Face analysis through semantic face segmentation

Article · May 2019

DOI: 10.1016/j.image.2019.01.005

CITATION
1

READS
857

5 authors, including:

Sergio Benini
Università degli Studi di Brescia
62 PUBLICATIONS **503** CITATIONS
SEE PROFILE

Khalil Khan
University of Azad Jammu and Kashmir
27 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Riccardo Leonardi
Università degli Studi di Brescia
263 PUBLICATIONS **2,497** CITATIONS
SEE PROFILE

Signal Processing: Image Communication 74 (2019) 21–31

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Face analysis through semantic face segmentation

Sergio Benini [a], Khalil Khan [a,b], Riccardo Leonardi [a], Massimo Mauro [a], Pierangelo Migliorati [a,*]

[a] Department of Information Engineering, University of Brescia, Brescia, Italy
[b] Department of Electrical Engineering, The University of Azad Jammu and Kashmir, Pakistan

ARTICLE INFO

Keywords:
Face segmentation
Head pose estimation
Gender recognition
Face expression classification

ABSTRACT

Automatic face analysis, including head pose estimation, gender recognition, and expression classification, strongly benefits from an accurate segmentation of the human face. In this paper we present a multi-feature framework which first segments a face image into six parts, and then performs classification tasks on head pose, gender, and expression. Segmentation is achieved by training a discriminative model on a manually labeled face database, namely FASSEG, which we extend from previous versions, and which we publicly share. Three kinds of features accounting for location, shape, and color are extracted from uniformly sampled square image patches. Facial images are then pixel-wise segmented into six semantic classes – hair, skin, nose, eyes, mouth, and background, – using a Random Forest classifier (RF). Then a linear Support Vector Machine (SVM) is trained for each face analysis task i.e., head pose estimation, gender recognition, and expression classification by using the probability maps obtained during the segmentation step. Performance of the proposed framework is evaluated on four face databases, namely Pointing'04, FEI, FERET, and MPI, with results which outperform the current state-of-the-art.

1. Introduction

Human Face Segmentation (HFS) is an active research area in computer vision and multi-media signal processing. It can be included in the more general problem of semantic segmentation, on which extensive research work is carried out in the context of the PASCAL VOC challenge [1].

Face segmentation often plays a crucial role in many directly face-related applications such as face detection and tracking [2], face recognition [3], expression analysis [4], head pose estimation [2,5], face swapping [6,7], etc. It is also useful for other applications which are not directly related to face analysis, such as the estimation of shot scale [8] or video affective characterization [9,10].

Psychology literature confirms that isolating single facial features, such as 'hair', 'nose', 'eyes', etc. is essential for the human visual system to recognize face identity [11,12]. Therefore the performance of all mentioned applications can be likely boosted if a properly segmented face is provided in input. However, variations in illumination and visual angle, complex background, different facial expressions and orientations make human face segmentation a challenging task.

In our previous work [13–15] we tackle the challenges posed by HFS and provide the first version of FASSEG dataset [16]. Differently from previous approaches which considered only three or four classes, in [13] we extend the labeled set to six semantic classes (i.e., 'hair', 'skin', 'nose', 'eyes', 'mouth', and 'background'), working on high resolution frontal images (FASSEG-frontal01 subset).

Following Haung et al. who first argue in [17] that head pose, gender and face expression can be more easily predicted if a prior accurately segmented face is provided in input, in the present paper we tackle the problems of pose, gender and expression classification starting from a face segmented image. Extending our work in [13], we here provide a new labeled subset of the FASSEG database (FASSEG-frontal02 subset), which improves the performance of the frontal face segmenter. Similarly to [13] we exploit color, shape, and location features to build a discriminative model using Random Forest (RF). As shown in Fig. 1 which illustrates the system workflow, the RF classifier returns a probability value and a class label for each pixel of the testing image among 'hair', 'skin', 'nose', 'eyes', 'mouth', or 'background'. Then using different combinations of pixel-wise probability maps obtained during segmentation, we train a linear Support Vector Machine (SVM) for each of the following tasks: head pose estimation (HPE), gender recognition (GR), and expression classification (EC). As a last contribution, we make available another subset of the FASSEG dataset, namely FASSEG-multipose01 subset, useful for the task of head pose estimation.

Novelties in a nutshell are therefore: an extended version of the FASSEG [16] dataset with two novel subsets for improving face segmentation and multiple head pose estimation, and an integrated framework for performing HPE, GR, and EC. Obtained results outperform

* Corresponding author.
E-mail addresses: sergio.benini@unibs.it (S. Benini), khalil.khan@ajku.edu.pk (K. Khan), riccardo.leonardi@unibs.it (R. Leonardi), massimo.mauro@unibs.it (M. Mauro), pierangelo.migliorati@unibs.it (P. Migliorati).

https://doi.org/10.1016/j.image.2019.01.005
Received 12 April 2018; Received in revised form 16 November 2018; Accepted 12 January 2019
Available online 18 January 2019

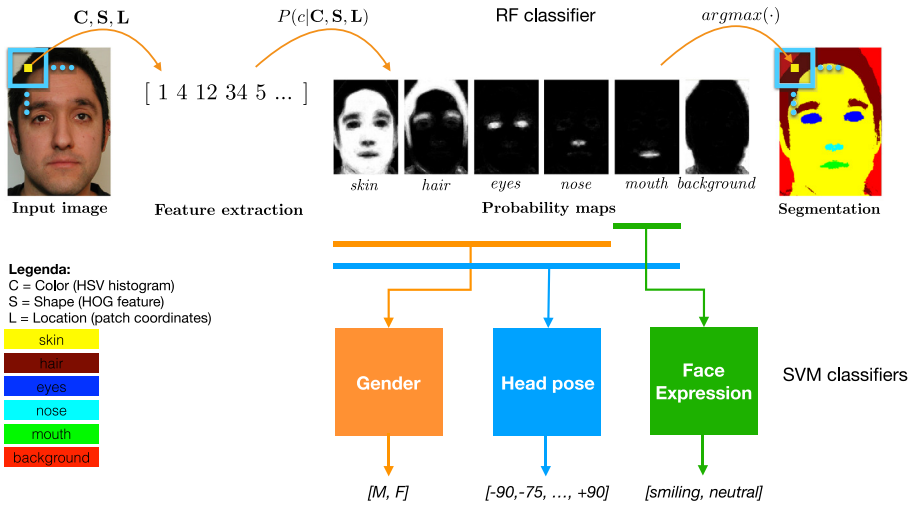

Fig. 1. Framework and workflow for head pose estimation, gender recognition, and expression classification based on previous segmentation of facial parts.

current state-of-the-art methods for HPE, and are comparable for GR and EC, when tested on standard databases such as the challenging Pointing'04 [18], FEI [19], FERET [20], and MPI [21].

The structure of the paper is as follow. In Section 2 we briefly review previous work on face segmentation, and give a wider overview on how HPE, GR and EC problems have been previously addressed. Section 3 presents the used databases. In Section 4 we describe the proposed HFS algorithm. Section 5 illustrates how we exploit segmentation to solve HPE, GR, and EC problems. In Section 6 we discuss the results. Section 7 gathers conclusions and outlines the future work.

2. Related works

In the last years an important research effort focused on algorithms to assign a semantic label to different facial parts (see for example [22–25]). Most of these methods for human face segmentation consider either three (i.e., 'skin', 'hair', and 'background') [17,26–28] or four classes [29–31], adding the label 'clothes' to the aforementioned ones. As an enrichment with respect to previous approaches, our work in [13] first proposes a six-class labeling, thus distinguishing among 'skin', 'nose', 'hair', 'mouth', 'eyes', and 'background'.

Other face analysis related tasks, such as head pose estimation, gender recognition, and expression classification, are historically handled separately from each other. A good survey on HPE can be found in [32]. Focusing our analysis on methods that are evaluated on the Pointing'04 dataset (which we also adopt for comparison), Stiefelhagen [33] extracts first order image derivatives to train a neural network which discriminates among poses. More recently, Ma et al. [34,35] estimate yaw using the symmetry principle by an effective combination of Gabor filters and covariance descriptors. Multi-scale gaussian derivatives are used in combination with SVM by Jain et al. [36] to implement a method which is suitable also for hand-held devices. Hara et al. [37] propose a regression method for head pose based on forests, where the best splitting is found from an already defined set of rules.

