feature Extraction and Fusion

In this article we use three different types of face features to classify gender. We extract intensity, shape, and texture features using three different spatial scales, see Figure 2. For the spatial scales we used the same as in Makinen and Raisamo (2008a) 24×24, 36×36, and 48×48. We also used 24×24 and 32×40 for the FERET database (Phillips et al. 1997) and WWW for comparsion with Makinen and Raisamo (2008b). Additionally we use face image dimensions of 20×20, 36×36, and 128×128 for the FERET database to compare with Alexandre (2010).

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 are computed using the masks $[-1, 0, 1]$ and $[-1, 0, 1]^T$. Consider that v and h are the vertical and horizontal edge values at any pixel obtained by convolution of the edge masks with the original image, respectively. The edge map is found using $\theta = \tan^{-1}\left(\frac{v}{h}\right)$ and the edge magnitude is given by $m = \sqrt{v^2 + h^2}$. The edge map is

Figure 2. Example images of image (*a*) intensity, (*b*) shape, and (*c*) texture.

discretized in 18 degree intervals. Each pixel adds its magnitude $m$to the bin that correspond to its edge directions θ. For N image windows, an image is represented by $20^{*}N$ real values. The N window sizes are specified in 2.4.2.

For the texture features, in this work we use the LBP transformation. LBP features are computed from pixel intensities in a neighborhood as in Ojala, Pietikainen, and Maenpaa (2002). If $h(I(xc,yc),I(x,y))$ is a comparison operator such that $h=1$ if $I(xc,yc)<I(x,y)$ and $h=0$ otherwise, then

$$LBP(x,y) = \bigotimes_{(x',y') \in N(x,y)} h(I(x,y), I(x',y')) \tag{10}$$

where $N(x, y)$ is a vicinity around (x, y) and U is the concatenation operator.

After the feature extraction, we fuse the information at the feature level by combining the feature vectors from different sources into a single feature vector that becomes the input to the feature selection method and then the selected features become the inputs to the classifier. The classifiers are trained with different features and with the fused features.

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. L1, L2, and L3 represent fusion of intensities, shape and texture for three different sizes, 20×20, 36×36, and 128×128. L4, L5, L6 represent the fusion of three different features (intensity, shape, and texture) for 20×20, 36×36, and 128×128 respectively. L7 represent fusion of three scales and three types of features, as shown in Figure 3.

2.3. Classifiers

2.3.1. Dataset experiment 1. The tests were performed on two internationally available face databases. These databases were used to train and test the MI feature selection method as well as the fusion method to allow comparison of results with those previously published. As in Makinen and Raisamo (2008b), faces of one image per person from the Fa and Fb subsets were used and duplications were eliminated. Therefore, 450 female and 450 male normalized images were used from the FERET database. The second database, the WWW image database,

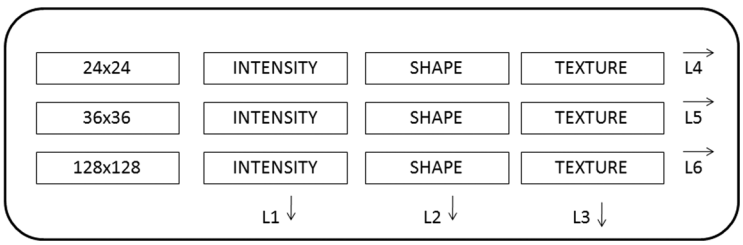

Figure 3. Diagram explaining the fusion of features made in experiment 2. It shows clearly how fusion is obtained from L1 to L6. In this diagram L1, L2, and L3 were obtained from vertical fusion of features and L4, L5, and L6 were performed by horizontal fusion of features.

contains 2,360 female and 2,360 male images, randomly collected from the World Wide Web and annotated by the researchers (Makinen and Raisamo 2008b). Faces were detected automatically by face Detector OpenCV 1.0 (Open CV 2005). To compare our results with those of Makinen and Raisamo (2008a; 2008b) two sets of image sizes were used. The first set was composed of 411 images (212 males and 199 females) and three image sizes (24×24, 36×36, and 48×48) from the subset Fa of the FERET database (Makinen and Raisamo 2008a). The second set was composed of 900 images of 24×24 and 760 images of 32×40 from the subset Fa of the FERET database (Makinen and Raisamo 2008b). We used a PC Intel I7 with 4 GB of memory Ram.

As in (Makinen and Raisamo 2008a; 2008b), both databases, FERET and WWW, were partitioned to have 80% training data and 20% testing data. All results were obtained with fivefold cross-validation, simulations. For all four models, a training set was used to determine the best number of features that were tested later on the test set for the FERET as well as the WWW databases.

In the FERET database, the eyes were located manually in the faces and face were rotated and aligned in the images so that each face had eyes in the same location and all faces were in the upright position (Makinen and Raisamo 2008b). Later the face areas were scaled to different sizes. The following rules were used for the calculation of the rectangular area as shown in Figure 4. The width for the rectangle is computed using the eye distance, *de*. The *de* is the distance between the detected locations of the left eye and the right eye. A space is left on both sides of the eyes which is $0.25^{*}de$ for each side of the eyes. The total width for the rectangle is $1.5^{*}de$. The space included above the eyes is $0.5^{*}de$. The height of the rectangle is $2.2^{*}de$. In the WWW database the images were not aligned and faces were detected

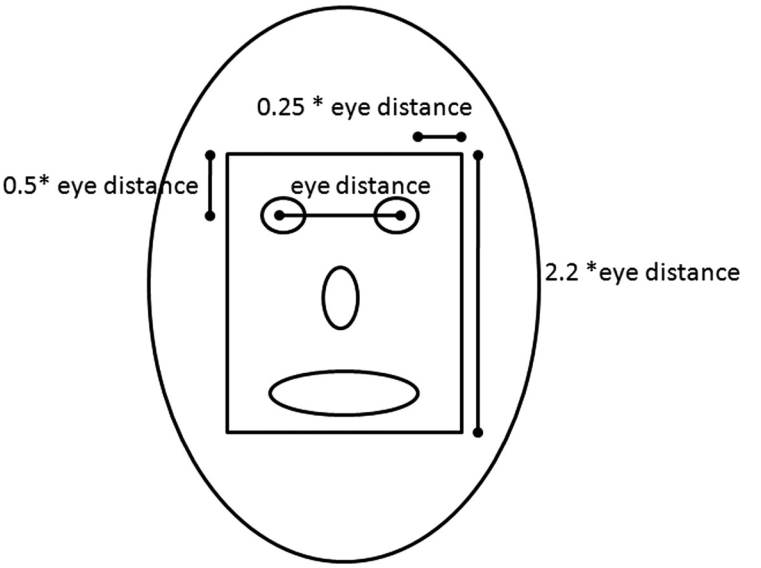

Figure 4. Graphic description of the measures used to standardize and align face images, when the coordinates of the eyes are used.

using the face detector of OpenCV (OpenCV, 2005). In the Morph-II face database, faces were detected and aligned using the face detection OpenCV.

2.3.2. Dataset experiment 2. In this experiment we used the same image sizes as those used in Alexandre (2010) to be able to compare our results of feature selection and fusion to those previously. We compare the results of our method to those reported in (Alexandre 2010; Makinen and Raisamo 2008a) using the same dataset from the FERET Database with sizes 20×20, 36×36, and 128×128. Three hundred and four images are used for training and 107 for testing. All results were obtained with fivefold cross-validation, simulations. For all four models, a training set was used to determine the best number of features that were tested later on the test set.

