Expert Systems with Applications 42 (2015) 2772–2784

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Boosting gender recognition performance with a fuzzy inference system

Taner Danisman *, Ioan Marius Bilasco, Jean Martinet

Lille 1 University Science and Technology, IRCICA, Parc Scientifique de la Haute Borne, 50 Avenue Halley, 59655 Villeneuve d'Ascq, France

ARTICLE INFO

Article history:
Available online 20 November 2014

Keywords:
Gender recognition
Fuzzy inference system
Fuzzy rules
Cross-database tests

ABSTRACT

In this paper, we propose a novel gender recognition framework based on a fuzzy inference system (FIS). Our main objective is to study the gain brought by FIS in presence of various visual sensors (e.g., hair, mustache, inner face). We use inner and outer facial features to extract input variables. First, we define the fuzzy statements and then we generate a knowledge base composed of a set of rules over the linguistic variables including hair volume, mustache and a vision-sensor. Hair volume and mustache information are obtained from Part Labels subset of Labeled Faces in the Wild (LFW) database and vision-sensor is obtained from a pixel-intensity based SVM + RBF classifier trained on different databases including Feret, Groups and GENKI-4K. Cross-database test experiments on LFW database showed that the proposed method provides better accuracy than optimized SVM + RBF only classification. We also showed that FIS increases the inter-class variability by decreasing false negatives (FN) and false positives (FP) using expert knowledge. Our experimental results yield an average accuracy of 93.35% using Groups/LFW test, while the SVM performance baseline yields 91.25% accuracy.

1. Introduction

Gender recognition is a challenging two-class classification task in computer vision to identify female and male faces. Visual gender recognition is a key component of demographic studies and focuses on gender, age and ethnicity analysis for targeted advertisement, electronic marketing, biometrics, and Human Computer Interaction. Gender recognition studies rely on different disciplinary fields using text, speech, image and video. Considering the visual gender recognition, the gender can be recognized from video, 2D images (e.g., color and intensity images), 2.5D images (e.g., RGB-D depth images) and 3D images. Besides that, there are studies relying on the whole body and gait sequences. However, in the literature, the main approach to gender recognition is 2D facial analysis. Therefore, our literature review focuses on 2D facial gender recognition (FGR) and FGR related studies.

FGR is not a trivial task and it holds known challenges (e.g., illumination, head-pose changes, and occlusions) of other face based pattern recognition problems. There are multiple factors that effect the FGR process. First groups of factors are created by the human such as head-pose changes, aging, make-up, ethnicity, accessories, occlusions, facial hair and expressions. The second group of factors is usually the external factors such as lighting, illumination conditions, camera resolution and perspective. Large intra-class variations in female and male subjects also bring further difficulties. In the literature, different preprocessing, normalization, feature extraction and classification techniques proposed to overcome these differences where majority of them are inspired from face recognition studies. A general processing chain for traditional 2D FGR methodologies is summarized in Fig. 1.

Initial studies in the domain considered appearance-based features like raw pixels such as in Golomb, Lawrence, and Sejnowski (1990), Gutta, Huang, Jonathon, and Wechsler (2000), Moghaddam and Yang (2002) and Walawalkar, Yeasin, Narasimhamurthy, and Sharma (2002). More recent studies have focused on feature based methods such as in Shan (2012), Santana, Lorenzo-Navarro, and Ramon-Balmaseda (2013), Ramon-Balmaseda, Lorenzo-Navarro, and Castrillon-Santana (2012) and Dago-Casas, Gonzalez-Jimenez, Yu, and Alba-Castro (2011). Histogram of Oriented Gradients (HOG) and Gabor filters for unconstrained gender recognition were studied in Santana et al. (2013). LBP operator (Ojala, Pietikainen, & Maenpaa, 2002) and its variants are also widely used in feature based methods. A recent survey explaining board range of methodologies for vision based gender recognition is presented in Ng, Tay, and Goi (2012). Since gender recognition is also a pattern recognition problem, Adaboost, nearest neighbor classifiers, neural networks, and SVM classifiers are widely used. According to the literature survey, SVM classifier with RBF kernel is the most common classifier used in gender recognition studies because of its high generalization ability.

* Corresponding author. Tel.: +33 362531546.
E-mail addresses: taner.danisman@lifl.fr (T. Danisman), marius.bilasco@lifl.fr (I.M. Bilasco), jean.martinet@lifl.fr (J. Martinet).

http://dx.doi.org/10.1016/j.eswa.2014.11.023
0957-4174/

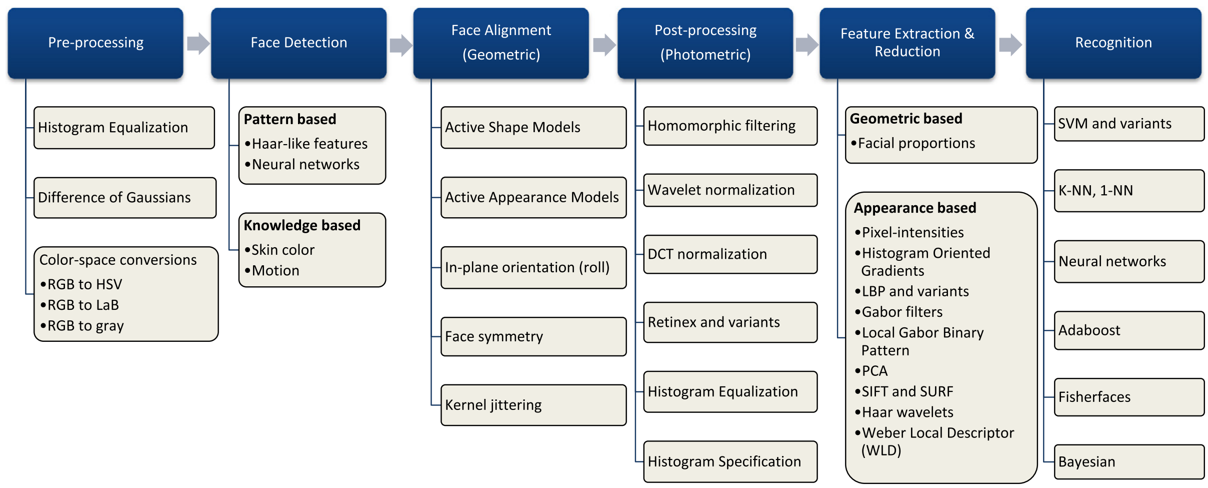


Fig. 1. General processing chain for 2D-FGR methods.

The main shortcoming of 2D FGR is that they focus on inner face area and ignore outer face segments and contextual information that may help improving the generalization ability of the methods. Majority of the literature on gender recognition focuses on extracting information from internal face area and several others focuses on multi-feature extraction by using external cues. Among others, effect of facial hair (e.g., hair, mustache and beard) to FGR is less studied. Current literature discusses only pixel-intensity based and feature-based (e.g., LBP, Gabor filters) evaluation of facial hair for FGR. Face and neck region is studied in Ueki and Kobayashi (2008), hair and upper body clothing is studied in Li, Lian, and Lu (2012), Head-shoulder based gender recognition is studied in Li, Bao, Dong, Wang, and Su (2013). Authors in Tome, Fierrez, Vera-Rodriguez, and Nixon (2014) studied soft biometrics such as gender, hair and arm length. Satta, Galbally, and Beslay (2014) also used contextual information to complement face features. Although these methods obtained better performance compared to the similar FGR methods, they use automatic techniques based on heuristics and localization. Therefore, the actual effect of the contextual information under perfect conditions is still unknown. According to the extensive experiments by Makinen and Raisamo (2008), face location normalization is more important than including facial hair for FGR. They also concluded that inclusion of hair does not guarantee a better classification rate when compared to the images without hair. However, their experiments are based on pixel intensities without considering any segmentation. Therefore, their results depend on the complexity of the background. For example, their experiments on Feret database showed that use of hair information has a positive effect on average FGR accuracy using different classifiers. This is because that the controlled background contributes to the FGR process and virtually provides the hair segmentation. On the other hand, their experiments on WWW images showed that use of hair information has negative effect on average FGR accuracy due to the complex and uncontrolled background. Therefore, there is a need to explore actual effect of contextual information to the FGR on a large scale annotated database. So far there was no such database available. The Part Labels (Kae, Sohn, Lee, & Learned-Miller, 2013) database is the first database providing manual annotations of face, facial hair and background based on superpixels. Using Part Labels database, low-level information extracted from images can be combined with rich contextual knowledge to include human reasoning in the decision process. For example, women's hair is longer than that of men in general as shown in Fig. 2.