The state of the art algorithm for HPE, at least until the advent of convolutional neural networks (CNN), is the Multivariate Label Distribution (MLD) proposed in [38]. By using histogram of gradient (HOG) features [39], MLD introduces the idea of soft labels: rather than explicit hard labels, every image is associated with a label distribution. Then the MLD approach tries to capture the correlation between neighboring poses in the label space.

Huang et al. [17] are among the first who suggest that HPE, GR, and EC are more easily predicted from a well segmented facial image. Using a rather small database of 100 images they segment facial images into three semantic classes: 'skin', 'hair' and 'background'. Then, using the segmented image, they estimate three simple head poses: 'frontal', 'right', and 'left' profile.

With the recent advent of deep learning methods, which have rapidly become omnipresent within the image analysis community, some works started to propose the use of CNN for face analysis and head pose estimation [2,5]. Segmentation-free approaches for HPE employing CNN can be found in [40], which classifies head pose with two degrees of freedom (pitch and yaw) using a low-resolution single image, and in [41], where the authors compare the performance of different network architectures using adaptive gradient methods and drop out on three different datasets, including Pointing'04 (referred to as "Prima" dataset, as the name of the research group [18]) thus enabling comparison with previous work.

A survey on gender recognition methods is provided in [42]. Many among the first methods rely on one or multiple combinations of well-known features such as SURF, LBP, SIFT, and HOG. In [21] authors instead exploit color histogram descriptors, texture, and geometrical information for clustering pixels using a k-means algorithm; then these information are input to a single layer perceptron that classifies each face as male or female. Other approaches to gender recognition exploit several facial binary attributes (e.g., long hair, white, etc.) such as [43]. More recent appearance-based methods are found in [44] and [45]. Also CNN are used as in [46,47] for the purpose of learning gender representations based on attributes.

Active Shape Modeling (ASM) [48] is often used to locate facial landmarks for both gender recognition and expression classification. A feature vector is created from landmarks and a machine learning tool is used for performing training and prediction. However ASM may fail to locate correct landmarks if lighting conditions change, or if facial expressions are complex, or if the pose of the image changes from frontal to an extreme profile. Other notable appearance-based methods for EC appear in [49,50]. In the specific [49] uses Gabor filters with a SVM classifier, while [50] proposes local binary pattern (LBP) as an alternative for feature extraction. Finally, hybrid methods for expression classification mixing appearance and geometric properties are proposed in [51] and [52].

3. Face databases

In this section we describe face data employed in this work.

3.1. FASSEG database

First introduced and made available in [13], the FAce Semantic SEGmentation (FASSEG) dataset initially contained only one set of

manually annotated segmentation masks of face images (FASSEG-frontal01 subset). Now FASSEG is composed by a total of three subsets: two for frontal face segmentation (FASSEG-frontal01 and FASSEG-frontal02), and one (FASSEG-multipose01) with labeled faces in multiple poses. In this work we use FASSEG-frontal02 for training the frontal face segmenter, and FASSEG-multipose01 for working on multiple head poses.

3.1.1. FASSEG-frontal02

The FASSEG-frontal02 subset is made of 70 manually labeled images taken from the MIT-CBCL [53] and the FEI [19] face databases. It contains refined versions of the segmentation masks provided in the former FASSEG-frontal01. Images are organized in two folders – train and test – matching the division we adopt in the paper. Faces present a moderate degree of variability, as it includes people of different ethnicity, gender, and age. Moreover faces are not perfectly aligned in position and scale. This makes the algorithm suitable for performing face segmentation on the bounding-boxes derived from a previous face detection.

3.1.2. FASSEG-multipose01

FASSEG-multipose01 contains more than 200 labeled faces in multiple poses. Original faces are taken from the Pointing'04 database which contains pictures from 15 subjects asked to gaze at 93 markers in a measurement room. The orientation of the head is determined by pan and tilt varying from −90° to +90° in vertical and horizontal orientations. Our data-set is limited only to horizontal orientation poses, and labeled using a step size of 15°.

3.2. Other datasets

3.2.1. Pointing'04

Available datasets for head pose estimation are limited and a precise ground truth of head pose is difficult to obtain. The widely used Pointing'04 head pose database [18] – which we also adopt for comparison – is collected by asking humans sitting at the same position of the room to look at the markers. This approach leads to approximate results: both the starting 3D head location and the direction of the head towards the markers are not perfectly aligned. As a result, the dataset contains a finite number of coarse poses, and considers only yaw and pitch angles. The database contains images of 15 people, with pan and tilt varying from −90° to +90° in both horizontal and vertical orientation. The phase shift for horizontal orientation is 15°, while for vertical orientation is 30°.

3.2.2. FEI and FERET

To evaluate the performance of gender recognition and expression classification we use other two databases, namely FEI [19] and FERET [20]. FEI database is composed of 200 subjects (100 males and 100 females), where each person appears in two frontal images (one with smiling expression, and one neutral). We used both image folders (namely FEI A containing all neutral expression images, and FEI B for all smiling faces) for a total number of 400 images, for both gender recognition and expression classification. Similarly 400 images coming from the FERET database are employed for both gender and expression classification tests.

3.2.3. MPI

The MPI dataset, first introduced in [21], contains 200 images of frontal faces (100 males and 100 females) collected in controlled lighting conditions. None of the MPI face images shows facial hair, glasses or any other accessories; beside this all faces have neutral expression, so the database is relatively easier when compared to FASSEG database. In our work we use it to evaluate and compare the performance of the proposed gender recognition method.

4. Face segmentation algorithm

The general workflow of the proposed segmentation algorithm is shown in the top part of Fig. 1. Section 4.1 illustrates how features are extracted from image patches, while Section 4.2 details the classification of facial parts.

4.1. Feature extraction

We use square patches as processing primitives, which are extracted from database images keeping a fixed step size of one pixel. Every patch is then classified and the class label transferred to the central pixel of the patch. In both training and testing we rescale the original images in order to have a constant height of $H = 512$ pixels, while width W is varied accordingly to keep the original image ratio. As a result, the type of content for a given patch dimension is comparable for different face images.

For our classification purposes, from each patch we extract color and local shape features, combined with spatial information.

As color features we adopt HSV color histograms, thus concatenating hue, saturation, and value histograms to form a single feature vector. To account for shape information we extract the widely used HOG feature [39]. As spatial information we use the relative location of the patch. Specifically we extract the coordinates (x, y) of the patch central pixel and represent them in relative terms as $f_{LOC} = [x/W, y/H] \in \mathbb{R}^2$, where W is the width, and H is the height of the image.

All three features (HSV, HOG, and spatial location) are then concatenated to form, for each patch $I(x, y)$, a single feature vector f^I. In Section 6 we explore different parameterizations for the patch dimension of each feature ($D_{HSV} = D_{HOG} = 16 \times 16, 32{\times}32$, and 64×64), for the number of color histogram bins ($N_{bins} = 16$, 32 and 64), and then compare generated feature vectors to find the parameters which ensure the best segmentation accuracy.

4.2. Classification of facial parts

A Random Forest (RF) classifier is then trained using the implementation in [54]. As the trained RF predicts for each pixel a probability value of belonging to each class c, the predicted class label $\hat{c}$ of each pixel is assigned based on maximum probability:

$$\hat{c} = \arg\max_{c \in \mathbb{C}} p(c|\mathbf{L}, \mathbf{C}, \mathbf{S}) \quad (1)$$

where $\mathbb{C} = \{skin, background, eye, nose, mouth, hair\}$ and $\mathbf{L}$, $\mathbf{C}$, and $\mathbf{S}$ are random variables for features f_{LOC} (relative location), f_{HSV} (HSV color), and f_{HOG} (shape local feature), respectively. Example of segmentation results are given in Fig. 2.