In experiment 2, we performed 7 tests, called L1 to L7, where L1 represents fusion of intensities of pixels from different size images, L2 represents fusion of shapes from different size images and L3 represents fusion textures from different size images (LBP). Different image sizes were 20×20, 36×36, and 128×128, respectively. L4 represents fusion of three features (intensity, shape, and texture) for size 20×20, L5, L6 represent the same fusion but sizes, 36×36 and 128×128, respectively. L7 represent fusion of three scales (20×20, 36×36, and 128×128) and three types of features (intensity, shape, and texture). Fusion is performed by simple concatenation of inputs to the classifier. Figure 3 illustrates the 7 different fusion schemes tested. Figure 2 shows example images for intensity, shape, and texture.

2.3.3. Cross database performance. The Labeled Faces in the Wild (LFW) and MORPH-II databases were used to test cross database performance (Dago-Casas et al. 2011). These databases were tested with the best cases selected for the FERET and WWW databases.

The LFW database contains faces of 5,749 individuals (4,263 male, 1,486 female) collected from the web using a Viola-Jones face detector. Of these, there are 1,680 people for which more than one image is available. This results in 10,256 male images and 2,977 female images. These color images have a resolution of 250×250 (Huang et al. 2007).

The MORPH-II database is composed of 55,608 color images of 13,673 subjects of the age between 16 and 99 years, where 47,057 images correspond to male persons and 8,551 to female persons. 42,897 of these images depict black faces, 10,736 white, 1,753 Hispanic, 160 Asian, 57 Indian and 5 faces are of other ethnicities. The images have varying resolutions of either 200×240 or 400×480 pixels. This dataset is highly imbalanced towards black male persons and missing images of persons below the age of 16 (Ricanek and Tesafaye 2006).

2.4. Classifiers

2.4.1. SVM model with pixel based inputs (SVM). An SVM model was used to classify gender from a selected set of pixels in the same manner as in Makinen and Raisamo (2008a; 2008b) for experiment 1 and as in Alexandre (2010) for experiment 2. The selected pixels were used as input to the SVM (Vankayalapati et al. 2011). We trained the SVM with histogram equalized image

pixels. The pixel intensities were scaled to range from −1 to 1 and transformed with the RBF kernel. The most relevant pixels were selected using mRMR, NMIFS, and CMIFS. In experiment 1, a number of selected features were in the range 50–500 for image size 24×24, 50–1,000 for 36×36, and 50–2,000 for 48×48. In experiment 2, the number of selected features was in the range 50–10,000 for L1, 50–5,000 for L2, 50–10,000 for L3, 50–1,000 for L4, 50–10,000 for L5, and 50–15,000 for L6. L7 represents the union of L1-L6 features, i.e., the union of mRMR for L1 to L6. The L7 (Best fea) represents the union of the best features of each one of L1 to L6, for each feature selection method with size of 33,800 features.

2.4.2. SVM model with LBP features (SVM_LBP). The LBP values were computed and used as inputs to the SVM with RBF kernel. For Experiment 1, as in Makinen and Raisamo (2008a; 2008b), the face image was divided in N, 8×8 blocks, and the LBP operator was applied to each block using 4-connected neighbors and a radius of one. Then a histogram with 16 bins was created for each block. We also performed the uniform *LBP* feature extraction with 8-connected neighbors and radius one. A 59-bin histogram was created for this *LBP*. The histograms were concatenated ($N^*16 + 59$) and the best features were selected using mRMR, NMIFS, CMIFS in the ranges 50–200 for image size 24×24, 50–300 for size 36×36, and 50–500 for size 48×48. In experiment 2, for the shape and texture features we chose a window size that yielded the best results in Alexandre (2010) for each image size of 128×128, 36×36, and 20×20. For 128×128 images the windows sizes was 16×16; for the 36×36 was 6×6, and for 20×20 images the window size was 10×10. In all cases the windows had 50% overlay, and the histograms were concatenated.

2.4.3. Neural network (NN) model. In experiment 1, a multi-layer Perceptron NN, trained with back-propagation method with equalized face images was used as in Makinen and Raisamo (2008b). Input features were used as inputs to the NN. Inputs were selected using mRMR, NMIFS, and CMIFS. Each node at the input layer represents one data value. The hidden layer with activation function has two nodes and the output layer has one node. The output of the NN is the classification result value in the range [−0.5, 0.5], where the value above zero was defined as male and below zero as female. Figure 5 shows the structure of the NN.

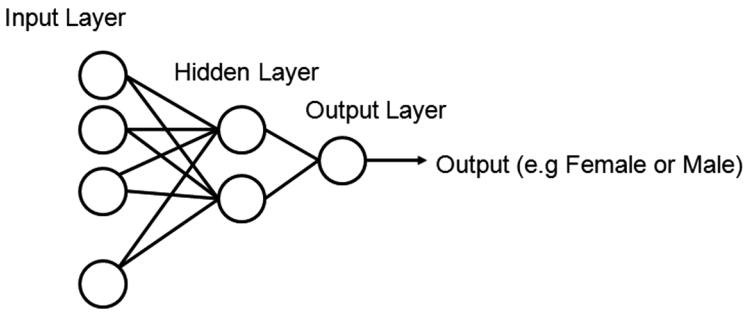

Figure 5. Example of a multi-layer Perceptron with one hidden layer and one output node.

2.4.4. Adaboost model. The other classifiers were: Support Vector Machine (SVM) with base input, different types of Adaboost, including Threshold Adaboost (ADA_TH), Gentle_Adaboost (ADA_GENTLE), Real Adaboost (ADA_REAL), Modest Adaboost (ADA_MOD) and SVM with local binary pattern (SVM_LBP) (Wu, Ai, and Huang 2003; Vezhnevets 2005; Qahwaji et al. 2008). In all cases the inputs were the selected features. The Adaboost model is based on a set of weak classifiers in cascade (Viola and Jones 2001) that are trained using the selected features extracted from the face using mRMR, NMIFS, and CMIFS. The number of selected features was in the range 50–500 for image size 24×24, 50–1,000 for 36×36, and 50–2,000 for 48×48. Different implementations of the Adaboost model (LUT, threshold, real, gentle, and modest) were tested (Wu, Ai, and Huang 2003; Vezhnevets 2005; Qahwaji et al. 2008; Ruan et al. 2010).

3. EXPERIMENTS AND RESULTS

The tests were performed in two internationally available face databases. These databases were used to train and test the MI feature selection method as well as the fusion method to allow comparison of results with those previously published. As in Makinen and Raisamo (2008b), faces of one image per person from the Fa and Fb subsets were used and duplications were eliminated. Therefore, 450 female and 450 male normalized images were used from the FERET database. The second database, the WWW image database, contains 2,360 female and 2,360 male images, randomly collected from the World Wide Web and annotated by the researchers (Makinen and Raisamo 2008b). Faces were detected automatically by face Detector OpenCV 1.0 (Open CV 2005).

To compare our results with those of Makinen and Raisamo (2008a; 2008b) two sets of image sizes were used. The first set was composed of 411 images (212 males and 199 females) and three image sizes (24×24, 36×36, and 48×48) from the subset Fa of the FERET database (Makinen and Raisamo 2008a). The second set was composed of 900 images of 24×24 and 760 images of 32×40 from the subset Fa of the FERET database (Makinen and Raisamo 2008b). We used a PC Intel I7 with 4 GB of memory Ram. As in (Makinen and Raisamo 2008a; 2008b), both databases, FERET and WWW, were partitioned to have 80% training data and 20% testing data. All results were obtained with fivefold cross-validation, simulations. For all four models, a training set was used to determine the best number of features that were tested later on the test set for the FERET as well as the WWW databases.

3.1. Results Experiment 1

In Table 1, four different methods for gender classifications were compared. The methods were: a multilayer Perceptron neural network (NN), Support Vector Machine (SVM) with base input, different kind of Adaboost, including Threshold Adaboost (ADA_TH), Gentle_Adaboost (ADA_GENTLE), Real Adaboost (ADA_REAL) and Modest Adaboost (ADA_MOD) and SVM with local binary pattern (SVM_LBP) (Wu, Ai, and Huang 2003; Vezhnevets 2005; Qahwaji et al. 2008; Ruan et al. 2010).