This common knowledge may provide additional information for existing classification systems. Although it is difficult to generate a rule covering all female and male subjects, a fuzzy inference system (FIS) can use generated rules based on expert knowledge. FIS are one of the most common applications of fuzzy logic to solve problems in pattern recognition such as in Melin, Mendoza, and Castillo (2010), Polat and Yildirim (2008) and Zadeh (2010). However, considering the visual gender recognition problem, there exist a few fuzzy logic studies. Authors in Leng and Wang (2008) used Fuzzy SVM to increase the generalization ability for gender classification. They used Learning Vector Quantization (LVQ) to generate fuzzy membership functions. Their experiments on different databases confirmed that Fuzzy SVM shows strong robustness to variations than traditional SVM, LDA and NN methods. Authors in Moallem and Mousavi (2013) used shape and texture information to design a fuzzy decision making system. They use Zernike moments to apply texture properties to FIS. They obtained 85.05% accuracy on Feret database including different pose and expressions.

The main advantage of a FIS is its ability to handle linguistic information and to perform nonlinear mappings between the input and output variables. Since FIS is designed from expert knowledge or from raw data, we can generate such rules to solve the gender recognition problem. However, expert knowledge only based FIS may show poor performance (Guillaume, 2001) due to the capacity of the expert to generalize the variability of the subjects. Therefore, it must be supported by additional inputs. Creation of a successful fuzzy system depends on the system design and optimization including quality of the input variables, fuzzy sets and appropriate rules.

In this study, we propose a novel gender recognition framework based on FIS. Our study aims to explore the effect of facial hair to the FGR using a FIS model where the hair is considered as a segmented region rather than pixel-intensities. Therefore, in this

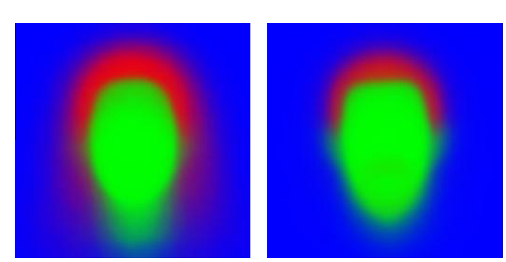


Fig. 2. Average female and male faces obtained from Part Labels database. Red color presents both hair and facial hair. Green color presents inner face and visible neck region. Blue color presents background. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

study, we focused on a set of human factors (e.g., FIS, facial hair) and the classification methodology. We used hair volume and mustache ratio as linguistic variables from our expert knowledge. The output of pixel-based SVM + RBF classifier is then used as a vision-sensor input to the FIS model with other input variables (e.g., hair, mustache). We defined a gender recognition knowledge base having six rules performing nonlinear mapping between the input and output variables. Since we used manually segmented hair information, our study explores actual effect of the use of hair for the FGR. Cross-database tests on LFW showed that FIS obtains better results than the performance baseline of single SVM + RBF approach.

In comparison to previous studies, the main contribution of this study is twofold. First, we showed that hair volume and mustache has positive effect on gender recognition results. We used manually annotated hair information which shows the real effect of the facial hair for the FGR independent from the possible errors in the feature extraction methodology. Second, FIS further improves classical SVM based recognition with proper membership functions and rules presenting human reasoning.

The remainder of this paper is organized as follows. Section 2 presents our methodology based on FIS. Section 3 present databases before discussing deeply experimental setup, evaluation metrics and results obtained on public databases and comparison with the state-of-the-art methods. The final section summarizes and concludes the study with future directions.

2. Methodology

The general framework of the proposed approach is shown in Fig. 3. We used hair and mustache information from Part Labels subset (Kae et al., 2013) of the LFW database (Huang, Ramesh, Berg, & Learned-Miller, 2007). Although we are using manually segmented annotations, methods for automating this process are available in the state of the art (Kae et al., 2013). In addition, we used pixel-intensity based SVM + RBF classifier from our previous work on gender recognition (Danisman, Bilasco, & Djeraba, 2014).

In order to obtain crisp input variables in Fig. 3(a), we performed a geometric and photometric normalization on the input images as described in Danisman et al. (2014). Then low-level and high-level information are extracted using both the annotations and the SVM classifier. The crisp values are then fed into the FIS model shown in Fig. 3(b). The fuzzification step evaluates the crisp input values by considering the corresponding input membership functions to obtain the fuzzy sets. Then, an inference engine evaluates the fuzzy sets and generates a fuzzy set output to be evaluated by the defuzzification step. Finally, we obtain the crisp output from the defuzzification process.

Cross-database tests are performed using different public databases: Feret (Phillips, Wechsler, Huang, & Rauss, 1998), GENKI-4K (http://mplab.ucsd.edu, 2011), Groups (Gallagher & Chen, 2009) and LFW. Individual SVM model training is performed on Feret, GENKI-4K and Groups databases. Optimized models are tested against LFW database by FIS.

2.1. Feature extraction

We extracted three linguistic variables: hair, mustache, and vision-sensor. Hair and mustache information is directly extracted from the Part Labels database. In order to obtain crisp hair ratio value, we normalized hair volume size by facial area. Similarly, we obtained the crisp mustache ratio value by coarse localization of nose and mouth area represented by the square (x = 16, y = 20, w = 8, h = 8) in 40 × 40 images as shown in Fig. 4. All images are normalized as described later in this section. Fig. 5 shows sample images from Part Labels and LFW databases.

Vision-sensor variable is obtained from the response of the SVM classifier where positive and negative responses show the gender information. We linearly extend the response range to [−10, +10] for a better visual display as shown in Fig. 6. In this particular Groups/LFW-P cross-database test, the vertical axis values are used as the vision-sensor value.

2.1.1. Face detection and alignment

We followed the same preprocessing steps described in our earlier study (Danisman et al., 2014). First, we detected the faces using well-known frontal haar-like features model Viola and Jones (2004) available in OpenCV (Bradski, 2000). Then, eye detection is performed to correct in-plane rotation of the face according to the vertical position of left and right pupil. We used the neural network-based eye detector Rowley, Baluja, and Kanade (1998) available in the Stacked Trimmed Active Shape Model (STASM) Milborrow and Nicolls (2008) library to locate the positions of the pupils. After that, a geometric normalization is performed. For the face alignment, we considered normalized *IPD* (Inter-Pupillary Distance) which is the Euclidean distance between the eye centers. Note that initial location of the OpenCV face detection results are updated according to the IPD distance using the following equations where F_x, F_y, F_w and F_h represents new x, y, *width* and *height* of the face. Eye_{Left_x} and Eye_{Left_y} are the x and y positions of the left eye with respect to upper left origin of the image.

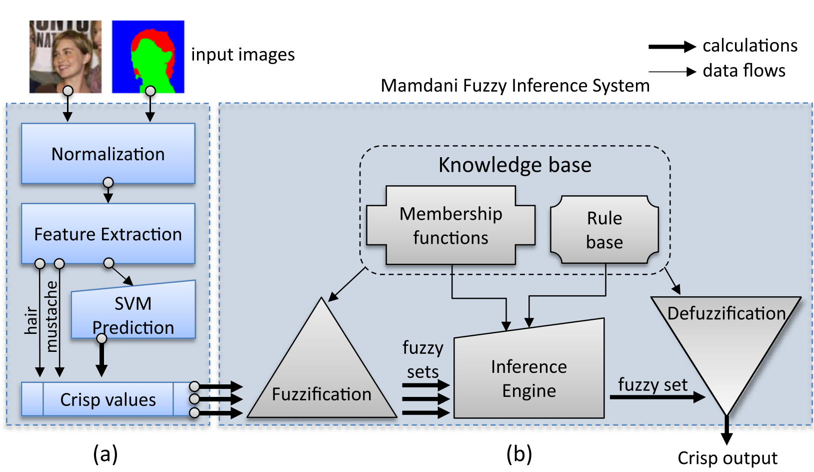


Fig. 3. General framework of the proposed approach. (a) Crisp linguistic variable computation. (b) Mamdani FIS model (Mamdani & Assilian, 1975).

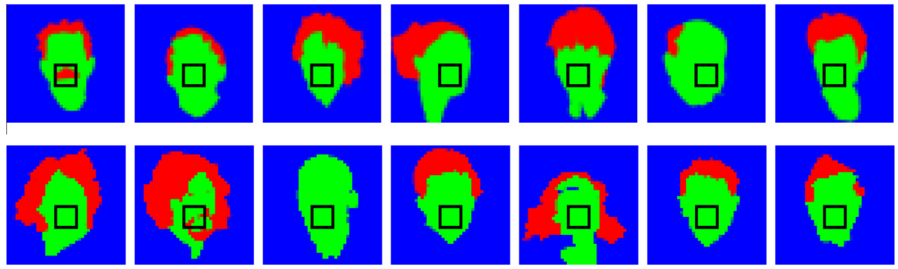


Fig. 4. Coarse mouth localization samples from Part Labels database.

Fig. 5. (a) Segmented female (first row) and male (second row) images from the Part Labels database. (b) Corresponding female and male images from LFW database.