5. Classification of head pose, gender, and expression

While for the segmentation task the method outputs the most likely class for each pixel, for the other classification tasks (HPE, GR, and EC) we exploit the entire probability maps generated during segmentation for each semantic class $c \in \mathbb{C}$, as shown at the bottom of Fig. 1. Probability maps associated to face parts ($\boldsymbol{P}_{skin}$, $\boldsymbol{P}_{hair}$, $\boldsymbol{P}_{eyes}$, $\boldsymbol{P}_{nose}$, $\boldsymbol{P}_{mouth}$, and $\boldsymbol{P}_{back}$) are generated by converting the probability of each pixel so as to obtain a grey scale image, where higher pixel intensities represent higher values of probability for that pixel of belonging to that class.

Fig. 3 shows two face images taken from Pointing'04 database, and their respective probability maps for five considered classes of facial parts.

For each task (HPE, GR, EC) we train a separate linear SVM classifier with a feature vector which is given by the concatenation of different probability maps. After a thorough investigation to select the best features for each task, HPE employs five probability maps ($\boldsymbol{P}_{skin}$, $\boldsymbol{P}_{hair}$, $\boldsymbol{P}_{eyes}$, $\boldsymbol{P}_{nose}$, and $\boldsymbol{P}_{mouth}$), GR exploits all previous maps but $\boldsymbol{P}_{mouth}$, while EC adopts only $\boldsymbol{P}_{mouth}$, as shown in Fig. 1. Specific binary SVM classifiers are trained for GR and EC, while a multi-class SVM is used for HPE.

Fig. 2. Example of segmentation results on FASSEG dataset. First row: original images. Second row: labeled ground truth (in FASSEG-frontal02). Third row: segmented images.

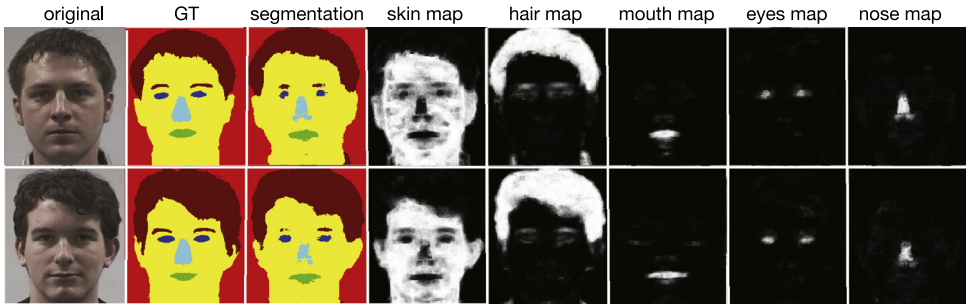

Fig. 3. Frontal images from Pointing'04 database. In columns (from left to right): (1) original images; (2) ground truth masks; (3) segmented images by our algorithm; (4) skin probability map; (5) hair probability map; (6) mouth probability map; (7) eyes probability map; (8) nose probability map.

5.1. Pose estimation

For the task of head pose estimation the RF classifier responsible for segmentation is first trained on 14 subjects from FASSEG-multipose01 whose face images are manually labeled for each of the 13 poses (from −90° to +90° with step size 15°).

The idea behind the algorithm is as follows: given a test image, we run the RF segmentation models for all face classes and for all different poses, obtaining, for every pixel, a set of class predictions with associated probabilities. The probability maps for a single subject are shown in Fig. 4 (5 classes, 13 poses).

As it is evident from Fig. 4, probability maps highly differ from pose to pose. Considering for example the first row (i.e., the skin probability map) in frontal images the forehead is more exposed to the camera, and consequently its size is larger. As a result, the brighter area of the probability map is concentrated in the center of the image. Similarly, probability maps are brighter on the right and left side of the image in the left profile poses (0° to −90°), and right profile ones (0° to +90°), respectively. On extreme poses high intensity values are bounded to a comparatively smaller area (e.g., as in Fig. 4, row 1).

Since skin and hair are in a way complementary, opposite observations can be made for hair: in frontal images hair is less, while as pose changes from frontal to profile, more hair is exposed to the camera. As a result, probability maps have larger brighter areas in extreme poses (as in Fig. 4, row 2).

Probability maps for mouth, nose, and eyes are also shown in Fig. 4 on rows 3, 4, and 5, respectively. A large variability can be noted for these probability maps as pose changes occur. For example, high intensity pixels are almost missing for eyes and mouth on extreme profile poses.

After investigating which facial parts are the most discriminative for understanding head pose (see Section 6 for details), we use the concatenation of all five probability maps (i.e., 'skin', 'hair', 'nose', 'eyes', and 'mouth') to form a single feature vector and use it to train a multi-class linear SVM classifier for head pose estimation. Performance evaluation is carried out using test images from Pointing'04 database. To increase the amount of testing data and validate the model more precisely, we use 10-fold cross validation.

5.2. Gender recognition

For the task of gender recognition the RF classifier responsible for segmentation is first trained with a total of 60 manually labeled images, 30 from FEI [19] (15 males, 15 females) and 30 from FERET [20] (15 males, 15 females). After a pool of experiments (see Section 6) facial parts that play a key role in male and female face differentiation are: 'skin', 'nose', 'eyes', and 'hair'.

In the following we give a short overview taken by gender recognition literature on the reasons why these parts are helpful. First male foreheads are usually larger than in females, especially because the hairline for men is higher than in women (and for bold men the hairline is completely missing). Then male necks are comparatively thicker than female ones. As a result, skin maps for men occupy larger areas and are brighter than in women, as shown for two subjects of the FERET database in Fig. 5, first column.

From our segmentation results we notice that male eyes are often more accurately segmented if compared to female ones. This probably is due to the fact that female eyelashes are usually longer and curly outwards: therefore they are mostly miss-classified with hair, and as a result, location for eyes are predicted with less accuracy. As compared to female, male eyelashes are hardly visible, so better segmentation results are produced (see Fig. 5, column 3).

It is also visible that female noses are on average smaller, having smaller bridge and ridge [55], while an average male nose is larger and longer. A valid reason reported in literature for this fact accounts for the fact that male body mass is comparatively larger than in women, and requires larger lungs and sufficient passage for air supply to the lungs. As a result, male nostrils are usually bigger, as it is evident for example in Fig. 5, column 4.

Hairstyle is another effective feature effective for gender discrimination. Despite its complicated geometry which varies from person to

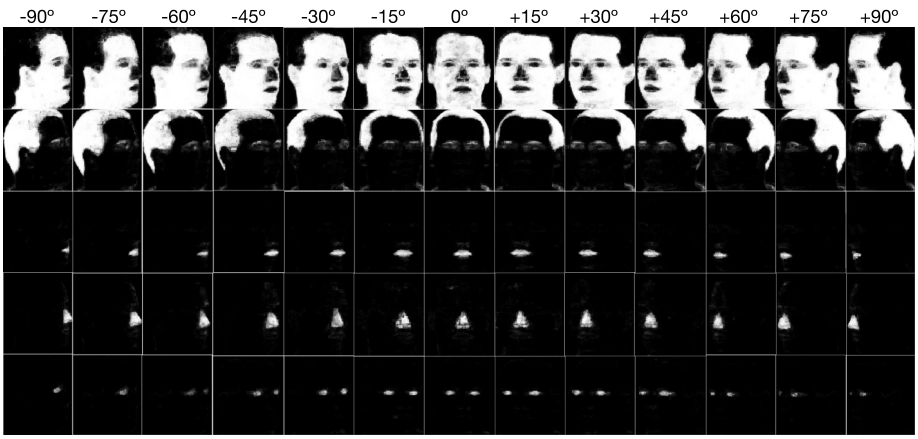

Fig. 4. Probability maps of one subject of Pointing'04. Poses vary from −90 to +90 with a step of 15°. In rows probability maps for: (1) skin, (2) hair, (3) mouth, (4) nose, and (5) eyes.