Table 1 compares previously published gender classification results for different image sizes on a subset of the FERET database to our results with feature selection

Table 1. Results of gender classification rates on the FERET database, subset Fa-411 images for different image sizes. The first four rows show classification rates published in the literature for 4 different methods. Rows 5-32 show our results with feature selection mRMR (MID and MIQ), NMIFS, and CMIFS for different classifiers. The number of selected features is shown in parentheses

Methods	Feret 24×24 [%]	Feret 36×36 [%]	Feret 48×48 [%]
SVM [1]	82.64+/−0.593 (576)	86.32+/−4.257 (1296)	84.12+/−4.939 (2304)
SVM_LBP [1]	76.90+/−0.806 (576)	80.00+/−0.931 (1296)	83.01+/−0.618 (2304)
NN [1]	84.73+/−0.505 (576)	86.97+/−1.999 (1296)	81.96+/−5.863 (2304)
ADA_TH [1]	82.35+/−1.107 (576)	84.25+/−0.672 (1296)	83.90+/−1.189 (2304)
SVM_MID	92.54+/−2.196 (350)	93.17+/−1.868 (300)	90.69+/−1.285 (850)
SVM_MIQ	91.72+/−0.831 (300)	91.26+/−2.066 (350)	89.19+/−2.369 (800)
SVM_NMIFS	92.53+/−1.408 (250)	91.89+/−1.203 (900)	89.83+/−0.789 (1000)
SVM_CMIFS	91.19+/−2.982 (450)	92.48+/−1.877 (600)	95.51+/−1.148 (750)
SVM_LBP_MID	90.15+/−1.234 (150)	92.07+/−0.753 (300)	92.10+/−0.763 (350)
SVM_LBP_MIQ	90.20+/−0.736 (150)	91.63+/−0.787 (300)	91.69+/−0.539 (350)
SVM_LBP_NMIFS	91.45+/−0.741 (100)	90.68+/−1.188 (100)	**96.78+/−1.504 (350)**
SVM_LBP_CMIFS	89.84+/−0.990 (200)	90.11+/−1.292 (100)	94.08+/−1.642 (350)
NN_MID	88.52+/−1.080 (300)	89.13+/−0.521 (550)	89.11+/−0.670 (750)
NN_MIQ	86.85+/−0.791 (300)	90.02+/−2.618 (550)	87.32+/−1.019 (800)
NN_NMIFS	89.83+/−0.635 (300)	89.94+/−0.743 (100)	89.49+/−0.961 (800)
NN_CMIFS	90.15+/−0.589 (400)	90.42+/−0.618 (400)	90.04+/−1.197 (800)
ADA_TH_MID	86.21+/−1.107 (250)	89.33+/−0.672 (350)	87.55+/−1.189 (900)
ADA_TH_MIQ	82.28+/−0.514 (250)	87.35+/−1.972 (350)	87.37+/−1.565 (850)
ADA_TH_NMIFS	91.21+/−0.644 (400)	92.18+/−0.984 (800)	92.62+/−1.177 (600)
ADA_TH_CMIFS	90.61+/−0.709 (400)	91.46+/−0.761 (600)	91.41+/−0.931 (800)
ADA_REAL_MID	88.83+/−1.346 (200)	90.04+/−1.558 (250)	86.53+/−1.349 (850)
ADA_REAL_MIQ	87.14+/−0.505 (250)	89.33+/−1.999 (300)	85.74+/−5.863 (900)
ADA_REAL_NMIFS	93.52+/−0.750 (500)	**95.58+/−1.934 (850)**	84.24+/−4.676 (500)
ADA_REAL_CMIFS	90.19+/−1.254 (550)	94.20+/−2.273 (900)	84.92+/−5.893 (200)
ADA_GENTLE_MID	89.85+/−0.806 (200)	89.54+/−3.022 (250)	87.97+/−4.576 (900)
ADA_GENTLE_MIQ	88.56+/−0.662 (300)	88.47+/−3.850 (400)	85.76+/−3.651 (850)
ADA_GENTLE_NMIFS	93.30+/−0.718 (400)	93.28+/−2.046 (400)	93.57+/−1.658 (200)
ADA_GENTLE_CMIFS	93.28+/−0.671 (150)	93.24+/−0.728 (950)	94.34+/−1.583 (200)
ADA_MOD_MID	91.55+/−1.367 (200)	89.53+/−2.211 (250)	86.78+/−5.016 (900)
ADA_MOD_MIQ	90.03+/−0.593 (250)	89.72+/−4.257 (300)	86.27+/−4.939 (1000)
ADA_MOD_NMIFS	93.86+/−0.794 (150)	94.83+/−1.872 (750)	91.70+/−1.324 (350)
ADA_MOD_CMIFS	**94.30+/−0.918 (400)**	93.13+/−1.493 (900)	94.47+/−2.567 (500)

[1]: (Makinen and Raisamo 2008a).

ADA: ADABOOST; TH: Threshold; MOD: Modest; NN: Neural Network; SVM: Support Vector Machine; MID: Mutual information Differential; MIQ: Mutual information Quotient; NMIFS: Normalized mutual information feature selection; CMIFS: Conditional mutual information feature selection.

based on MI and four different classifiers. Results represent the fivefold cross-validation, simulations of the database. The first four rows of Table 1 show the results of the best classification rates published in Makinen and Raisamo (2008a) for classifiers SVM, SVM_LBP, NN, and threshold Adaboost, for three image sizes: 24×24, 36×36, and 48×48. Each column shows the average classification rate for 5 simulations, the standard deviation, and in parenthesis, the number of selected features for each model. Rows 5-32 show the results of the same classifiers but using our proposed feature selection mRMR (MID and MIQ), NMIFS and CMIFS. The best result of

94.3% correct gender classification on the FERET database, 24×24 face size, was obtained for the Modest Adaboost with CMIFS feature selection and 400 features. This result was 9.6% higher than the best result previously published with 576 features. The best result of 95.6% correct gender classification for the FERET database, 36×36 face size, was obtained for the Real Adaboost model with NMIFS feature selection and 850 features. This result was 8.6% higher than those previously published with 1296 features using an SVM model. In the case of the FERET database, 48×48 face size, the highest result was 96.78% reached by the SVM_LBP model using NMIFS feature selection and 350 features. This result was 12.7% higher than those previously published for the SVM model and 2,304 features. In summary, for all image sizes, the classification rate is significantly better with CMIFS and NMIFS feature selection than the best results previously published in the literature and with significantly fewer numbers of features.

3.2. Results Experiment 2

In Table 2, four different methods for gender classifications were compared. The methods were: a multilayer Perceptron neural network (NN), Support Vector Machine (SVM) with base input, Look-up Table Adaboost (ADA_TH), and SVM with local binary pattern (SVM_LBP). These classifiers had inputs from three different feature selection methods (mRMR-MID, mRMR-MIQ NMIFS, CMIFS). Table 2 shows the best results on gender classification published to date for the FERET and WWW databases for two different face sizes, 24×24 and 32×40. Results are the average of 5 simulations with different partitions of the database. The first four rows of Table 2 show the results published in Makinen and Raisamo (2008b) for SVM, SVM_LBP, NN, and LUT Adaboost classifiers. Rows 5-20 show the same classifiers but using our proposed feature selection mRMR (MID and MIQ), NMIFS, and CMIFS. Each column shows the average classification rate for five simulations, the standard deviation, and in parenthesis, the number of selected features for each model.