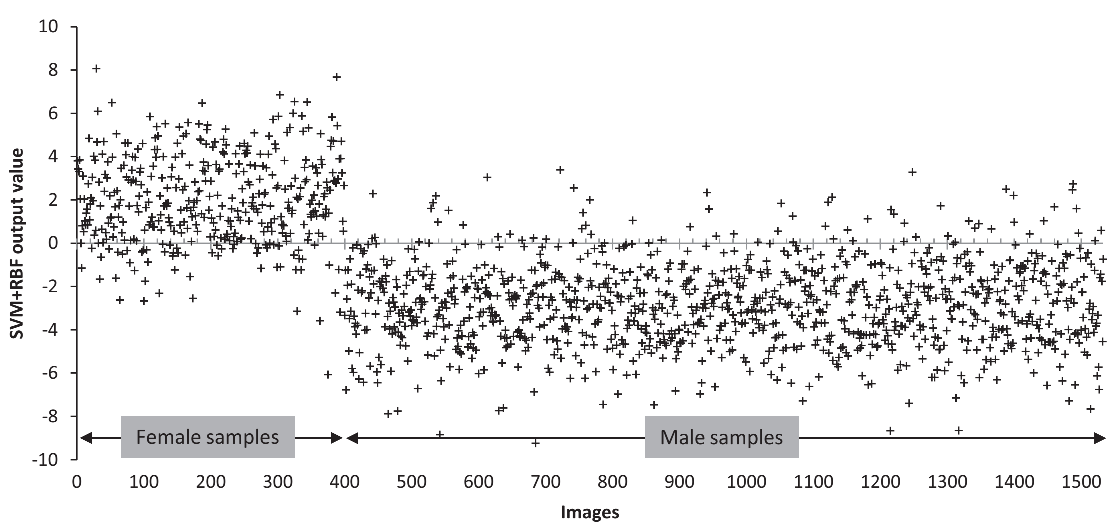


Fig. 6. Cross-database Groups/LFW-P test results (vision-sensor) using SVM + RBF method (Accuracy = 91.25%).

$$F_x = Eye_{Left_x} - IPD/4.0 \quad (1)$$

$$F_y = Eye_{Left_y} - IPD \quad (2)$$

$$F_w = IPD \times 1.5 \quad (3)$$

$$F_h = IPD \times 2.5 \quad (4)$$

Scalar values $4.0, 1.5$ and 2.5 are selected according to experimental observations. Aligned face is then resized to 20×24 image. Finally, a photometric normalization is applied using histogram specification to overcome illumination differences. Fig. 7 shows the initial and cropped face region after the use of Eqs. (1)–(4).

2.1.2. Histogram specification

Histogram specification and histogram equalization are fundamental image enhancement techniques used in image processing. Histogram equalization assigns equal number of pixels to all gray levels. However, this method does not consider common facial appearance. Histogram specification is a generalization of histogram equalization where the image is normalized with respect to a desired probability density function (pdf). Since we know an average human face, we can apply the histogram extracted from the average face to all normalized images. Fig. 8 shows the effect of histogram specification for a given face image using the histogram of the average face.

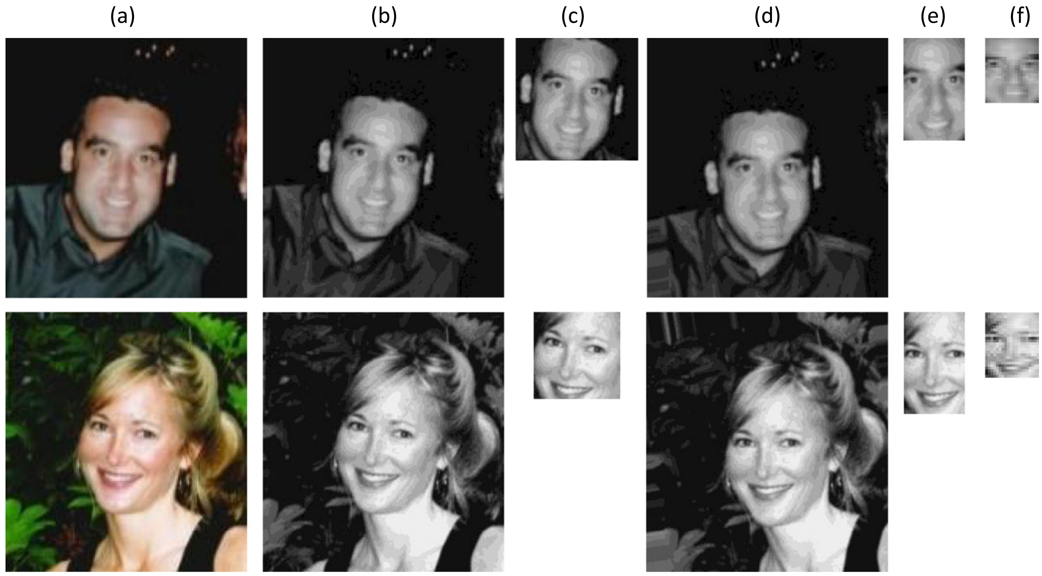


Fig. 7. Geometric normalization and alignment of the face. (a) Original image (b) Gray-level image (c) Face detection (d) Eye detection and in-plane orientation correction according to eye levels (e) Face cropping with respect to the IPD (f) Face scaling by 20×24.

As seen on Fig. 8(c), estimated new histogram is more close to the histogram of the average face. This feature provides better correction of the image histogram in case of different illumination conditions.

2.2. Parameter selection

After the feature extraction step, normalized faces of size 20×24 are used both for training the SVM with RBF kernel and to select the optimal cost (C) and gamma (γ) parameters. We used fivefold cross-validation method on GENKI-4K database for the parameter selection with the *easy* tool present in LibSVM (Chang & Lin, 2011) where a grid search is applied. Since the combination of large γ and large C leads to overfitting, we selected $C = 4$ and $\gamma = 0.03125$ as the optimum values which are similar to the selected parameters obtained in Makinen and Raisamo (2008). This setting provides 90.34% accuracy using fivefold cross-validation on GENKI-4K. Fig. 9 shows the parameter space and corresponding accuracies obtained from fivefold cross-validation.

Hair and mustache boundary values are selected considering visual analysis of the extracted values as presented in Figs. 10 and 11.

2.3. Fuzzy inference system

A FIS is a way of mapping input space variables to one or more output space variables using fuzzy logic. Basic components of a FIS are present in Fig. 3(b). The most common types of the fuzzy systems are Mamdani (Mamdani & Assilian, 1975) and Takagi–Sugeno models (Takagi & Sugeno, 1985) and they are denoted as expert systems. The main difference between the two FIS models is the form of the consequents. In Mamdani model, the output member function can be evaluated independently from the input variables, while in Takagi–Sugeno model, the output member function is a

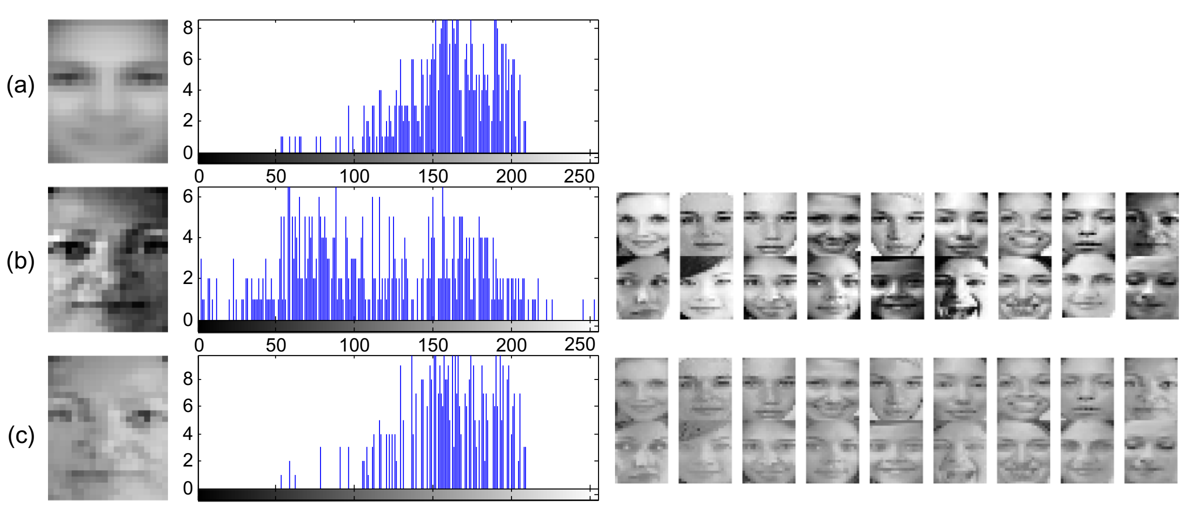


Fig. 8. (a) Average image obtained from web database and its histogram. (b) Example test image and corresponding histogram. (c) Result of histogram specification on (b) using the histogram of (a).