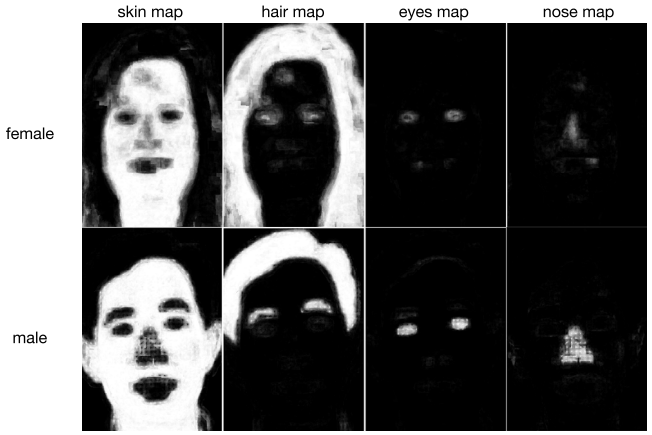

Fig. 5. Probability maps from FERET database used for gender recognition (row 1: female; row 2: male). Columns contain probability maps: (1) skin, (2) hair, (3) eyes, and (4) nose.

person, the proposed segmentation algorithm, getting 95.14% of pixel-wise accuracy on hair, is able to locate hair very precisely. Sometime eyebrows, which share the same hair label, also help in gender discrimination: on average female eyebrows are thinner and longer, while male eyebrows are shorter and wider.

Mouth is reported to be very gender distinctive in literature. Female lips are usually more clear as compared to male ones. In some cases the upper lip is completely missing in men. However our segmentation algorithm does not help much on this, since we do not notice any improvement in classification results using mouth information.

Thus the probabilistic RF classification approach is used to get probability maps for each of the four facial features: 'skin', 'nose', 'hair', and 'eyes'. These probability maps are then concatenated to form a single feature vector for each training image. A binary linear SVM classifier is trained to differentiate between the two genders. To measure the classification performance, we adopt 10-fold cross validation.

5.3. Expression classification

Our work on expression classification is limited to a binary classification problem i.e., differentiating between smiling and neutral facial expressions.

In general face parts endowed of dynamics are usually more relevant than static parts for the task of expression classification. Limiting the classification to two expressions only, we find out that mouth pixels are sufficient to distinguish whether a face is smiling or neutral (as those shown in Fig. 6), as reported in experiments in Section 6.

Table 1
Impact of D_{HSV} color parameter.

Features	Settings	Accuracy
HSV + LOC	$\mathbf{D_{HSV} = 16 \times 16}$, $N_{bins} = 16$	**92.27%**
HSV + LOC	$\mathbf{D_{HSV} = 32 \times 32}$, $N_{bins} = 16$	91.70%
HSV + LOC	$\mathbf{D_{HSV} = 64 \times 64}$, $N_{bins} = 16$	90.25%

Table 2
Impact of N_{bins} color parameter.

Features	Settings	Accuracy
HSV + LOC	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 16}$	92.03%
HSV + LOC	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 32}$	**92.27%**
HSV + LOC	$D_{HSV} = 16 \times 16$, $\mathbf{N_{bins} = 64}$	91.71%

Tests on expression classification are carried out by using two standard face databases: FEI and FERET. For the task of expression classification the RF classifier responsible for segmentation is first trained with a total of 60 manually labeled images, 30 from FEI [19] (15 smiling, 15 neutral) and 30 from FERET [20] (15 smiling, 15 neutral).

From experiments reported in Section 6, the probability map of mouth provides enough information to perform expression classification. Thus we first get probability maps for mouth using the probabilistic RF classification. Then a binary linear SVM classifier is trained to differentiate between the two expressions. To increase the amount of testing data and validate the model more precisely, we use 10-fold cross validation.

6. Experiments and discussion

6.1. HSV feature parameters

The HSV color feature has two important parameters to be considered: the patch dimension D_{HSV} on which the histogram is computed and the number of bins N_{bins} of the histogram itself. To evaluate the impact of both, a first stage of experiments is performed on `FASSEG-frontal01` as in [13], by only using location and color features and ignoring shape. We consider all the nine combinations of values from the sets $D_{HSV} = \{16 \times 16, 32 \times 32, 64 \times 64\}$ and $N_{bins} = \{16, 32, 64\}$. We find that the best accuracy - 92.27% - is achieved with $D_{HSV} = 16 \times 16$ and $N_{bins} = 32$. Results are reported in Tables 1 and 2.

6.2. HOG feature parameters

We then introduce HOG feature and run a second experiment to evaluate the impact of the patch dimension D_{HOG}. We find that the best accuracy - 92.95% - is achieved with $D_{HOG} = 64 \times 64$. Results are reported in Table 3.

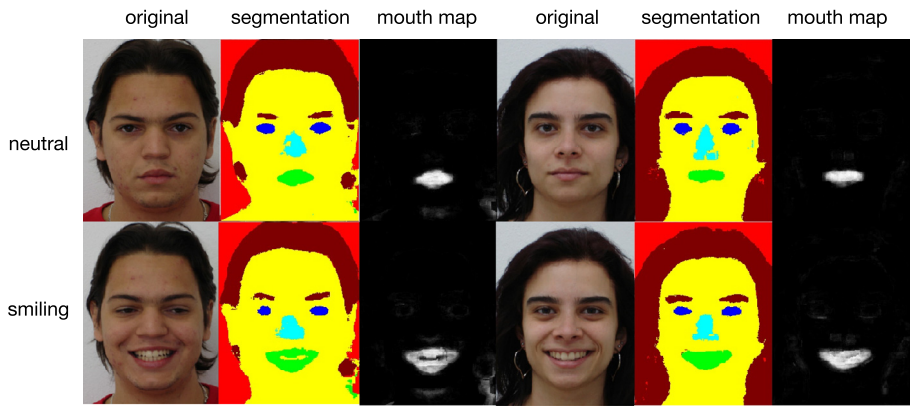

Fig. 6. Expression classification on two subjects. In columns: (1) original images; (2) segmented images; (3) probability maps for mouth; (4) original images; (5) segmented images; (6) probability maps for mouth.

Table 3
Impact of D_{HOG} shape parameter.

Features	Settings	Accuracy
HSV + HOG + LOC	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 16 \times 16}$	92.44%
HSV + HOG + LOC	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 32 \times 32}$	92.82%
HSV + HOG + LOC	$D_{HSV} = 16 \times 16$, $N_{bins} = 16$, $\mathbf{D_{HOG} = 64 \times 64}$	**92.95%**

6.3. Discussion on feature parameters

A few considerations emerge from the results. The first is that the right choice of features and parameters matters. Among all the configurations we experimented there is indeed a big difference between the worst – which has an accuracy of 79.89% (not shown here) – and the best, which achieves 92.95% and is obtained with $D_{HSV} = 16 \times 16$, $N_{bins} = 32$ (i.e., $f_{HSV} \in \mathbb{R}^{96}$), $D_{HOG} = 64 \times 64$ (i.e., $f_{HOG} \in \mathbb{R}^{1764}$), which returns the final feature vector $f^I \in \mathbb{R}^{1862}$.

A second observation is that the HOG feature boosts the accuracy from 92.27% to 92.95%. This may seem a small improvement, but it corresponds to a 9% reduction in error rate. Moreover, the classes which benefit the most from the introduction of HOG are 'eyes', 'nose' and 'mouth', which are mostly distinguishable from their shape, as shown in [13]. Since these classes are the least frequent, they have a small impact on the accuracy despite their importance for further face analysis tasks.

6.4. Head pose estimation

In our experiments on head pose estimation we use two evaluation measures: Mean Absolute Error (MAE) and Pose Estimation Accuracy (PEA). MAE is a regression measure which returns the absolute error between the estimated pose angle and the ground truth angle, averaged on the whole test set [32]. PEA is a classification measure which, considering every pose angle as a different class, corresponds to the ratio of the number of correct prediction to the total number of predictions (simply referred to as *accuracy* in other works e.g., [36,38]).