The best result of 94.08% correct gender classification rate on the FERET database, 24×24 face size, was obtained for the SVM model with NMIFS feature selection and 400 features. This result was 2.3% higher than the best result previously published with 576 features (third row of Table 2). The best result of 94.41% correct gender classification for the FERET database, 32×40 face size, was obtained for the SVM model with NMIFS feature selection and 950 features. This result was 1.2% higher than those previously published with 1,280 features using a LUT Adaboost model.

The best result of 83.86% correct gender classification rate on the WWW database, 24×24 face size, was obtained for the SVM_LBP model with MID feature selection and 150 features. This result was 4.1% higher than the best result previously published with 576 features (first row of Table 2). The best result of 86% correct gender classification for the WWW database, 32×40 face size, was obtained for the SVM_LBP model with NMIFS feature selection and 150 features. This result was 8.9% higher than those previously published with 1,280 features using an SVM_LBP model. In summary, our proposed feature selection reduced the number of features significantly and improved the classification improvement by 8.9% on the WWW database compared to the best results published previously in the literature.

Table 2. Results of gender classification rates on the FERET database, subset Fa and WWW database for two image sizes. The first four rows show classification rates published in the literature for 4 different methods. Rows 5-20 show our results with feature selection mRMR (MID and MIQ), NMIFS, and CMIFS. The number of features is shown in parentheses

Methods	Feret 24 × 24 [%]	Feret 32 × 40 [%]	WWW 24 × 24 [%]	WWW 32 × 40 [%]
SVM [2]	87.15+/−0.102 (576)	81.29+/−0.111 (1240)	79.74+/−0.077 (576)	75.77+/−0.077 (1240)
SVM_LBP[2]	81.12+/−0.155 (576)	91.80+/−0.107 (1240)	73.84+/−0.095 (576)	77.07+/−0.086 (1240)
NN [2]	91.79+/−0.107 (576)	90.16+/−0.070 (1240)	72.61+/−0.066 (576)	61.94+/−0.070 (1240)
LUT_ADA[2]	89.89+/−0.117 (576)	93.24+/−0.050 (1240)	74.47+/−0.096 (576)	76.63+/−0.084 (1240)
SVM_MID	94.00+/−0.006 (200)	94.26+/−0.020 (800)	81.09+/−0.036 (400)	79.89+/−0.017 (500)
SVM_MIQ	93.30+/−0.018 (250)	93.86+/−0.011 (900)	79.74+/−0.026 (400)	79.22+/−0.021 (500)
SVM_NMIFS	**94.08+/−0.022 (400)**	**94.41+/−0.015 (950)**	79.95+/−0.014 (400)	80.33+/−0.045 (900)
SVM_CMIFS	93.26+/−0.023 (350)	93.22+/−0.011 (850)	81.83+/−0.036 (550)	76.67+/−0.046 (650)
SVM_LBP_MID	89.69+/−0.012 (150)	92.68+/−0.012 (300)	**83.86+/−0.027 (150)**	83.09+/−0.019 (300)
SVM_LBP_MIQ	90.27+/−0.011 (150)	92.36+/−0.011 (300)	82.40+/−0.026 (200)	81.15+/−0.018 (300)
SVM_LBP_NMIFS	86.28+/−0.007 (150)	90.59+/−0.011 (300)	78.71+/−0.019 (200)	**86.00+/−0.017 (150)**
SVM_LBP_CMIFS	91.00+/−0.015 (150)	92.28+/−0.010 (300)	79.39+/−0.025 (200)	80.42+/−0.013 (300)
NN_MID	91.57+/−0.551 (200)	91.22+/−0.101 (250)	79.11+/−0.080 (300)	70.43+/−0.045 (500)
NN_MIQ	89.52+/−0.627 (250)	90.29+/−0.099 (300)	78.37+/−0.081 (350)	70.71+/−0.102 (500)
NN_NMIFS	89.39+/−0.542 (450)	90.47+/−0.121 (450)	79.83+/−0.091 (450)	80.26+/−0.051 (450)
NN_CMIFS	89.52+/−0.839 (400)	91.17+/−0.105 (400)	75.12+/−0.106 (450)	77.05+/−0.050 (500)
ADA_LUT_MID	89.51+/−0.802 (400)	92.87+/−0.143 (350)	78.17+/−0.053 (400)	79.40+/−0.067 (350)
ADA_LUT_MIQ	89.40+/−1.116 (450)	90.67+/−0.103 (450)	77.10+/−0.097 (400)	78.36+/−0.070 (500)
ADA_LUT_NMIFS	90.41+/−0.797 (350)	91.43+/−0.134 (500)	77.03+/−0.090 (350)	77.11+/−0.057 (500)
ADA_LUT_CMIFS	90.05+/−1.069 (350)	91.26+/−0.087 (500)	77.21+/−0.066 (350)	76.49+/−0.054 (500)

[2]: (Makinen and Raisamo 2008b).

ADA: ADABOOST; LUT: Look-up Table; NN: Neural Network; SVM: Support Vector Machine; MID: Mutual information Differential; MIQ: Mutual information Quotient; NMIFS: Normalized mutual information feature selection; CMIFS: Conditional mutual information feature selection.

3.3. Result Experiment 3

Table 3 compares previously published gender classification results for different image sizes on a subset of the FERET database to our results with feature selection based on MI and SVM classifiers. Results represent the average of 5 cross-validations simulations. The first column show results of the method, the second row shows the vector size for fused features, and the third column shows the results of the best classification rates published in Alexandre (2010) for classifiers SVM, for three image sizes 20×20, 36×36, and 128×128. Each row shows the average classification rate for five simulations and in parenthesis, the number of selected features for each model.

L1, L2, and L3 represent fusion of intensities, shape and texture for three different sizes, 20×20, 36×36, and 128×128. L4, L5, and L6 represent the fusion of three different features (intensity, shape, and texture) for 20×20, 36×36, and 128×128, respectively. L7 represents the fusion of three scales and three types of features; the total number of inputs was increased nearly nine-fold. Columns 4-7 show the results of the same classifier but using our proposed feature selection mRMR, NMIFS, and CMIFS. Figure 4 show the features selected on the face by the feature selection methods L3 on images of 20×20, 30×30, and 128×128. In the case of L1, intensity levels were used, in L2 the histogram distribution of shapes is used and in L3 the histogram distribution of textures is used. The intensity of the color in L2 and L3 determine the number of bins that were selected in this area. The results obtained in the FERET database with our method are better than those published in Alexandre (2010) which is shown in Table 3.

In Table 3 it can be observed in L1 that the feature selection method, CMIFS reaches the best classification performance with 93.82% and 14800 features. In L2, the best feature selection method was CMIFS with 450 features, equivalent to 50% size of the original vector. In L3 the best method was mRMR with 98.76% and 1,200 selected features, which is equivalent to 1.7% of the original vector. In L4 and L5 the best results were 92.59% and 95.06% with CMIFS method selecting 800 and 5,650 features. These are equivalent to a reduction to 41% and 52% of the original vectors, respectively. In L6 the best feature selection method reached

Table 3. Results of gender classification rates on the FERET database, for fusion of different features. The third column show classification rates published in the literature for SVM and different feature fusion. Columns 4-6 show our results with the feature selection methods mRMR, NMIFS, and CMIFS. The number of selected features is shown in parentheses

Fusion	Vector	FERET (%) [3]	mRMR (%)	NMIFS (%)	CMIFS (%)
Intensity (L1)	18,000	95.33	92.59 (9800)	93.82 (14,800)	93.82 (14,800)
Shape (L2)	7,100	96.26	86.41 (2050)	81.48 (2,050)	89.95 (450)
Texture (L3)	20,945	93.46	**98.76 (1200)**	**95.06 (6,700)**	**97.53 (750)**
20×20 (L4)	1,831	85.98	91.35 (850)	91.35 (500)	92.59 (800)
36×36 (L5)	10,855	91.59	93.82 (7650)	93.82 (7,650)	95.06 (5,650)
128×128 (L6)	34,159	95.33	95.06 (1000)	91.35 (9,950)	97.53 (11,000)
All (L7)	46,845	99.07	96.30 (19700)	95.06 (33,400)	97.53 (33,450)
Best_fea			**99.13 (33,800)**		

[3]: (Alexandre 2010).