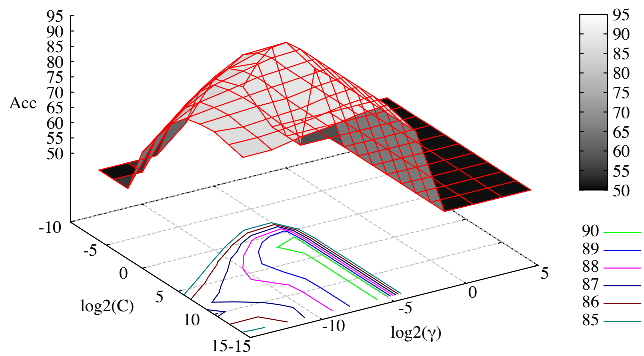


Fig. 9. SVM + RBF parameter space and corresponding fivefold cross-validation accuracy for GENKI-4K. Selected parameters $C = 4$ ($\log_2(C) = 2$) and $\gamma = 0.03125$ ($\log_2(\gamma) = -5$).

function of its inputs. We selected to use Mamdani type FIS for its ability to have independent output member functions, simple structure of min–max operations and wide acceptance for capturing expert knowledge.

A FIS is composed of the following components:

- **fuzzification:** modifies the crisp inputs using the input membership functions (MFs) so that they can be used by the rule base.
- **knowledge base:** consists of a rule base and a database storing the MFs. The rule base consists of a set of rules of type IF-THEN. A rule is also called fuzzy implication having an antecedent and a consequence. The database is a collection of MFs used by both of the fuzzification and defuzzification methods.
- **inference engine:** evaluates relevant rules according to the current input variables.
- **defuzzification:** converts outputs of the inference engine into the outputs of the fuzzy system using a specified defuzzification technique. Center of Gravity (COG), Center of Sums (COS) and Mean of Maximum (MOM) are well-known defuzzification techniques in the literature.

The Fuzzy Logic Toolbox of the Matlab software is used for creating the Mamdani FIS model in Multi Input Single Output (MISO) scheme. We used gaussian combination membership function to define the fuzzy sets. Gaussian combination membership function is a smooth MF that depends on four parameters $\sigma_1, c_1, \sigma_2, c_2$, to define two gaussians as given by:

$$\mu(x;\sigma_1,c_1,\sigma_2,c_2) = \begin{cases} \exp\left[\frac{-(x-c_1)^2}{2\sigma_1^2}\right] & : x < c_1 \\ 1 & : c_1 \leqslant x \leqslant c_2 \\ \exp\left[\frac{-(x-c_2)^2}{2\sigma_2^2}\right] & : c_2 < x \end{cases} \quad (5)$$

where σ_1 and c_1 define the leftmost curve and σ_2 and c_2 define the rightmost curve. Fig. 12 shows MF plots of input and output variables.

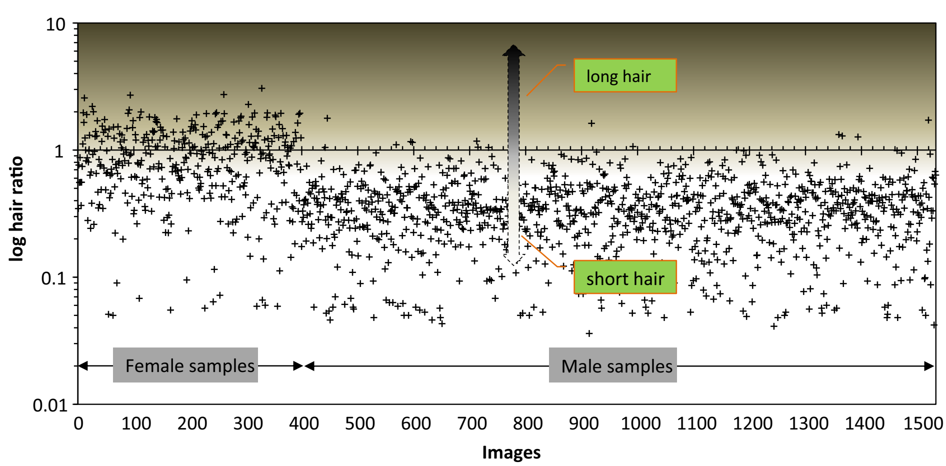


Fig. 10. Logarithmic scale representation of the hair ratio in Part Labels database.

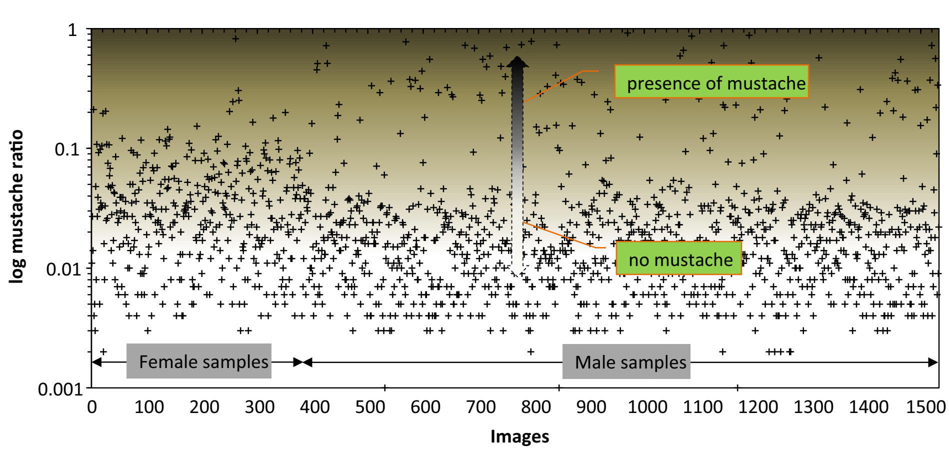


Fig. 11. Logarithmic scale representation of the mustache ratio in Part Labels database.

A fuzzy set A in X is a set of ordered pairs:

$$A = \{(x, \mu_A(x))|x \in X\} \tag{6}$$

where μ_A is the MF, $\mu_A : X \rightarrow M$, M is the membership space where each element of X is mapped to. Therefore, $\mu_A(x)$ presents the degree of membership of x in A, which maps X to the membership space. Considering the Eq. (5), Table 1 presents the details of each gaussian combination MF used in the framework.

We created six rules defining the mapping logic between the input and output variables as shown in Fig. 13. In order to handle fuzzy logic in a rule base system, the "AND" operator is handled as the intersection of the corresponding MF such that, for two fuzzy sets A and B:

$$A \cap B, \mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x)) \tag{7}$$

According to the defined variables and corresponding MFs, Fig. 14 shows the surface view of input and output variables in our Mamdani FIS model. The relation among the hair, mustache and gender is present in Fig. 14(a). Since the mustache value is obtained by measuring the body hair in the localized mouth area, it may include noisy information due to head pose changes and hair occlusions present in this region. This may also happen in frontal upright faces as well. It generates information where the subject has long hair and high mustache ratio at the same time. In such cases, the decision is given by the vision-sensor considering the inner face area. See Fig. 13 (rule 1 and rule 2). Fig. 14(b) shows the relation among hair, vision-sensor and gender. High hair ratio and high vision-sensor response indicates high probability of a female subject. Therefore, in Fig. 14(b) lower part of the surface is activated.

In defuzzification step, we selected COG method to find the point where a vertical line slices the aggregate set into two equal masses as shown in Eq. (8). COG method finds a point representing the center of gravity of the fuzzy set A on the interval ab. Fig. 15 shows examples of female and male inputs and corresponding COG outputs (bold vertical red line on gender column) from the FIS.

$$COG = \frac{\int_a^b \mu_A(x) x dx}{\int_a^b \mu_A(x) dx} \tag{8}$$

Each row in the Fig. 15 shows the evaluation of a single rule from the rule base with respect to the corresponding input values. According to the final evaluation by the COG defuzzification method, final gender decision is given. A female response is given when the COG output value is less than 0.5 and a male response is given when the COG value is greater than 0.5.

3. Experiments

In order to demonstrate the effectiveness of the proposed method, we performed quantitative experiments on variety of databases including Feret, GENKI-4K, Groups and LFW. Experiment databases are selected from publicly available databases and manually annotated into female and male classes, except the Groups database.

3.1. Databases

We select training databases that give the lowest accuracies in cross-database tests. Among others, we used GENKI-4K which is an unconstrained and balanced (in terms of female to male ratio) database for parameter selection as described in Section 2.2. The same parameter set (C and γ) is applied for all train and test experiments.

Table 2 summarizes the characteristics as well as initial and normalized population of the databases used in the experiments. The variety of the features of the selected databases guarantees a basic validation of our method in a wide collection of settings.