By trying all possible combinations of six facial features 'skin', 'hair', 'eyes', 'nose', 'mouth', and 'background', best results (MAE=2.79° and avg. PEA=81%) are obtained by concatenating five facial features, that are 'skin', 'hair', 'eyes', 'nose', and 'mouth'. For better inspecting HPE results, our findings are also presented as confusion matrix in Fig. 8, in which it is evident that most errors are concentrated in contiguous classes (i.e., close to the diagonal), with a degradation of performance in classes +60° and +75°.

For assessing how much each facial part is specific for the task, we exploit the feature importance measure returned by the Random Forest implementation in [56], which evaluates how much each feature decreases the weighted impurity in a tree. The feature importances of the five facial parts used for the head pose estimation task are given in Fig. 7(a) in absolute values.

Table 4
Head Pose Estimation comparison of various methods on Pointing'04 database.

Method	Reference	MAE (°)	PEA (%)
Our approach	–	**2.79**	**81.00**
ML Distribution	MLD-wJ [38]	4.24	73.30
MG Derivative	Jain et al. [36]	6.90	64.51
kCovGa	Bingpeng et al. [34]	6.24	–
CovGa	Bingpeng et al. [34]	7.27	–
Neural Networks	Stifelhagen et al. [33]	9.50	52.00
Random forest regression	Li et al. [29]	9.60	–
Human performance	Gourier et al. [57]	11.40	40.70
High-order SVD	Tu et al. [58]	12.90	49.25
CNN	Lee et al. [40]	5.17	69.88
CNN adaptive grad.	Patacchiola et al. [41]	7.74	66.60

A comparison with state-of-the-art is instead given in Table 4. From this table, it is clear that our reported results outperform previous results for both MAE and PEA. Even when considering recent approaches such as those adopting convolutional neural networks such as [40] and [41], our classical approach with hand-crafted features still constitutes the best solution to the problem, probably because the problem is within a limited data scenario, while a full deep learning case would need more data to be fully effective.

Please notice that some of the previous results may be obtained with different experimental setup such as validation protocols. For example MLD methods are validated by 5-fold cross validation experiments. For more details about each algorithm and experimental setup, the corresponding papers can be explored.

To better compare results with state-of-the-art, we also report PEA by other methods on different poses. As shown in Fig. 9 our PEA is higher on all angles except at +30°, +60°, and +75°, while MLD shows comparatively better results on extreme right profile poses (+60° and +75°).

Similarly, Fig. 10 shows a comparison on MAE with previously reported methods under all 13 pose angles: our algorithm is again the best on all poses except +30° and +75° angles. As compared to PEA results, our MAE performance is more impressive. The MLD results are good on frontal poses but as orientation of the pose changes from frontal to profile, its MAE increases. Finally reported results for methods MGD [36] and kVoD [35] are less uniform, and abrupt changes can be seen in the form of spikes in Fig. 10.

6.5. Gender recognition

By trying all possible combinations of six facial features 'skin', 'hair', 'eyes', 'nose', 'mouth', and 'background', best results are obtained by

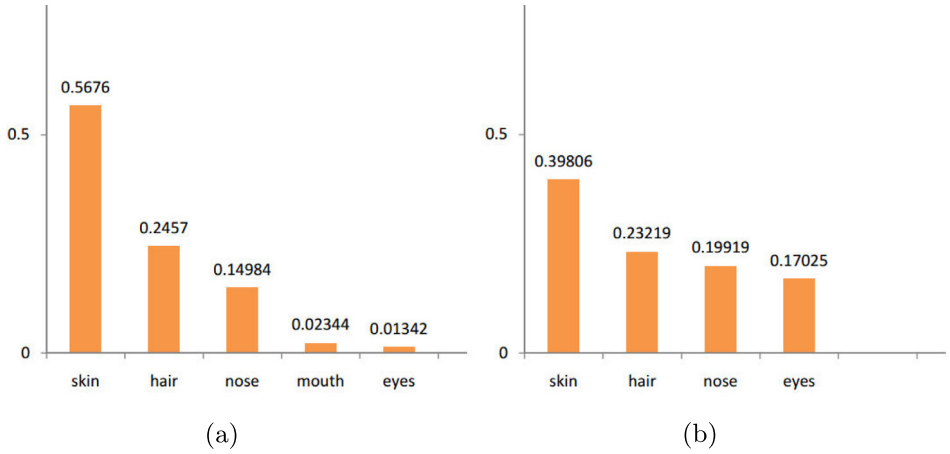

Fig. 7. (a) Feature importances of 'skin', 'hair', 'eyes', 'nose', and 'mouth' for head pose estimation. (b) Feature importances of 'skin', 'hair', 'eyes', and 'nose' for gender recognition.

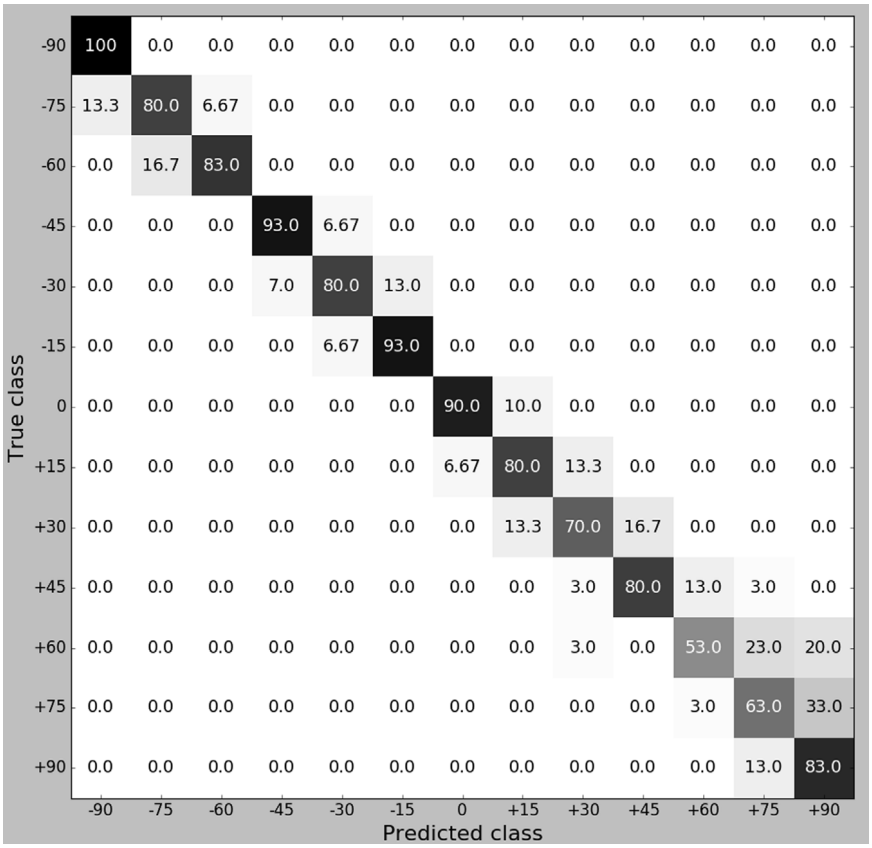

Fig. 8. HPE confusion matrix (%) for Pointing'04 database on the yaw angles.

concatenating four facial features, that are 'skin', 'hair', 'eyes', and 'nose'. Tests are performed on 800 frontal images, 400 from FEI and 400 from FERET databases. The obtained confusion matrices for gender recognition on FEI and FERET are reported in Fig. 11, with similar performance on the two datasets.

For assessing how much a certain facial part (among those four used for GR) is specific, we exploit the feature importance measure returned by the Random Forest implementation in [56]. Fig. 7(b) shows the feature importances of the four employed facial parts for the gender recognition task in absolute values.

Table 5 shows instead a comparison with state-of-the-art performed on the same set of images, and evaluated in terms of Classification Rate (CR), which reveals how many images are correctly classified by the framework.

A fair comparison is however hard to achieve, for many reasons, since authors use different image settings. For example Preeti et al. [59] use the same set of images, but the validation protocol is 2-fold.

In order to obtain a precise statistical significant comparison of our methods with respect to others in Table 5, we would need to estimate confidence intervals in the performance metric by running all the methods many times by randomizing data subsets and/or random seeds of the involved algorithms. While we can do this for our approach, this is however not possible at the moment for other competing algorithms, since this would require to reimplement the other methods from scratch.