97.53% correct gender classification with CMIFS. The number of selected features was 11,000 features which are only 33% of the original vector size. Two tests were performed in L7, the first is the union of features for each of the selection methods from L1 to L6 (i.e., the union of mRMR for L1 to L6), named "All-L7" in Table 3. The method reaches 97.53% with 11,700 features of the total 46,845 which is equivalent to 25% of the original vector. The second test named Best_fea, represents the union of the best features of each (L1 to L7), for each feature selection method. The sum of all characteristics is 33,800 with a classifications rate of 99.13%.

3.4. Computational Time

In Table 4 it is shown the computational time for classification of the best methods for different feature selection on the FERET database with 24×24 and 32×40 aligned images for the WWW database with 24×24 and 32×40 misaligned images. Table 4 compares the computational time for the classifiers: SVM, SVM_LBP, NN, and Adaboost using raw data compared to the feature selection

Table 4. Results to test cross database performance for gender classification performed on the databases LFW, MORPH-II with parameters selected from best results from the FERET and WWW. Column 1 shows the best classifiers obtained for each test. Columns 2-5 show the results for each of the databases and standard deviation

	LFW	MORPH-II	WWW	FERET
FERET 24×24 **ADA-MOD-CMIFS** BEST TABLE 1	90.10+/−0.125	91.70+/−0.025	81.00+/−0.058	94.30+/−0.918
FERET 36×36 **ADA-REAL-NMIFS** BEST TABLE 1	92.25+/−0.070	92.15 + / 0.057	83.05+/−0.023	95.58+/−1.934
FERET 48×48 **SVM-LBP-NMIFS** BEST TABLE 1	91.78+/−0.980	93.01+/−0.990	84.05+/−1.001	96.78+/−1.504
FERET-24×24 **SVM-NMIFS** BEST TABLE 2	90.25+/−0.102	92.03+/−0.055	80.00+/−0.078	94.08+/−0.022
FERET-32×40 **SVM-NMIFS** BEST TABLE 2	92.00+/−0.250	93.01+/−0.570	81.05+/−0.098	94.41+/−0.015
WWW-24×24 **SVM-LBP-MID** BEST TABLE 2	89.35+/−0.120	91.05+/−0.032	83.86+/−0.027	85.01+/−1.010
WWW-32×40 **SVM-LBP-NMIFS** BEST TABLE 2	91.05+/−0.101	92.00+/−0.212	86.00+/−0.017	86.25+/−0.976
FERET **BEST_FEA** BEST TABLE 3	**93.60+/−0.125**	**94.00+/−0.033**	**90.03+/−0.225**	**99.13 + /0.015**

ADA: ADABOOST; LUT: Look-up Table; NN: Neural Network; SVM: Support Vector Machine; MID: Mutual information Differential; MIQ: Mutual information Quotient; NMIFS: Normalized mutual information feature selection; CMIFS: Conditional mutual information feature selection.

methods. Besides the improvements in classification performance results, our proposed feature selection method reduces the number of features from 576 (24×24) and 1,280 (32×40) to up to 400 in the FERET database and to 150 features in the WWW database. Therefore, computational time can be significantly reduced for real-time implementations to 69.4% in the FERET 24×24 face size and to 74.2% in the FERET 32×40 face size cases. In the case of the WWW database the ratio can be reduced to 26% for 24×24 face size and to 11.7% for 32×40 face size. Computational time could be of significant commercial interest if the gender classification method is applied in real-time, for example in electronic product advertisement at retail stores. Therefore, feature selection is highly desirable.

Table 5 shows the computational time for the best result in experiment 2 with 1,200 features of 20,945 and the feature selection method, mRMR, NMIFS, CMIFS for L3 on the FERET database.

3.5. Results Analysis

In Figures 7–10 are shown the selected features for all different methods. The marks in the figures represent those pixels or area of the face that improves the gender classification for different classifiers. The feature selection methods measure both the highest mutual information features and those that are more distant from each class. The features selected with mutual information methods allow the improvement of gender classification when the selected features are used in the four types of classifiers; a NN and an SVM with pixel-based inputs, an Adaboost, and an SVM with Local Binary Pattern (LBP) features as inputs. It is important to note that the selected features are common to both genders, however allow making a difference between both classes. The feature selection methods determined the smallest optimal subset of features that maximize the rate of classification. If we increase the number of features the classification performance decreases because the new features may introduce noisy information which is detrimental for the classifier performance.

Figures 7–9 show examples of the selected features for the best results obtained by the MID, NMIFS, and CMIFS methods. Figures 7–9 show two images (male and female) from the FERET database and two images from the WWW database. Figure 7 shows the 300 features selected by the MID method for the FERET and WWW databases. Figure 8 shows 950 selected features for the NMIFS for the FERET database and 150 features selected for the WWW database. Figure 9 shows the CMIFS feature selection for the 850 features for the FERET database and 300 features selected for the WWW database.

Figure 10 show examples of selected features for the best result obtained for experiment 2, with feature selection method, mRMR. Figure 10 show 2 images (male and female) from FERET database, this represents the 1,200 features selected from the fusion histogram of textures (LBP) and distributed in three images. The square shows the selected area where the intensity toward black represents the number of bins that were selected in this area. If the area was not selected no square is shown.

Figure 6 shows the best results for experiment 2. Figure 5 shows the correct gender classification rate as a function of the number of features selected with three

Table 5. Results of the computational time for the best results for different feature selection methods on the FERET database with 24×24 and 32×40 aligned images and for the WWW database with 24×24 and 32×40 misaligned images. Time-Class is the computational time employed by all images and the last column shows the computational time per image. Raw-Data includes all features

FERET 24×24						
Classifier	Result (%)	Methods	N° Images	N° Features	time-Class(sec)	time-per Image
SVM	94.08	NMIFS	180	400	0.2401	0.00060
SVM-LBP	91.00	CMIFS	180	150	0.2166	0.00144
NN	91.57	CMIFS	180	200	0.1945	0.00097
ADA-LUT	90.41	NMIFS	180	350	0.2245	0.00064
SVM	87.15	Raw-Data	180	576	0.4975	0.00086
FERET 32×40						
Classifier	Result (%)	Methods	N° Images	N° Features	time-Class(sec)	time-per Image
SVM	94.41	NMIFS	152	950	0.61	0.00401
SVM_LBP	92.68	MID	152	300	0.5	0.00328
NN	93.22	MID	152	250	0.67	0.00440
ADA-LUT	92.87	MID	152	350	0.59	0.00388
SVM	81.29	Raw-Data	152	1,240	1.19	0.00782
WWW 24×24						
Classifier	Result (%)	Methods	N° Images	N° Features	time-Class(sec)	time-per Image
SVM	78.71	CMIFS	944	550	0.31	0.00032
SVM_LBP	83.86	MID	944	150	0.183	0.00019
NN	79.83	NMIFS	944	450	0.294	0.00031
ADA-LUT	82	CMIFS	944	500	0.34	0.00036
SVM	79.74	Raw-Data	944	576	0.76	0.00081
WWW 32×40						
Classifier	Result (%)	Methods	N° Images	N° Features	time-Class(sec)	time-per Image
SVM	80.33	NMIFS	762	900	1.58	0.00207
SVM_LBP	83.09	MID	762	300	1.31	0.00172
NN	80.26	NMIFS	762	450	1.35	0.00177
ADA-LUT	79.40	MID	762	350	0.98	0.00128
SVM	75.77	Raw-Data	762	1280	2.41	0.00316

ADA: ADABOOST; LUT: Look-up Table; NN: Neural Network; SVM: Support Vector Machine; MID: Mutual information Differential; MIQ: Mutual information Quotient; NMIFS: Normalized mutual information feature selection; CMIFS: Conditional mutual information feature selection.