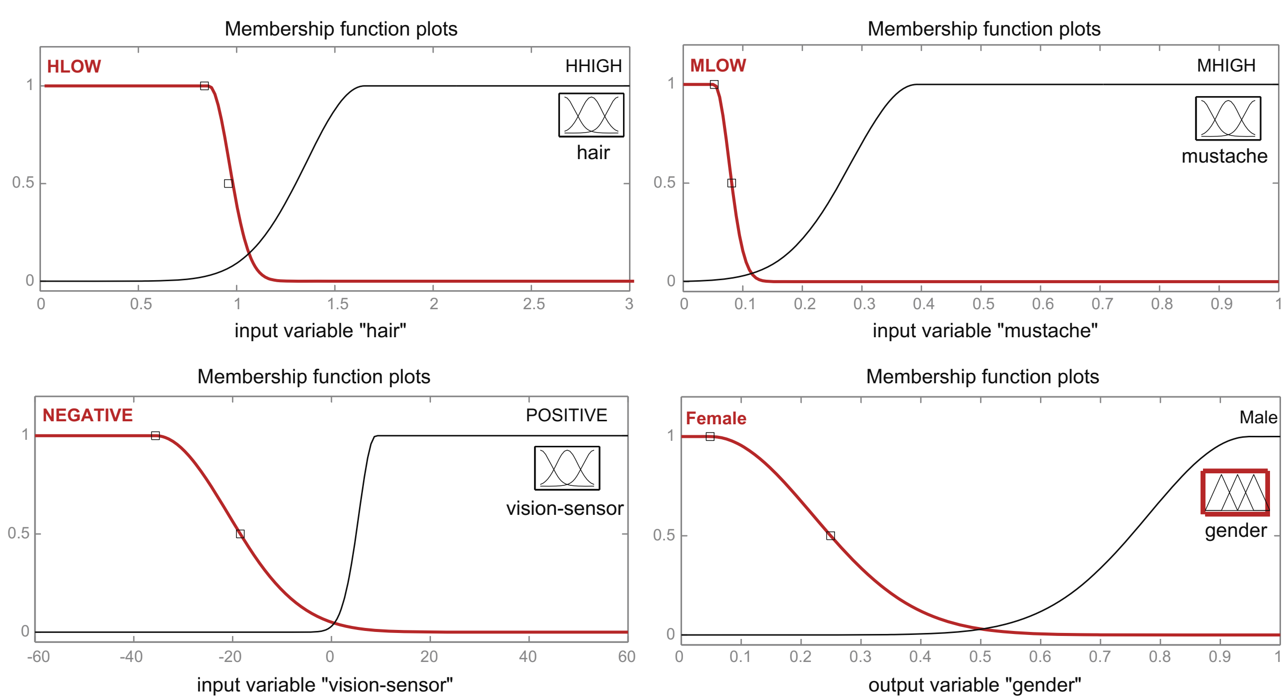


Fig. 12. Input and output membership function plots for hair, mustache, vision-sensor and gender.

Table 1
Details of the MF database. LV = linguistic variable, MF = membership function.

LV	Type	MF	Interval	σ_1	c_1	σ_2	c_2
Hair	Input	hlow	[0,3]	0.251	−0.824	0.103	0.836
Hair	Input	hhigh	[0,3]	0.298	1.652	0.067	3.800
Mustache	Input	mlow	[0,1]	0.217	−0.223	0.024	0.052
Mustache	Input	mhigh	[0,1]	0.110	0.392	0.227	1.157
Vision-sensor	Input	neg.	[−60,60]	13.690	−76.730	14.580	−35.610
Vision-sensor	Input	pos.	[−60,60]	3.397	9.000	8.901	73.020
Gender	Output	Female	[0,1]	0.338	−0.101	0.170	0.0481
Gender	Output	Male	[0,1]	0.168	0.949	0.391	1.114

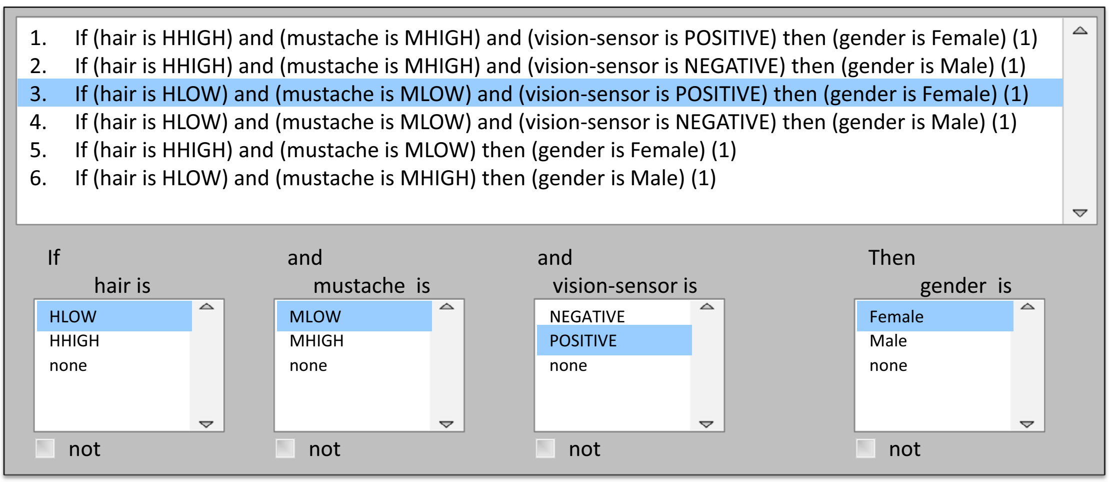


Fig. 13. Details of the rule base.

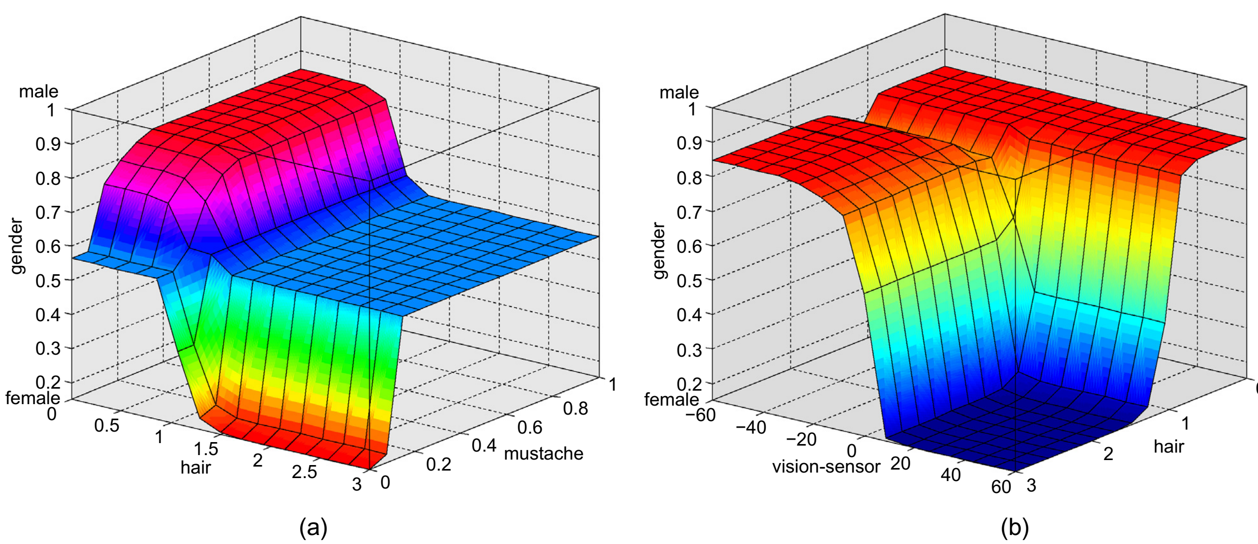


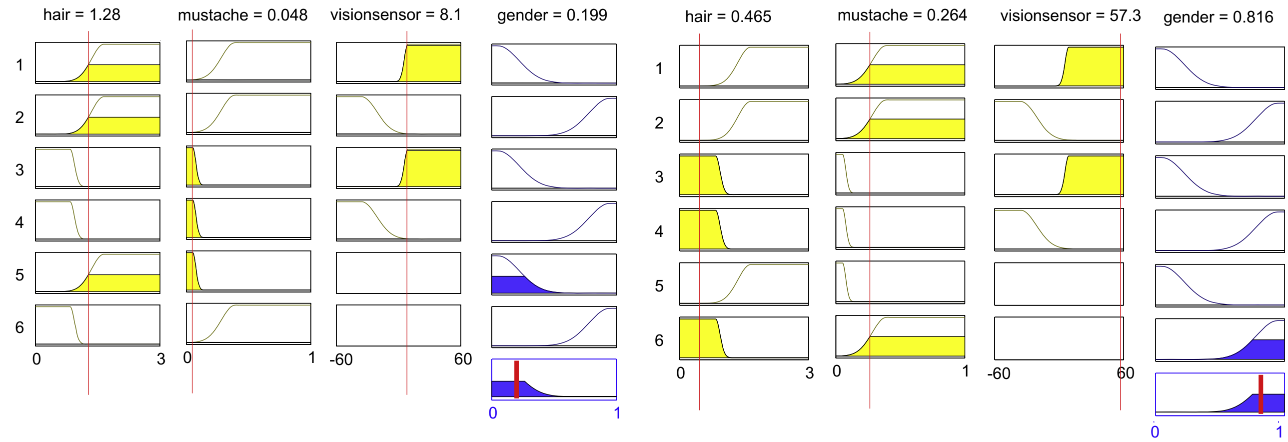
Fig. 14. Surface view of input and output variables. (a) Relation of the hair, mustache and gender. (b) Relation of the vision-sensor, hair and gender.