In any case Table 5 shows that classification results achieved with the proposed algorithm are in general better or at least competitive with respect to those reported, except for one case (on FEI database), where [60] outperforms our approach by 0.5%.

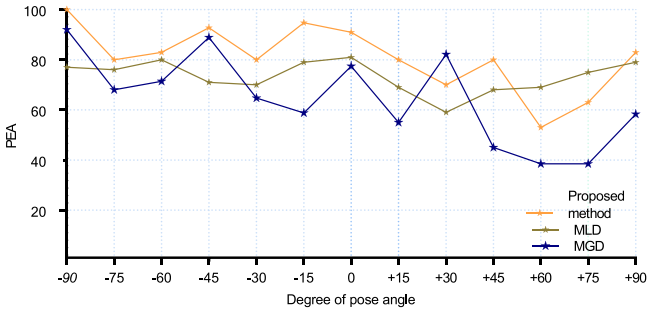

Fig. 9. Comparison of PEA by different algorithms and different poses on Pointing'04 database (averages are in Table 4).

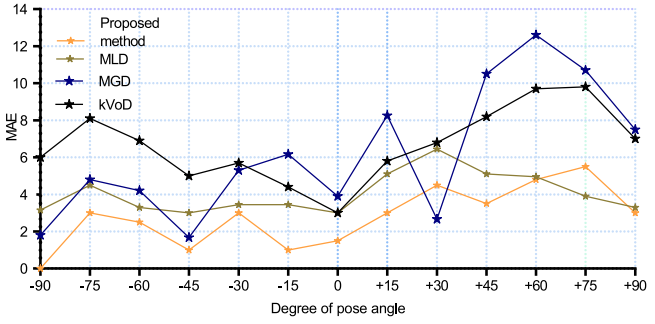

Fig. 10. Comparison of MAE by different algorithms and different poses on Pointing'04 database (averages are in Table 4).

An additional comparative test is carried out on the MPI dataset and the method used in [21]. To perform the experiment we rescale all images to the same size as in our approach ($H = 512$ pixels, while width W is varied accordingly to keep the original image ratio). All 200 resized images of MPI dataset are used for testing, using the model trained on the FASSEG database. The best result reported in [21] in various experiments is 94% in accuracy. On the same dataset we obtain a performance of 96.4% on the same indicator. It should be noted that since in the training session we do not use any image of the MPI dataset, better results could be expected if some images from the MPI dataset were included in the training phase.

The proposed gender recognition algorithm is also robust to illumination and expression variations. However, further investigations are needed to know how much CR will drop with change of pose or with more complicated facial expression.

Table 5
Comparative experiments on gender recognition.

Method	Database	CR (%)
A priori-driven PCA [60]	FEI	≈99.00
Proposed approach	**FEI**	**98.50**
2D Gabor+2DPCA [59]	FEI	96.61
Proposed approach	**FERET**	**96.50**
A priori-driven PCA [60]	FERET	≈84.00
Proposed approach	**MPI**	**96.40**
Multi-feat. K-means [21]	MPI	94.00

Table 6
Comparative experiments on expression classification.

Method	Database	CR (%)
Proposed approach	**FEI**	**96.25**
A priori-driven PCA [60]	FEI	≈95.00
Proposed approach	**FERET**	**83.25**
A priori-driven PCA [60]	FERET	≈74.00

6.6. Expression recognition

By trying all possible combinations of six facial features, best results are obtained by using only 'mouth' class. Tests are performed on 800 frontal images, 400 from FEI and 400 from FERET databases. The related confusion matrices are reported in Fig. 12, where best performance is obtained on FEI dataset.

As for the previous task, we compare our algorithm with state-of-the-art methods on the same set of images. In Table 6 methods are evaluated in terms of Classification Rate (CR), showing that our methods provides the best results till date on the same set of images and same databases, obtaining an average CR of 96.25% on FEI and 83.25% on FERET.

As in the previous case, this comparative analysis suffers by some limitations on the statistical estimation of confidence intervals in the performance of obtained results, which could be overcome only by reimplementing all other methods from scratch.

As another limitation to this study, six facial expression are usually studied in literature (neutral, smiling, anger, fear, sadness, and surprise). Since we limit our investigation to smiling and neutral expressions only, these experiments should be considered as a corollary to the other classification tasks as their purpose is not a detailed insight on facial expression recognition, but demonstrating that mid-level vision problems (such as pose and gender classification) are better solved starting from accurately segmented facial images.

6.7. Towards fully automated face analysis

The images employed in previous experiments are almost centered with respect to the background and with small background area. In

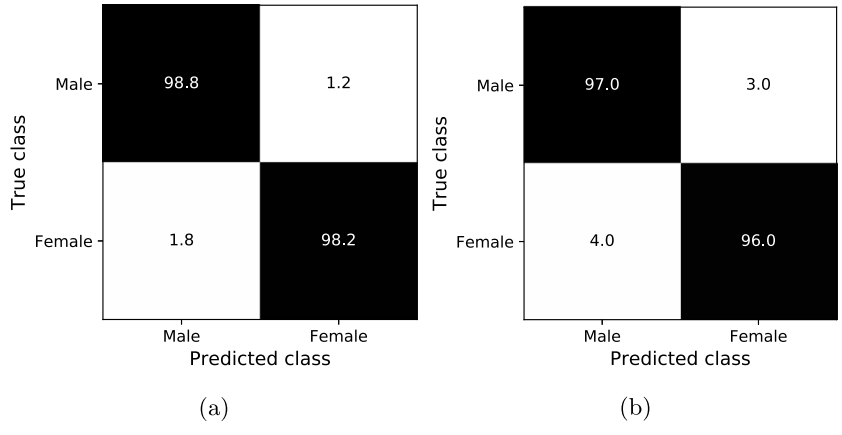

Fig. 11. (a) Confusion matrices obtained on gender recognition task on (a) FEI database and (b) FERET database.

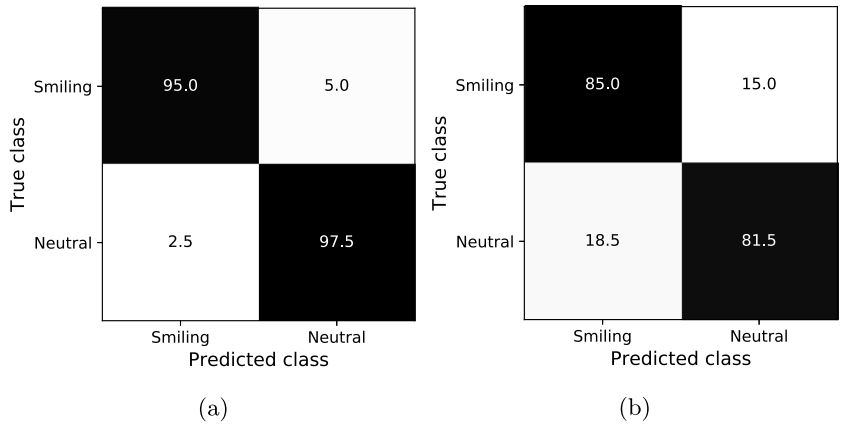

Fig. 12. (a) Confusion matrices obtained on expression classification task on (a) FEI database and (b) FERET database.

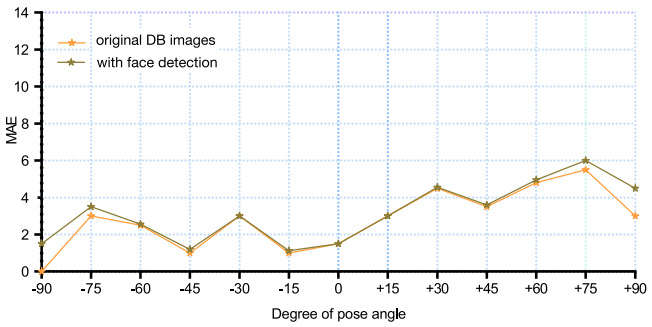

Fig. 13. Comparison of MAE on Pointing'04 database between the algorithm working on original images and the same images in their non-normalized version (by automatic face detection and normalization).