MI selection methods. Figure 6(*a*) is for mRMR, (*b*) for CMIFS and (*c*) for NMIFS. The best gender classification rate was reached with 1,200 features.

We measured the computational time employed with the selected features in gender classification and is reported in Tables 4–5. This time can be compared to the computational time without feature selection. Our experiments were performed on a 2.5 GHz I7 PC with 4 GB of memory using Matlab. In Table 5, it can be observed that in the FERET database with 24×24 size the shortest computational

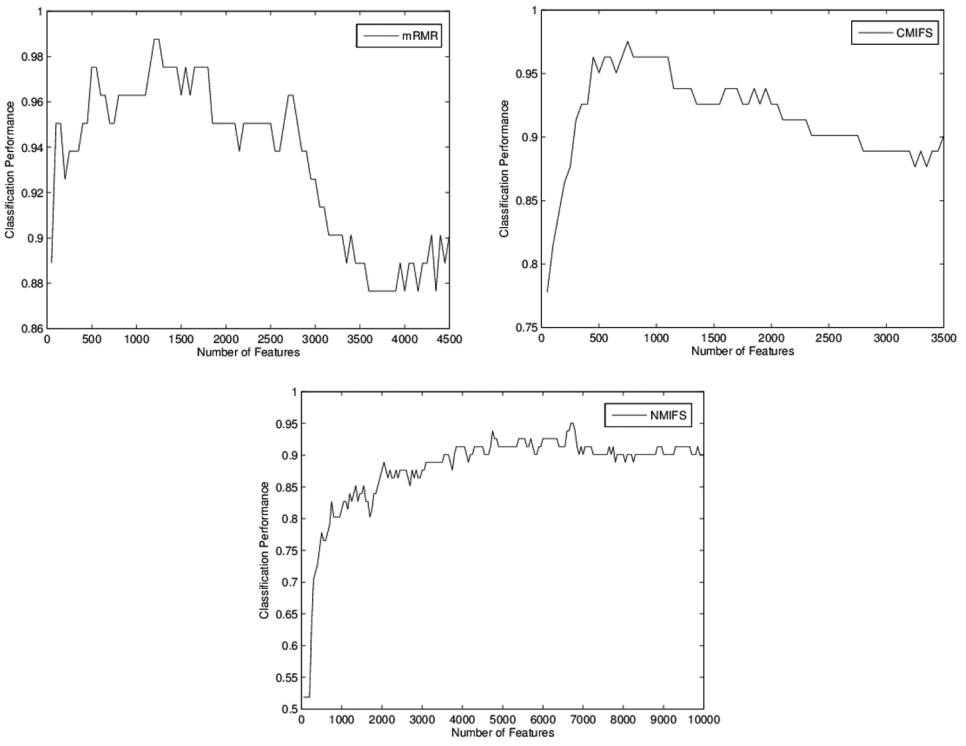

Figure 6. Classification performance as a function of the number of features for experiment 2. The feature selection methods (*a*) mRMR, (*b*) CMIFS, and (*c*) NMIFS are compared for the feature fusion L3. The best result was reached by mRMR with 1,200 features.

employed in gender classification for 180 test images was using a NN with 0.19 s for all images and 0.97 ms per image when using 200 features selected by CMIFS. This is 50% of the time employed without feature selection. In the case of the FERET database with image size of 32×40, the computational time employed in gender classification for 152 test images using SVM-LBP classifier was 0.5 s, and 3.2 per image using 300 features selected by mRMR-MID. This is a 58% reduction in computational time. For images in the WWW database of sizes 24×24, the shortest computational time employed in gender classification for 944 test images was 183 with SVM-LBP classifier and 0.19 per image with 150 features selected by mRMR-MID. This represents 23% of the computational time employed without feature selection. For the WWW database and image size of 32×40 the shortest computational time employed in 762 test images was ADA-LUT with 0.98 s and 1.2 per images with 350 features selected by mRMR-MID. This represents a 62% reduction in computational time when feature selection is used. Table 6 shows the computational time employed for L3 in gender classification when the selected features are fused. The computational time for the best feature selection method –mRMR- with 1,200 features was 1.6 per image. Computational time shown in Tables 5 and 6 can be further improved in the future by implementing the methods in C or by parallel computations.

Figure 7. Two original images (left), male and female, from the FERET database (Fa), size 32×40 with 300 features selected for gender classification with MID-LBP and two original images (right), male and female, from the WWW database, size 32×40 with 300 features selected for gender classification with MID are shown in the bottom row.

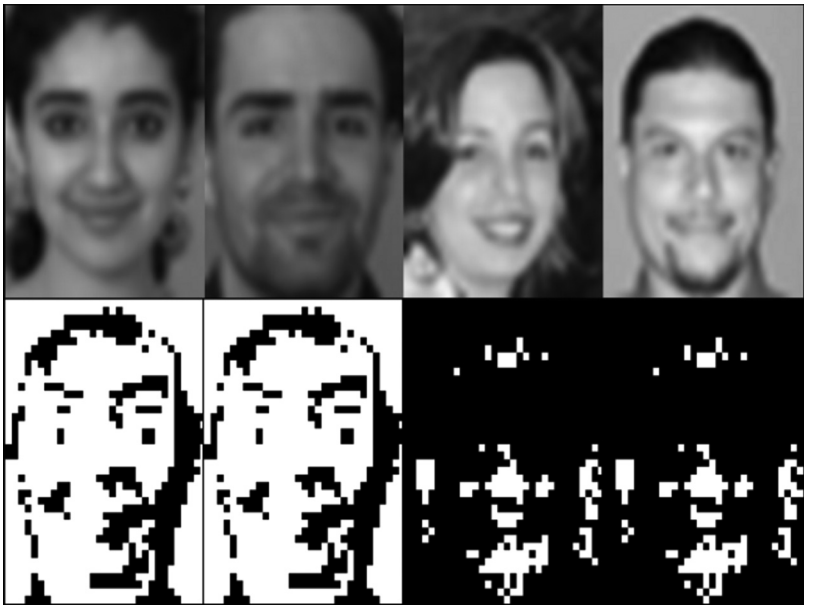

Figure 8. Two original images (left), male and female, from the FERET database (Fa), size 32×40 with 950 features selected for gender classification with NMIFS and two original images (right), male and female, from the WWW database, size 32×40 with 150 features selected with NMIFS-LBP for gender classification are shown in the bottom row.

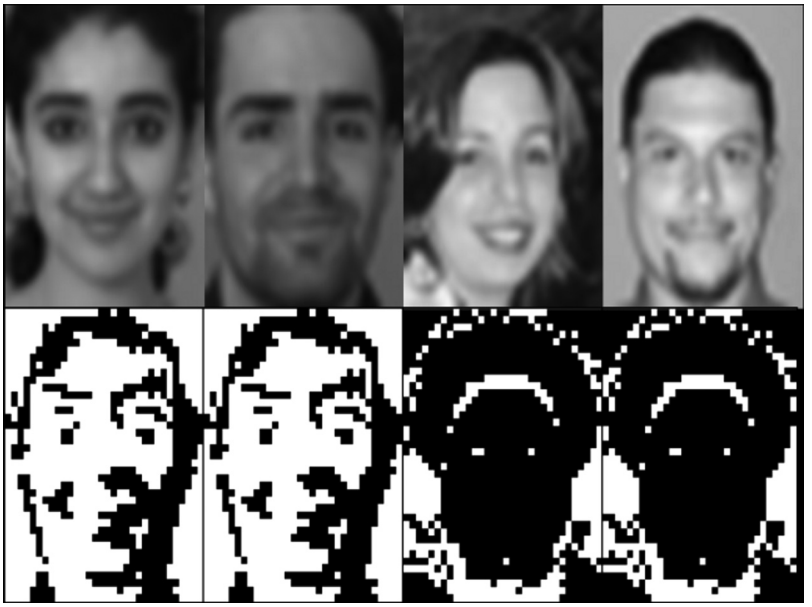

Figure 9. Two original images (left), male and female, from the FERET database (Fa), size 32×40 with 850 features selected for gender classification with CMIFS and two original images (right), male and female, from the WWW database, size 32×40 with 300 features selected with CMIFS-LBP for gender classification are shown in the bottom row.