3.1.1. Genki-4K subset

GENKI-4K database mainly used in facial expression studies containing 4000 face images labeled as either smiling or non-smiling. It involves wide range of subjects, facial appearance, illumination, backgrounds, imaging conditions, and camera model. However, it does not include gender labels. Therefore, we manually labeled the images as female and male classes for our experiments. After the geometric normalization step, we obtained 1539 females and 1506 males.

3.1.2. Image of Groups (Groups)

Groups database (Gallagher & Chen, 2009) includes 5080 images having 28,231 faces labeled with the age and gender categories. It involves wide range of illumination, ethnicity, ages, facial expressions, in-plane and out-of-plane poses. Manually labeled eye positions are provided for all faces. However, we automatically detect the faces and eyes using the methods described in Section 2.1. We obtained a total of 19,835 faces (10,303 female, 9532 male).



(a) Female sample input [1.28, 0.048,8.1] (b) Male sample input [0.465, 0.264, 57.3]

Fig. 15. Example of crisp input and output for different genders.

Table 2
Summary of the databases. AG = different age groups, E = different ethnicities, FAP = total number of faces after the preprocessing and normalization step, FE = facial expressions, I = illumination changes, SEGI = segmented color image, STDI = standard color image, U = unconstrained.

Database	Type	Number of faces	FAP	Female faces	Male faces	Normalized size ($w \times h$)
Feret	STDI, E	2369	2337	908	1429	20×24
GENKI-4K	STDI, FE, E, I, U	4000	3045	1539	1506	20×24
Groups	STDI, FE, AG, E, I, U	28231	19835	10303	9532	20×24
LFW	STDI, FE, AG, E, I, U	13236	11106	8539	2567	20×24
LFW-P	STDI, FE, AG, E, I, U	2927[a]	1533[b]	399	1134	20×24
Part Labels	SEGI	2927[a]	1533[b]	399	1134	40×40

[a] Multiple faces per identity.
[b] Single face per identity.

3.1.3. Labeled Faces in the Wild (LFW) and LFW-P

LFW database (Huang et al., 2007) contains 13,236 labeled images from 5749 individuals mainly actors, politicians and sport players. We automatically select detected faces where eye detection is successful (11,106) and then manually group them into male (8539) and female (2567) categories.

LFW-P is a subset of LFW database that contains the same identities as Part Labels database. Therefore, it contains 2927 face images. After the geometric normalization step, we obtained 399 female and 1134 male faces. LFW-P and Part Labels databases are used together as a test database in cross-database experiments.

3.1.4. Part Labels subset

Part Labels database (Kae et al., 2013) contains labeling of 2927 face images into Hair/Skin/Background labels. The face images are a subset of the Labeled Faces in the Wild (LFW) funneled images. Each image is segmented into superpixels and then these superpixels are manually labeled.

Since this database is originally proposed for image segmentation and labeling problem, to the best of our knowledge, this is the first work which uses the Part Labels database for gender recognition.

3.2. Experimental results

Cross-database and cross-validation based testing are the two common evaluation methodology for the FGR. Majority of the research in FGR uses cross-validation methodology to evaluate the performance. On the other hand, generalization ability of an FGR method is better represented by cross-database evaluations due to the independence of individual identities in training and test sets. Therefore, FGR is a challenging problem under unconstrained settings when cross-database evaluation is applied. As explained in Danisman et al. (2014), average accuracy of the FGR on controlled databases is much higher when a cross-validation scheme is used. The same methods give poor results under cross-database evaluation showing the lack of inconsistency of the generalization ability of the models across different databases.

3.2.1. Baseline performances

We defined three baseline results by using both cross-validation and cross-database tests. Cross-validation baseline results are obtained by evaluating pixel information with the SVM + RBF classifier. Since Part Labels and LFW-P databases are equal in terms of identity, we performed one cross-validation experiment on each of these databases. Using fivefold cross-validation technique, we obtained 88.91% accuracy ($\log_2(C) = 2$ and $\log_2(\gamma) = -9$) on Part

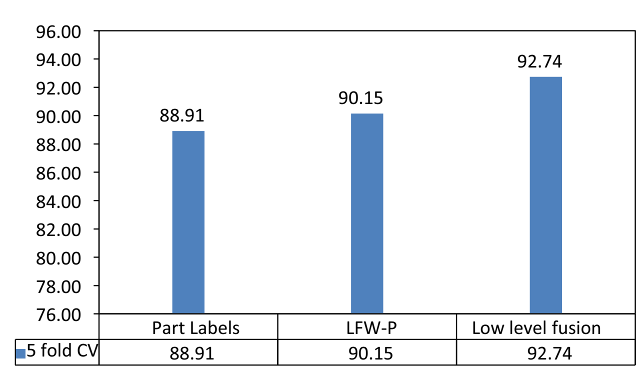


Fig. 16. Baseline performances using fivefold cross-validation method.

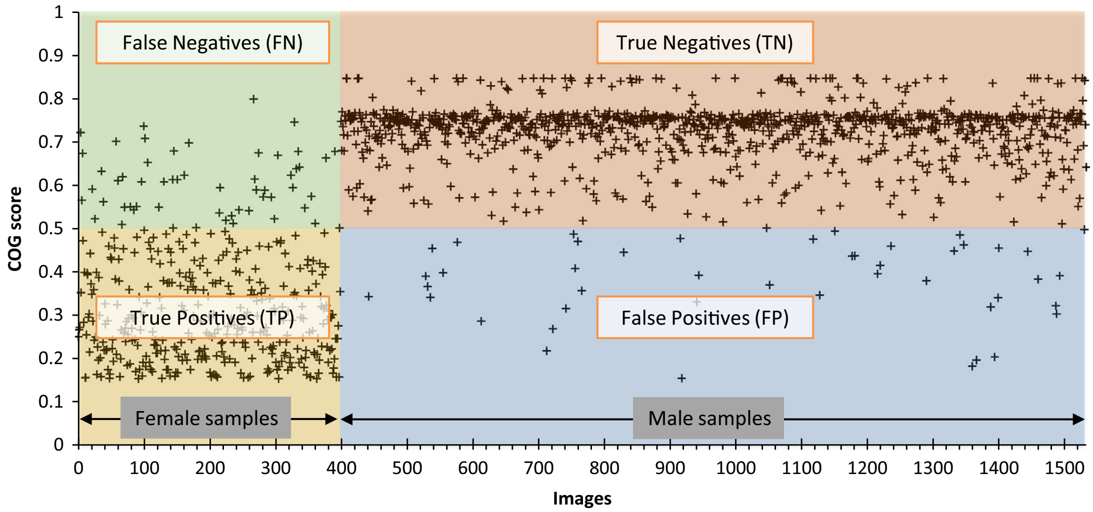


Fig. 17. COG scores obtained from cross-database Groups/LFW-P test using FIS.

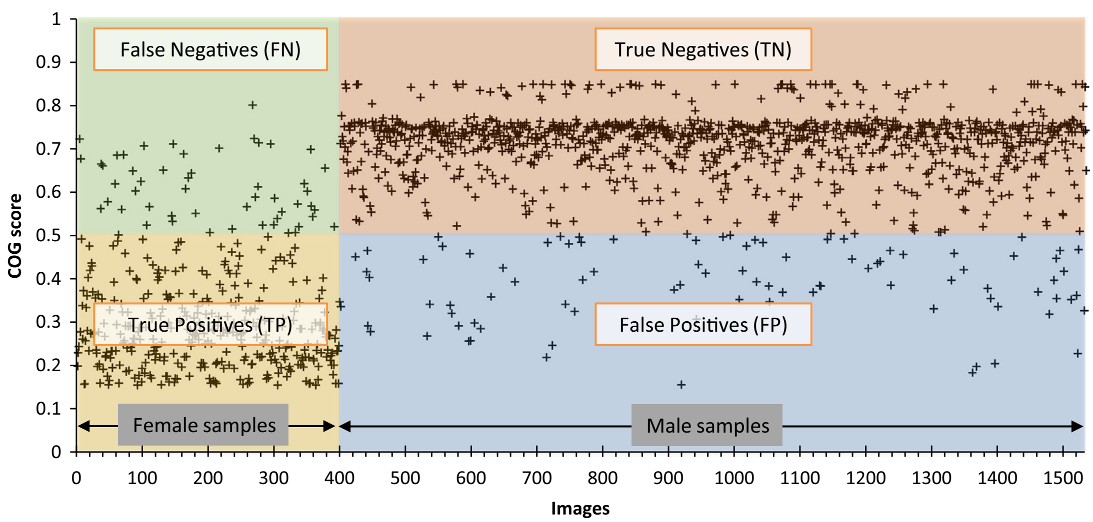


Fig. 18. COG scores obtained from cross-database GENKI-4K/LFW-P test using FIS.