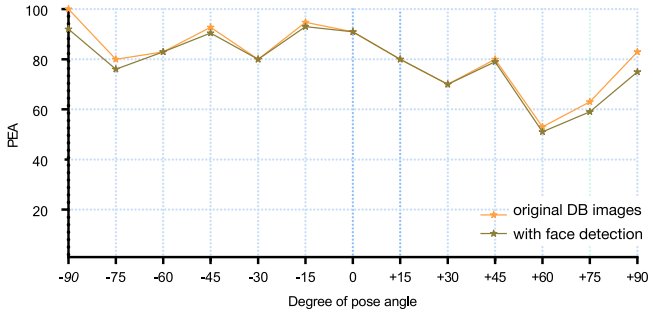

Fig. 14. Comparison of PEA on Pointing'04 database between the algorithm working on original images and the same images in their non-normalized version (by automatic face detection and normalization).

general, there are also small variations in face scales. Therefore, as an attempt to automatize the process of face analysis also to non-normalized and non-centered images, we show how the application of a face detection algorithm, followed by a size normalization step, does not severely affect the performance of the three face analysis tasks.

Since Pointing'04 and FEI databases are also available in their non-normalized versions, we consider them for these preliminary experiments. In particular for head pose estimation we use the non-normalized Pointing'04, while we perform tests on expression and gender recognition on the non-normalized FEI database. For face detection we use the algorithm described in [61], and we then normalize the face bounding-box to the fixed dimension suitable for feature extraction (height of $H = 512$ pixels, while width W is varied accordingly to keep the original image ratio) so as the content for a given patch dimension is again comparable across different face images.

As all face images are captured in a controlled lab environment, we expect little effect from the introduction of the face detector in the framework — particularly in the cases of gender recognition and expression classification. However, for head pose estimation of extreme left and right profile images, face detection will be probably not as efficient as in the frontal case.

In fact, for what concerns head pose estimation, small drops in both PEA and MAE are observed, as shown in Figs. 13 and 14. As expected, the loss of performance is less in the case of frontal faces, when the face detector works best. Conversely, for extreme profile images, faces are not as accurately detected as compared to frontal images, hence performance of the head pose estimation drops in terms of both MAE and PEA to some extent.

For expression classification, the fact that all images in FEI database are frontal causes an almost negligible difference in performance with the case not employing a face detector, as can be seen in the confusion matrix in Fig. 15(a).

Conversely, a small drop in the case of gender recognition is registered, probably since in this case four face parts are involved in the classification procedure (versus five parts used for head pose), hence the lower performance as depicted in Fig. 15(b).

In conclusion the proposed face analysis algorithms are apparently suitable for being applied also in the non-normalized case i.e., working on the resized bounding-box derived from a face detection process, with no substantial loss.

7. Conclusion and future work

In this paper we propose a method which tries to solve the three challenging problems of pose estimation, gender recognition, and expression classification, in a unified framework. We have also extended and made publicly available two novel subsets of FASSEG, a face database which contains segmentation masks of frontal and multi-pose images. The proposed methods are compared with state-of-the-art on standard databases obtaining results which are competitive or superior to state-of-the-art. Regarding the method, a face segmentation algorithm first segments a face image into six semantic classes. The probabilistic classification strategy used in segmentation is also exploited to generate probability maps for each semantic class. Experiments are carried out to identify which probability maps are more helpful for pose estimation, gender recognition, and expression classification; the idea is to use them to generate feature vectors to train a linear SVM classifier for each task. Our segmentation results provide sufficient information for various hidden variables in face and provide a route towards solving more difficult classification problems. We are planning to add more complicated face expressions to our framework. Our future work also

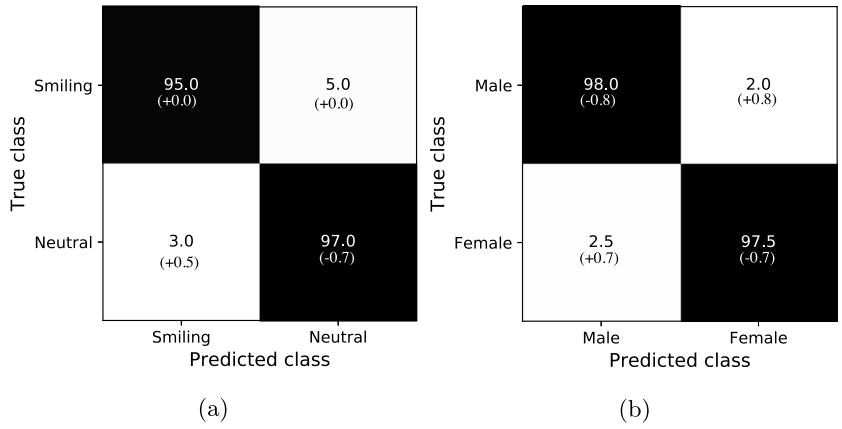

Fig. 15. Confusion matrices obtained for (a) the expression classification task and for (b) the gender recognition task, both using automatic face detection (and normalization) on the non-normalized version of the FEI database. Differences with respect to results obtained with the original dataset are shown in brackets.

includes: an investigation on the problems of age and race estimation; the improvement of the segmentation model by integrating the pixel probabilities with Conditional Random Fields (CRF) to enhance labeling consistency; the full automatization of the face analysis pipeline.

References

[1] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, http://dx.doi.org/10.1007/s11263-009-0275-4.

[2] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2019) 121–135, http://dx.doi.org/10.1109/TPAMI.2017.2781233.

[3] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[4] S.L. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE Trans. Affective Comput. 6 (1) (2015) 1–12, http://dx.doi.org/10.1109/TAFFC.2014.2386334.

[5] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 483–499.

[6] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S.K. Nayar, Face swapping: automatically replacing faces in photographs, ACM Trans. Graph. 27 (3) (2008) 39:1–39:8, http://dx.doi.org/10.1145/1360612.1360638.

[7] Laan Labs, Face swap live, 2016, http://faceswaplive.com (accessed January 24, 2018).

[8] S. Benini, M. Svanera, N. Adami, R. Leonardi, A.B. Kovács, Shot scale distribution in art films, Multimedia Tools Appl. 75 (23) (2016) 16499–16527, http://dx.doi.org/10.1007/s11042-016-3339-9.

[9] Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen, Affective video content analysis: a multidisciplinary insight, IEEE Trans. Affective Comput. 9 (4) (2018) 396–409, http://dx.doi.org/10.1109/TAFFC.2017.2661284.

[10] L. Canini, S. Benini, R. Leonardi, Affective analysis on patterns of shot types in movies, in: 2011 7th International Symposium on Image and Signal Processing and Analysis, ISPA, 2011, pp. 253–258.

[11] G. Davies, H. Ellis, J. Shepherd, Perceiving and remembering faces, Academic Press, 1981.

[12] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: Nineteen results all computer vision researchers should know about, Proc. IEEE 94 (11) (2006) 1948–1962, http://dx.doi.org/10.1109/JPROC.2006.884093.

[13] K. Khan, M. Mauro, R. Leonardi, Multi-class semantic segmentation of faces, in: 2015 IEEE International Conference on Image Processing, ICIP, 2015, pp. 827–831, http://dx.doi.org/10.1109/ICIP.2015.7350915.

[14] K. Khan, M. Mauro, P. Migliorati, R. Leonardi, Head pose estimation through multi-class face segmentation, in: Proc. ICME-2017, IEEE, 2017, pp. 253–258.

[15] K. Khan, M. Mauro, P. Migliorati, R. Leonardi, Gender and expression analysis based on semantic face segmentation, in: Image Analysis and Processing - ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, 2017.

[16] S. Benini, K. Khan, M. Mauro, R. Leonardi, M. Pierangelo, FASSEG: The face semantic segmentation repository, mendeley data, v1, 2019, http://dx.doi.org/10.17632/8yjv7b3hgr.1.