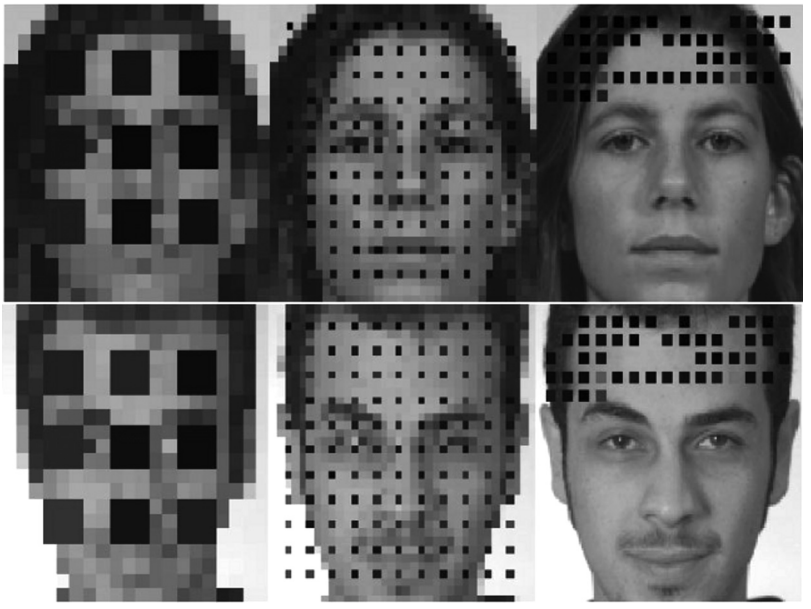

Figure 10. Original Images, male and female, from the FERET database. Feature fusion L3 with 1,200 features selected using mRMR which reached the best results for experiment 2. The fusion considered three scales for image sizes: 20×20, 36×36, and 128×128. The squares show the selected features at each of the three scales.

Table 6. Computational time required to classify one image from the FERET database with L3 feature fusion in experiment 2. The best classification performance was reached with 1200 features selected by the three feature selection methods; mRMR, NMIFS, and CMIFS

Number of features	mRMR [s]	NMIFS [s]	CMIFS [s]
100	0.00083	0.00341	0.00536
200	0.00057	0.00341	0.00568
300	0.00088	0.00463	0.00601
400	0.00073	0.00479	0.00609
500	0.00129	0.00512	0.00625
600	0.00099	0.00503	0.00633
700	0.00150	0.00520	0.00634
800	0.00112	0.00524	0.00635
900	0.00144	0.00528	0.00636
1,000	0.00108	0.00536	0.00635
1,100	0.00183	0.00544	0.00633
1,200	**0.00166**	0.00552	0.00617
1,300	0.00207	0.00552	0.00617
1,400	0.00208	0.00544	0.00619
1,500	0.00171	0.00552	0.00619
1,600	0.00182	0.00544	0.00620
1,700	0.00246	0.00527	0.00620
1,800	0.00213	0.00552	0.00621
1,900	0.00249	0.00560	0.00621
2,000	0.00207	0.00576	0.00629

3.6. Statistical Analysis

We use the ANOVA (analysis of variance) multi-comparison test to determine whether or not the differences among results were statistically significant. The p-value indicates that differences between means are highly significant, e.g., $p < 0.05$ (Bradstreet 2006). We compare the results of the classifiers without feature selection and the results of the classifiers with feature selection using mutual information. In Table 1, for FERET database and image sizes of 24×24, 36×36, 48×48 the ANOVA showed that the following seven classifiers have means significantly different from those of the SVM with raw data: SVM-NMIFS, SVM-CMIFS, SVM-LBP, ADA-GENTLE-NMIFS, ADA_GENTLE-CMIFS, ADA-MOD-NMIFS, and ADA-MOD-CMIFS. In all cases p was lower than 1.27e-5 ($p < 0.05$) which is highly statistically significant. The best result was obtained by ADABOOST-MODEST-CMIFS with 94.30% $+/-0.918$ and 400 features. This case was the best in comparison to the four classifiers without feature selection (Table 1, lines 1 to 4).

In Table 2, for the FERET database we also compared the results of all classifiers with no feature selection versus those of the same classifiers with feature selection using mutual information. The ANOVA multi-comparison test yielded a $p = 0.047$ ($p < 0.05$). The best method was the feature selection using NMIFS with SVM classifier with 400 features. All variants of SVM with feature selection methods (MID, MIQ, NMIFS, and CMIFS) showed significantly different results than those obtained by the same classifiers without feature selection (Table 2, lines 1 to 4).

In Table 2, for WWW database we also obtained statistical significant differences when comparing the results of the classifiers without feature selection versus those with feature selection (MID, MIQ, NMIFS, and CMIFS). The ANOVA multi-comparison test yielded a $p = 0.0158$, ($p < 0.05$). Although statistically significant, the NN classifier with raw data yielded the lowest classification results. The best results were obtained with the SVM-LBP-MID yielding 86.00% $+/-0.017$ and 150 selected features.

After analyzing the results, it can be concluded that feature selection improved significantly the performance of gender classification. The quality of the face images in the FERET database, which is much better than those of the WWW database, may account for the improved results obtained in the FERET database compared to those results obtained in the WWW database. The results also show that selected features from the WWW database generalize quite well on the images of the LFW and MORPH-II databases.

From the results it can be inferred that feature selection improves significantly the performance of gender classification. There are differences between the FERET and the WWW databases. The FERET database shows better image quality relative to the WWW database. It is also shown that selected features from the WWW database generalize quite well on the images of the databases LFW and MORPH-II. In contrast, FERET has images from controlled environments and a smaller number in the training y testing sets. After performing a cross check among databases, the selected features as well as fusion of features with MI generalize quite well.

4. CONCLUSIONS

In this article we report for the first time the use of feature selection based on MI and fusion of three methods of features for gender classification. Starting from a large number of features (all information of image) extracted from input data the best subset of relevant features which contains only the useful information for distinguishing one class from the other. This allows the representation of the data in a lower dimensional space, and classification in less time. This is significantly better than all previously published, reducing time of classifiers input on the same databases. Computational time could be of significant commercial interest, if the gender classification is applied in real time later on stage of face detection. Therefore feature selection methods and fusion of features are highly desirable.

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. Feature selection has the additional improvement of reducing computational time which is central for many real-time applications. 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 the case 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 up to date in standard databases.

A new method for gender classification from faces is proposed using MI to select facial features. Two forms of combining relevance and redundancy using MID

and MIQ were employed (mRMR) as well as NMIFS, CMIFS measures. Our results show that gender classification results can be significantly improved, up to 12.7%, by feature selection. Reducing the number of features has the additional benefit of reducing the required number of computations, making implementation of the method in real-time possible. The best gender classification performance for experiment 1 was obtained with NMIFS and showed improvement of 12.7% on the FERET database and 11.7% on the WWW database compared to previous publications.

In experiment 2 the best performance was obtained with the fusion of features (Best_Fea) with a classification rate of 99.13%. This is the best of all result reported for gender classification on the FERET database.