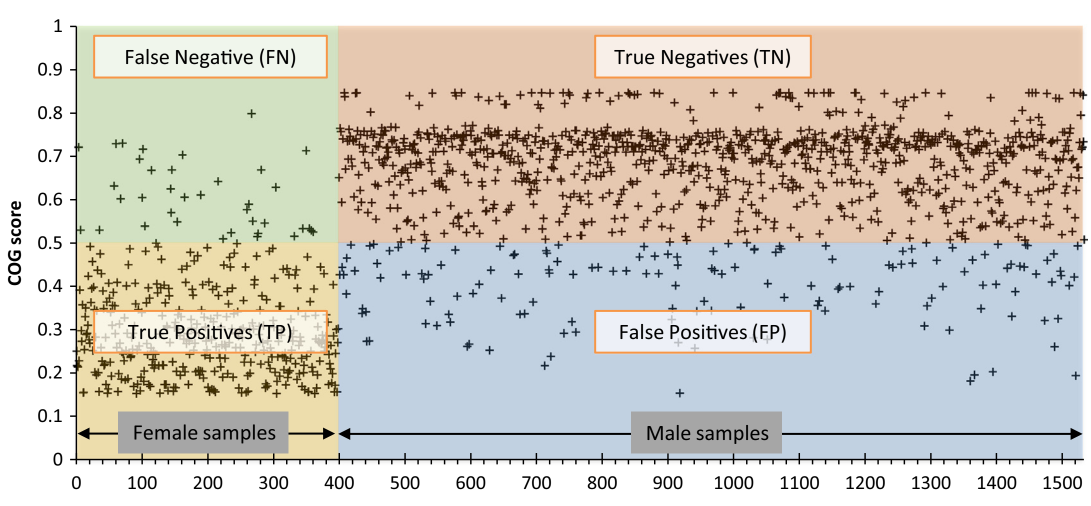


Fig. 19. COG scores obtained from cross-database Feret/LFW-P test using FIS.

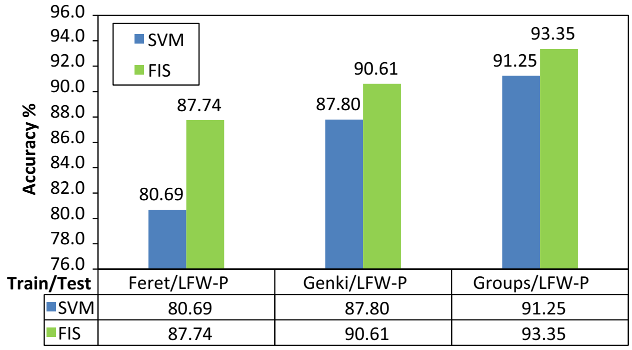


Fig. 20. Cross-database test results on LFW-P.

Labels database and 90.15% accuracy ($\log_2(C) = 2$ and $\log_2(\gamma) = -7$) on LFW-P database using optimized SVM + RBF classifier. Fig. 16 shows baseline performances using fivefold cross-validation method.

We also considered the effect of low-level fusion by concatenating raw pixel information from normalized Part Labels and LFW-P databases. We obtained 92.74% accuracy ($\log_2(C) = 4$ and $\log_2(\gamma) = -11$) which is higher than that of individual cross-validation tests.

In order to obtain the cross-database baseline, we performed three tests: Feret/LFW-P, GENKI-4K/LFW-P and Groups/LFW-P on the LFW-P database. From these experiments, we obtained $80.69\%, 87.80\%$ and 91.25% accuracies respectively. According to the results, the lowest accuracy is obtained from the Feret database. On one hand Feret is a constrained database recorded in controlled environment; therefore it does not perform well on unconstrained LFW-P database. On the other hand, GENKI-4K and Groups databases provide better results than Feret. Compared to the GENKI-4K, the Groups database contains more training samples and support vectors than GENKI-4K which allows more chance for SVM to provide correct results from the soft margin. Therefore, we selected Groups/LFW-P test with 91.25% accuracy as the cross-database baseline performance.

3.2.2. Fuzzy inference system experiments

After obtaining the baseline results, we performed cross-database experiments using the FIS model described in Section 2.3. Figs. 17–19 plot the outputs of the FIS on LFW-P database using Groups, GENKI-4K and Feret databases respectively. Compared to the best cross-database baseline plot in Fig. 6, female and male samples are more far away from each other in FIS outputs.

Since, Groups and GENKI-4K are unconstrained databases; they provide better cross-database results than Feret database on LFW-P. Compared to others, Feret is a restricted database (see Table 2) collected in controlled environment which limits overall accuracy. Our experiment also showed that, databases containing more samples tend to give higher accuracies for gender recognition.

Fig. 20 compares FIS based results to the SVM based results. For each experiment, FIS provides better results than SVM only results. Using FIS, we obtain 93.35% accuracy from cross-database Groups/LFW-P test which is higher than other cross-validation results reported in Fig. 16. Details of all cross-database tests are presented in Table 3.

Experiments showed that the main advantage of FIS over the traditional methods is its ability to perform nonlinear mapping between the input and output. When a FIS is used with a classifier (e.g., SVM), it further eliminates the false positive and false

Table 3
Detailed summary of experiments for LFW and LFW-P databases. H = hair, M = mustache, P = pixel-intensities, CDB = cross-database, CV = 5-fold cross-validation. Bold numbers present the highest accuracies obtained from cross-database tests.

Study	Train/test	Eval.	Model	Test size	Acc. (%)
Dago-Casas et al. (2011)	Groups/LFW[a]	CDB	LBP + PCA + SVM	13,088	89.77
Ramon-Balmaseda et al. (2012)	Morph/LFW[b]	CDB	LBP + SVM + Linear	1149	75.10
Bekios-Calfa et al. (2014)	Groups/LFW[a]	CDB	PCA + LDA + KNN	13,233	79.11
Bekios-Calfa et al. (2014)	Groups[c]/LFW[a]	CDB	PCA + LDA + KNN	13,233	79.53
Danisman et al. (2014)	WebDB/LFW[a]	CDB	P + SVM	11,106	91.87
Danisman et al. (2014)	Groups/LFW[a]	CDB	P + SVM	11,106	91.62
Our CDB baseline	Groups/LFW-P[b]	CDB	P + SVM	1533	91.25
Our CDB baseline	GENKI-4K/LFW-P[b]	CDB	P + SVM	1533	87.80
Our CDB baseline	Feret/LFW-P[b]	CDB	P + SVM	1533	80.69
Our CV baseline	Low-level fusion[b]	CV	H + M+P + SVM	1533	92.74
Our CV baseline	Part Labels[b]	CV	P + SVM	1533	88.91
Our CV baseline	LFW-P[b]	CV	P + SVM	1533	90.15
Our method	Groups/LFW-P[b]	CDB	H + M+P + FIS	1533	**93.35**
Our method	GENKI-4K/LFW-P[b]	CDB	H + M+P + FIS	1533	**90.61**
Our method	Feret/LFW-P[b]	CDB	H + M+P + FIS	1533	**87.74**

[a] Multiple faces per identity.
[b] Single face per identity.
[c] Training step without children faces.

Table 4
Detailed summary of cross-database experiments. TP = true positives, FP = false positives, TN = true negatives, FN = false negatives, PPV = positive predictive value, NPV = negative predictive value. Bold numbers present the highest accuracies obtained from cross-database tests.

Train/test	Model	TP	FP	TN	FN	PPV (%)	NPV (%)	Accuracy (%)
Feret/LFW-P	FIS	361	149	984	39	70.78	96.19	**87.74**
Feret/LFW-P	SVM	361	257	876	39	58.41	95.74	80.69
GENKI-4K/LFW-P	FIS	346	90	1043	54	79.36	95.08	**90.61**
GENKI-4K/LFW-P	SVM	340	127	1006	60	72.80	94.37	87.80
Groups/LFW-P	FIS	343	45	1088	57	88.40	95.02	**93.35**
Groups/LFW-P	SVM	336	70	1063	64	82.75	94.32	91.25

negative results using the knowledge base where the classifier fails. Table 4 shows more detailed results of SVM and FIS. Use of FIS provides reduction in FP and FN while increase number of TP and TN.

4. Conclusion

The current study presents an FGR framework based on FIS for still images using inner and outer facial cues. We present a reliable assessment of the robustness of the presented framework by performing cross-database experiments. We showed that external cues improves the classification performance in both cross-database and cross-validation tests. To deal with the influence of facial hair on FGR we performed tests on LFW and Part Labels database with and without the facial hair feature. LFW database become a standard test database for unconstrained facial gender recognition. Our study is the first to use facial hair information from the Part Labels database for gender recognition purpose. We have confirmed previous results reporting the positive effect of hair to the FGR. However, hair is not a good choice in case of a raw pixel based feature extraction. In addition, unconstrained databases having more training samples provides better results than that of constrained databases since more visual variation covers more area in the solution domain.