[17] G.B. Huang, M. Narayana, E.G. Learned-Miller, Towards unconstrained face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2008, Anchorage, AK, USA, 23-28 June, 2008, IEEE Computer Society, 2008, pp. 1–8, http://dx.doi.org/10.1109/CVPRW.2008.4562973.

[18] N. Gourier, D. Hall, J.L. Crowley, Estimating face orientation from robust detection of salient facial structures, in: FG Net Workshop on Visual Observation of Deictic Gestures, FGnet (IST–2000–26434) Cambridge, UK, 2004, pp. 1–9.

[19] Centro Universitario da FEI, FEI face database, http://www.fei.edu.br/~cet/facedatabase.html.

[20] Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image Vis. Comput. 16 (5) (1998) 295–306, http://dx.doi.org/10.1016/s0262-8856(97)00070-x.

[21] A. Nestor, M.J. Tarr, The segmental structure of faces and its use in gender recognition, J. Vision 8 (7) (2008) 1–12, http://dx.doi.org/10.1167/8.7.7.

[22] A. Kae, K. Sohn, H. Lee, E.G. Learned-Miller, Augmenting crfs with Boltzmann Machine Shape Priors for Image Labeling, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, IEEE Computer Society, 2013, pp. 2019–2026, http://dx.doi.org/10.1109/CVPR.2013.263.

[23] S.M.A. Eslami, N. Heess, C.K.I. Williams, J.M. Winn, The shape boltzmann machine: A strong model of object shape, Int. J. Comput. Vis. 107 (2) (2014) 155–176, http://dx.doi.org/10.1007/s11263-013-0669-1.

[24] M. Svanera, U.R. Muhammad, R. Leonardi, S. Benini, Figaro, hair detection and segmentation in the wild, in: 2016 IEEE International Conference on Image Processing, ICIP, 2016, pp. 933–937, http://dx.doi.org/10.1109/ICIP.2016.7532494.

[25] U.R. Muhammad, M. Svanera, R. Leonardi, S. Benini, Hair detection, segmentation, and hairstyle classification in the wild, Image Vis. Comput. (2018).

[26] Y. Yacoob, L.S. Davis, Detection and analysis of hair, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1164–1169, http://dx.doi.org/10.1109/TPAMI.2006.139.

[27] C. Lee, M.T. Schramm, M. Boutin, J.P. Allebach, An algorithm for automatic skin smoothing in digital portraits, in: 2009 16th IEEE International Conference on Image Processing, ICIP, 2009, pp. 3149–3152, http://dx.doi.org/10.1109/ICIP.2009.5414430.

[28] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289, http://dl.acm.org/citation.cfm?id=645530.655813.

[29] Y. Li, S. Wang, X. Ding, Person-independent head pose estimation based on random forest regression, in: 2010 IEEE International Conference on Image Processing, 2010, pp. 1521–1524, http://dx.doi.org/10.1109/ICIP.2010.5652915.

[30] C. Scheffler, J.-M. Odobez, Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2011, pp. 53.1–53.11, http://dx.doi.org/10.5244/C.25.53.

[31] M. Ferrara, A. Franco, D. Maio, A multi-classifier approach to face image segmentation for travel documents, Expert Syst. Appl. 39 (9) (2012) 8452–8466.

[32] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626.

[33] R. Stiefelhagen, Estimating head pose with neural networks, results on the pointing'04 ICPR Workshop Evaluation Data, in: Pointing'04 ICPR Workshop of the Int. Conf. on Pattern Recognition, 2004.

[34] B. Ma, A. Li, X. Chai, S. Shan, Covga: a novel descriptor based on symmetry of regions for head pose estimation, Neurocomputing 143 (2014) 97–108.

[35] B. Ma, R. Huang, L. Qin, VoD: A novel image representation for head yaw estimation, Neurocomputing 148 (2015) 455–466.

[36] V. Jain, J.L. Crowley, Head pose estimation using multi-scale gaussian derivatives, in: 18th Scandinavian Conference on Image Analysis, Espoo, Finland, Springer, 2013, pp. 319–328.

[37] K. Hara, R. Chellappa, Growing regression forests by classification: applications to object pose estimation, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 552–567.

[38] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1837–1842, http://dx.doi.org/10.1109/CVPR.2014.237.

[39] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, IEEE Computer Society, Washington, DC, USA, 2005, pp. 886–893, http://dx.doi.org/10.1109/CVPR.2005.177.

[40] S. Lee, T. Saitoh, Head pose estimation using convolutional neural network, in: K.J. Kim, H. Kim, N. Baek (Eds.), IT Convergence and Security 2017, Springer Singapore, Singapore, 2018, pp. 164–171.

[41] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recognit. 71 (2017) 132–143, http://dx.doi.org/10.1016/j.patcog.2017.06.009.

[42] E. Mäkinen, R. Raisamo, An experimental comparison of gender classification methods, Pattern Recognit. Lett. 29 (10) (2008) 1544–1556, http://dx.doi.org/10.1016/j.patrec.2008.03.016.

[43] N. Kumar, P.N. Belhumeur, S.K. Nayar, Facetracer: a search engine for large collections of images with faces, in: The 10th European Conference on Computer Vision, ECCV, 2008.

[44] J. Zheng, B.-L. Lu, Asupport vector machine classifier with automatic confidence and its application to gender classification, Neurocomputing 74 (11) (2011) 1926–1935, http://dx.doi.org/10.1016/j.neucom.2010.07.032.

[45] J. Bekios-Calfa, J.M. Buenaposada, L. Baumela, Revisiting linear discriminant techniques in gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2011) 858–864, http://dx.doi.org/10.1109/TPAMI.2010.208.

[46] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 3730–3738, http://dx.doi.org/10.1109/ICCV.2015.425.

[47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, PANDA: Pose aligned networks for deep attribute modeling, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 1637–1644, http://dx.doi.org/10.1109/CVPR.2014.212.

[48] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models - their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59, http://dx.doi.org/10.1006/cviu.1995.1004.

[49] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, vol. 2, 2005, pp. 568–573, http://dx.doi.org/10.1109/CVPR.2005.297.

[50] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816, http://dx.doi.org/10.1016/j.imavis.2008.08.005.

[51] S. Chen, Y. Tian, Q. Liu, D.N. Metaxas, Recognizing expressions from face and body gesture by temporal normalized motion and appearance features, Image Vision Comput. 31 (2) (2013) 175–185, http://dx.doi.org/10.1016/j.imavis.2012.06.014.

[52] T.H.H. Zavaschi, A.S. Britto Jr., L.E.S. Oliveira, A.L. Koerich, Fusion of feature sets and classifiers for facial expression recognition, Expert Syst. Appl. 40 (2) (2013) 646–655, http://dx.doi.org/10.1016/j.eswa.2012.07.074.

[53] M.I.T. Center for Biological and Computational Learning (CBCL) MIT-CBCL database, http://cbcl.mit.edu/software-datasets/FaceData2.html.

[54] S. Bochkanov, ALGLIB, http://www.alglib.net.

[55] D. Enlow, R. Moyers, W. Merow, Handbook of Facial Growth, Saunders, 1982, https://books.google.it/books?id=SA5NAQAAIAAJ.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[57] N. Gourier, J. Maisonnasse, D. Hall, J.L. Crowley, Head pose estimation on low resolution images, in: Multimodal Technologies for Perception of Humans, Springer, 2007, pp. 270–280.

[58] J. Tu, Y. Fu, Y. Hu, T. Huang, Evaluation of head pose estimation for studio data, in: Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006, Revised Selected Papers, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 281–290, http://dx.doi.org/10.1007/978-3-540-69568-4_25.

[59] P. Rai, P. Khanna, An illumination, expression, and noise invariant gender classifier using two-directional 2DPCA on real Gabor space, J. Vis. Lang. Comput. 26 (Suppl. C) (2015) 15–28, http://dx.doi.org/10.1016/j.jvlc.2014.10.016.

[60] C. Thomaz, G. Giraldi, J. Costa, D. Gillies, A priori-driven PCA, in: Computer Vision-ACCV 2012 Workshops, Lecture Notes in Computer Science, Springer, 2013, pp. 236–247.

[61] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 21–37.