Another important result of the feature selection method is that, depending on the image size, the total number of features selected was reduced to at least 74.2% on the FERET database and to 26.04% on the WWW database. Therefore, computational time is significantly reduced, which makes real-time applications of gender classification feasible.

ACKNOWLEDGMENTS

This research was funded by FONDECYT, grant No. 1080593 by FONDEF D08I1060 and the Department of Electrical Engineering, Universidad de Chile.

REFERENCES

Akadi, A. El., A. Amine, A. El Ouardighi, and D. Aboutajdine. 2009. A new gene selection approach based on Minimum Redundancy-Maximum Relevance (MRMR) and Genetic Algorithm (GA). *Proceedings of the IEEE ACS Conf. Computer System and Applications*. 69–75.

Alexandre, Luís A. 2010. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters* 31 (11): 1422–1427.

Bekios-Calfa, J., J. M. Buenaposada, L. Baumela. 2011. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (4): 858–869.

Bradstreet, T. E. 2006. *Journal of the American Statistical Association* 101 (474): 848–849.

Brunelli, R. and T. Poggio. 1995. HyberBF networks for gender classification. *Proc. DARPA Image Understanding Workshop* 311–314.

Cheng, Hongrong, Zhiguang Qin, Weizhong Qian, and Wei Liu. 2008. Conditional mutual information based feature selection. *Proc. Int. Symp. Knowledge Acquisition and Modeling* 103–107.

Chow, T. W. S. and D. Huang. 2005. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural Networks* 16 (1): 213–224.

Chu, W.-S., C. R. Huang, and C. S. Chen. 2010. Identifying gender from unaligned facial images by set classification. *Proc. 20th Int. Conf. on Pattern Recognition (ICPR)* 2636–2639.

Dago-Casas, P., D. Gonzalez-Jimenez, L. Long-Yu, J. L. Alba-Castro. 2011. Single and cross- database benchmarks for gender classification under unconstrained settings. *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies* 2152–2159.

Ding, C. and H. Peng. 2003. Minimum redundancy feature selection from microarray gene expression data. *IEEE Bioinformatics Conf.* 523–528.

Estévez, P. A., M. Tesmer, C. A. Perez, and J. M. Zurada. 2009. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* 20 (2): 189–201.

Huang, J., Y. Cai, and X. Xu. 2006. A filter approach to feature selection based on mutual information. *Proc. 5th IEEE Int. Conf. Cognitive Informatics ICCI* 84–89.

Huang G. B., M. Ramesh, T. Berg, and E. Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amberst, *Tech. Rep. 07* 49, October 2007.

Irick, K., M. DeBole, V. Narayanan, R. Sharma, Hankyu Moon, and S. Mummareddy. 2007. A unified streaming architecture for real time face detection and gender classification. *Proc. Int. Conf. Field Programmable Logic and Applications* 267–272.

Jun, B., T. Kim, and D. Kim. 2011. A compact local binary pattern using maximun of mutual information for face analysis. *Pattern Recognition* 44 (3): 532–543.

Lu, L., Z. Xu, and P. Shi. 2009. Gender classification of facial images based on multiple facial regions. *World Congress on Computer Science and Information Engineering* 6: 48–52.

Makinen, E. and R. Raisamo. 2008a. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (3): 541–547.

Makinen, E. and R. Raisamo. 2008b. An experimental comparison of gender classifications methods. *Pattern Recognition Letters* 29 (10): 1544–1556.

Mayo, M. and E. Zhang. 2008. Improving Face gender classification by adding deliberately misaligned faces to the training data. *Proc. 23rd Int. Conf. Image and Vision Computing NZ, IVCNZ08* 1–5.

Moghaddam, Baback and Yang Ming-Hsuan. 2002. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (5): 707–711.

Ojala, T., M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (7): 971–987.

Peng, Hanchuan, Fuhui Long, and C. Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1226–1238.

Perez, C. A., V. A. Lazcano, P. A. Estevez, and C. M. Estevez. 2004. Real-time iris detection on faces with coronal axis rotation. *Proc. IEEE International Conference Systems, Man and Cybernetics* 7: 6389–6394.

Perez, C. A., G. D. Gonzalez, L. E. Medina, and F. J. Galdames. 2005. Linear versus nonlinear neural modeling for 2-D pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 35 (6): 955–964.

Perez, C. A., V. A. Lazcano, and P. A. Estévez. 2007. Real-time iris detection on coronal-axis-rotated faces. *IEEE Transactions on Systems, Man and Cybernetics – Part C-Applications and Reviews* 37 (5): 971–978.

Perez, C. A., L. E. Castillo, L. A. Cament, P. A. Estevez, and C. M. Held. 2010a. Genetic optimisation of illumination compensation methods in cascade for face recognition. *Electronics Letters* 46 (7): 498–500.

Perez, C. A., C. M. Aravena, J. I. Vallejos, P. A. Estevez, and C. M. Held. 2010b. Face and iris localization using templates designed by particle swarm optimization. *Pattern Recognition Letters* 31 (9): 857–868.

Perez, C. A., L. A. Cament, and L. E. Castillo. 2011. Methodological improvement on local Gabor face recognition based on feature selection and enhanced Borda count. *Pattern Recognition* 44 (4): 951–963.

Phillips, P. J., Hyeonjoon Moon, P. Rauss, and S. A. Rizvi. 1997. The FERET evaluation methodology for face-recognition algorithms. *Proc. of the IEEE Computer Society Conf. Computer Vision and Pattern Recognition Conference* 137–143.

Qahwaji, R., M. Al-Omari, T. Colak, and S. Ipson. 2008. Using the real, gentle and modest AdaBoost learning algorithms to investigate the computerised associations between Coronal Mass Ejections and filaments. *Proc. Mosharaka Int. Conf. Communications, Computer and Applications* 37–42.

Ricanek K. Jr., and T. Tesafaye. 2006. MORPH: A longitudinal image database of normal adult age-progression. *Proc. IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR)* 341–345.

Ruan, Chengxiong, Qiuqi Ruan, and Xiaoli Li. 2010. Real Adaboost feature selection for face recognition. *IEEE Int Signal Processing* 1402–1405.

Shakhnarovich, G., P. A. Viola, and B. Moghaddam. 2002. A unified learning framework for real time face detection and classification. *Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition* 14–21.

Shan C. 2012. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters* 33 (4): 431–437.

Sun, Z., G. Bebis, and R. Miller. 2004. Object detection using feature subset selection. *Pattern Recognition* 37 (11): 2165–2176.

Vankayalapati, H. D., L. N. P. Boggavarapu, R. S. Vaddi, and K. R. Anne. 2011. Extraction of facial features for the real-time human gender classification. *Proc. Int. Conf. Emerging Trends in Electrical and Computer Technology* 752–757.

Vezhnevets, A. and V. Vezhnevets. 2005. Modest Adaboost- teaching Adabost to generalize better. *Proc. Graphicon* 1–12.

Vinh, L. T., N. D. Thang, and Y.-K. Lee. 2010. An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. *Proc. 10th Int. Symp. IEEE/IPSJ Applications and Internet (SAINT)* 395–398.

Viola, P. and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. *Proc. 2001 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)* 511–518.

Wang, Y., H. Ai, B. Wu, and C. Huang. 2004. Real time facial expression recognition with AdaBoost. *Proc. 17th Int. Conf. Pattern Recognition ICPR* 3: 926–929.

Wu, B., H. Ai, and C. Huang. 2003. Lut-based adaboost for gender classification. *Proc. 4th Int. Conf. on Audio and Video Based Biometrics Person Authentication* 104–110.

Wu, Jing, William A. P. Smith, and Edwin R. Hancock. 2010. Facial gender classification using shape-from-shading. *Image and Vision Computing* 28 (6): 1039–1048.