Compared to the SVM based approaches that use LBP features (e.g., Dago-Casas et al., 2011; Ramon-Balmaseda et al., 2012) the proposed method provides better results. Although LBP is a powerful texture descriptor, this is an expected result since LBP does not perform well in low resolution images and its performance depends on the quality of the image. Researchers usually require at least 100×100 faces to apply the LBP due to the required region histograms from grid content. Since the images in LFW database are low resolution images, performance of the LBP is limited for the LFW database. On the other hand we used raw pixel based input which performs better than LBP at low resolutions. Our fuzzy model further improves the baseline with the linguistic variables.

The main advantage of the proposed framework is improved generalization ability which is known to be one of the most important features of an FGR system. The use of the high-level knowledge with the FIS improves existing results obtained from traditional methods thus improving the generalization ability. However, the advantage of the proposed framework depends on the distribution of the subjects affected by the fuzzy rules. Another say, amount of the subject with long hair or mustache determines overall success of the proposed framework. When it is tested with subjects without these features (e.g., short hair and no mustache), then output will be theoretically equivalent to the output of SVM based method. Considering the overall system performance, it is obvious that bringing the human reasoning in the decision process by means of nonlinear mapping of input and output variables is advantageous. In addition, the numerical interpretation of the linguistic information requires less computational effort than traditional methods.

Main limitation of the proposed framework is the manual segmentation of the facial hair. Considering the fact that the automatic segmentation systems tend to provide less accurate results than manual systems, future direction is to investigate the overall gain brought by these automatic segmentation methods.

Further research should be conducted in multiple directions. First, the proposed framework can be extended to perform automatic hair segmentation which will provide complete automation of the proposed framework. Part of the experiments in this paper is based on the manually labeled Part Labels database. There is still much research to be done in the area of hair segmentation. Thus, the main challenge will be to identify a robust facial hair segmentation method. With the recent advancements in the well-known super-pixel methods (e.g., Simple Linear Automatic Clustering (SLIC) (Achanta et al., 2012)) in combination with inference algorithms (e.g., Grabcut (Rother, Kolmogorov, & Blake, 2004), SVM) will make it possible to implement an automatic facial hair segmentation system. The overall robustness of proposed framework may be further tested on other large databases when the automatic segmentation is provided.

Second, other complementary information including clothing and accessories can also be considered. Existing studies (e.g., Chen, Gallagher, & Girod, 2012) showed that use of clothing attributes in combination with the facial data further improves the classification performance. In this context, the proposed framework can contribute in a significant manner to further improve existing FGR studies.

Third, different weights of the fuzzy rules and different membership functions can be further analyzed to improve the overall performance of the proposed framework. More generally, the current framework can be extended to explore the effect of multiple vision-sensors obtained from different type of classifiers (e.g., Neural networks, Adaboost).

Acknowledgment

This study is supported by TWIRL (ITEA2 10029 – Twinning Virtual World Online Information with Real-World Data Sources) project.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 2274–2282.

Bekios-Calfa, J., Buenaposada, J. M., & Baumela, L. (2014). Robust gender recognition by exploiting facial attributes dependencies. *Pattern Recognition Letters, 36*, 228–234.

Bradski, G. (2000). The opencv library. *Doctor Dobbs Journal, 25*, 120–126.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:1–27:27.

Chen, H., Gallagher, A., & Girod, B. (2012). Describing clothing by semantic attributes. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision ECCV 2012. Lecture notes in computer science* (Vol. 7574, pp. 609–623). Berlin, Heidelberg: Springer.

Dago-Casas, P., Gonzalez-Jimenez, D., Yu, L. L., & Alba-Castro, J. (2011). Single and cross-database benchmarks for gender classification under unconstrained settings. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 2152–2159).

Danisman, T., Bilasco, I., & Djeraba, C. (2014). Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings. In *22nd IEEE international conference on pattern recognition, ICPR 2014*. Stockholm, Sweden.

Gallagher, A., & Chen, T. (2009). Understanding images of groups of people. *2013 IEEE Conference on Computer Vision and Pattern Recognition, 0*, 256–263.

Golomb, B. A., Lawrence, D. T., & Sejnowski, T. J. (1990). Sexnet: A neural network identifies sex from human faces. In *Proceedings of the 1990 conference on advances in neural information processing systems 3 NIPS-3* (pp. 572–577). San Francisco, CA, USA: Morgan Kaufman.

Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems, 9*, 426–443.

Gutta, S., Huang, J., Jonathon, P., & Wechsler, H. (2000). Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Transactions on Neural Networks, 11*, 948–960.

<http://mplab.ucsd.edu> (2011). The MPLab GENKI Database, GENKI-4K Subset.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07–49 University of Massachusetts, Amherst.

Kae, A., Sohn, K., Lee, H., & Learned-Miller, E. (2013). Augmenting CRFs with Boltzmann machine shape priors for image labeling.

Leng, X., & Wang, Y. (2008). Improving generalization for gender classification. In *15th IEEE international conference on image processing, 2008. ICIP 2008* (pp. 1656–1659).

Li, M., Bao, S., Dong, W., Wang, Y., & Su, Z. (2013). Head-shoulder based gender recognition. In *ICIP* (pp. 2753–2756). IEEE.

Li, B., Lian, X.-C., & Lu, B.-L. (2012). Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing, 76*, 18–27. Seventh international symposium on neural networks (ISNN 2010) advances in web intelligence.

Makinen, E., & Raisamo, R. (2008). An experimental comparison of gender classification methods. *Pattern Recognition Letters, 29*, 1544–1556.

Mamdani, E., & Assilian, S. (1975). A fuzzy logic controller for a dynamic plant. *International Journal of Man-Machine Studies, 7*, 1–13. cited by (since 1996) 3.

Melin, P., Mendoza, O., & Castillo, O. (2010). An improved method for edge detection based on interval type-2 fuzzy logic. *Expert Systems with Applications, 37*, 8527–8535. cited by (since 1996) 55.

Milborrow, S., & Nicolls, F. (2008). Locating facial features with an extended active shape model. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer vision ECCV 2008*. *Lecture notes in computer science* (Vol. 5305, pp. 504–513). Berlin/ Heidelberg: Springer.

Moallem, P., & Mousavi, B. S. (2013). Gender classification by fuzzy inference system. *International Journal of Advanced Robotic Systems, 10*(89).

Moghaddam, B., & Yang, M.-H. (2002). Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 707–711.

Ng, C., Tay, Y., & Goi, B.-M. (2012). Recognizing human gender in computer vision: A survey. In P. Anthony, M. Ishizuka, & D. Lukose (Eds.), *PRICAI 2012: Trends in artificial intelligence*. *Lecture notes in computer science* (Vol. 7458, pp. 335–346). Berlin, Heidelberg: Springer.

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

Phillips, P., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The {FERET} database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16*, 295–306.

Polat, O., & Yildirim, T. (2008). Genetic optimization of GRNN for pattern recognition without feature extraction. *Expert Systems with Applications, 34*, 2444–2448.

Ramon-Balmaseda, E., Lorenzo-Navarro, J., & Castrillon-Santana, M. (2012). Gender classification in large databases. In L. Alvarez, M. Mejail, L. Gomez, & J. Jacobo (Eds.), *Progress in pattern recognition, image analysis, computer vision, and applications*. *Lecture notes in computer science* (Vol. 7441, pp. 74–81). Berlin, Heidelberg: Springer.

Rother, C., Kolmogorov, V., & Blake, A. (2004). GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics, 23*, 309–314.

Rowley, H., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 23–38.

Santana, M. C., Lorenzo-Navarro, J., & Ramon-Balmaseda, E. (2013). Improving gender classification accuracy in the wild. In J. Ruiz-Shulcloper & G. S. di Baja (Eds.), *CIARP (2)*. *Lecture notes in computer science* (Vol. 8259, pp. 270–277). Springer.

Satta, R., Galbally, J., & Beslay, L. (2014). Children gender recognition under unconstrained conditions based on contextual information. In *22nd IEEE International Conference on Pattern Recognition, ICPR 2014*. Stockholm, Sweden.

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

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 116–132. SMC-15.

Tome, P., Fierrez, J., Vera-Rodriguez, R., & Nixon, M. (2014). Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security, 9*, 464–475.

Ueki, K., & Kobayashi, T. (2008). Gender classification based on integration of multiple classifiers using various features of facial and neck images. *Information and Media Technologies, 3*, 479–485.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57*, 137–154. http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb.

Walawalkar, L., Yeasin, M., Narasimhamurthy, A., & Sharma, R. (2002). Support vector learning for gender classification using audio and visual cues: A comparison. In S.-W. Lee & A. Verri (Eds.), *Pattern recognition with support vector machines*. *Lecture notes in computer science* (Vol. 2388, pp. 144–159). Berlin, Heidelberg: Springer.

Zadeh, A. E. (2010). Automatic recognition of radio signals using a hybrid intelligent technique. *Expert Systems with Applications, 37*, 5803–5812.