TSINGHUA SCIENCE AND TECHNOLOGY
ISSNll1007-0214 08/10 pp333–345
DOI: 10.26599/TST.2018.9010090
Volume 24, Number 3, June 2019

Exploiting Effective Facial Patches for Robust Gender Recognition

Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang*

Abstract: Gender classification is an important task in automated face analysis. Most existing approaches for gender classification use only raw/aligned face images after face detection as input. These methods exhibit fair classification ability under constrained conditions, in which face images are acquired under similar illumination with similar poses. The performances of these methods may deteriorate when face images show drastic variances in poses and occlusion as routinely encountered in real-world data. The reduction in the performances of current gender classification methods may be attributed to the sensitiveness of features to image translations. This work proposes to alleviate this sensitivity by introducing a majority voting procedure that involves multiple face patches. Specifically, this work utilizes a deep learning method based on multiple large patches. Several Convolutional Neural Networks (CNN) are trained on individual, predefined patches that reflect various image resolutions and partial cropping. The decisions of each CNN are aggregated through majority voting to obtain the final gender classification accurately. Extensive experiments are conducted on four gender classification databases, including Labeled Face in-the-Wild (LFW), CelebA, ColorFeret, and All-Age Faces database, a novel database collected by our group. Each individual patch is evaluated, and complementary patches are selected for voting. We show that the classification accuracy of our method is comparable with that of state-of-the-art systems. This characteristic validates the effectiveness of our proposed method.

Key words: gender classification; Convolutional Neural Network (CNN); majority voting

1 Introduction

This work considers the task of gender classification. Given an input face image, a binary decision is rendered, i.e., male or female. Gender classification is a long-studied task and has extensive applications in identity authentication, suspect tracking, and other fields. It is a challenging task because face images exhibit extensive variations in illumination conditions, poses, and other features.

Numerous methods for gender classification have been proposed over the past several decades. Classical methods include the Artificial Neural Network (ANN)[1], EigenFace[2], and Linear Discriminant Analysis (LDA)[3,4]. ANN, EigenFace, and LDA correspond to the three mainstream methods of deep learning, eigenvector, and feature mapping, respectively. These classical methods exhibit impressive accuracy on strictly constrained databases, such as Face Recognition Technology (FERET)[5], PIE[6], and some self-collected databases. When tested on real-life face images wherein faces are not all frontally and similarly posed, however, their performances drastically deteriorate[7].

Image variations caused by pose changes and occlusion are crucial and frequently encountered challenges in face gender classification. In recent years, benchmarks with increased realism, such as Labeled Face in-the-Wild (LFW) and CelebA, have emerged in the gender classification task. New recognition methods have also been proposed, including Facial Attributes[7],

• Jingchun Cheng, Yali Li, Jilong Wang, and Shengjin Wang are with Tsinghua University, Beijing 100084, China. E-mail: chengjingchun14@163.com; liyali@ocrserv.ee.tsinghua.edu.cn; wjl@cernet.edu.cn; wgsgj@tsinghua.edu.cn.

• Le Yu is with China Mobile Information Security Center, Beijing 100084, China.

∗ To whom correspondence should be addressed.

Manuscript received: 2018-01-02; accepted: 2018-03-01

AdaBoost[8], and deep learning[9, 10]. These methods improve the performance of gender classification on in-the-wild databases to a large extent. Nevertheless, most of these methods recognize gender on the basis of a single input image. This approach may easily result in erroneous detection when poses change or occlusion occurs.

In this work, we attempt to address the problems posed by perspective changes and occlusion. We show the framework in Fig. 1, examples of errors that can be corrected by our proposed method in Fig. 2, and the differences between face images captured under constrained/unconstrained conditions in Fig. 3. We demonstrate that training multiple Convolutional Neural Networks (CNNs) based on face patches with multiple cropping positions improve the accuracy of gender classification (Fig. 1). We define front-face, left-face, right-face, and lower-face patches for which 11, 4, 4, and 2 different patches are cropped, respectively. During training, four CNN models are trained with the four types of face patches comprising multiple crops. For testing, given an input face image, the 21 patches are cropped and fed into their corresponding CNN models for gender recognition. The final classification result is obtained on the basis of the majority voting of the 21 binary decisions. We have collected and annotated All-Age Faces (AAF), a new gender classification database. It is used as validation set for the selection of hyper-parameters in the CNN models. We show that the accuracy of our proposed method on three test databases, i.e., LFW, CelebA, and ColorFeret, is comparable with that of state-of-the-art methods. To summarize, our main contributions are as follows:

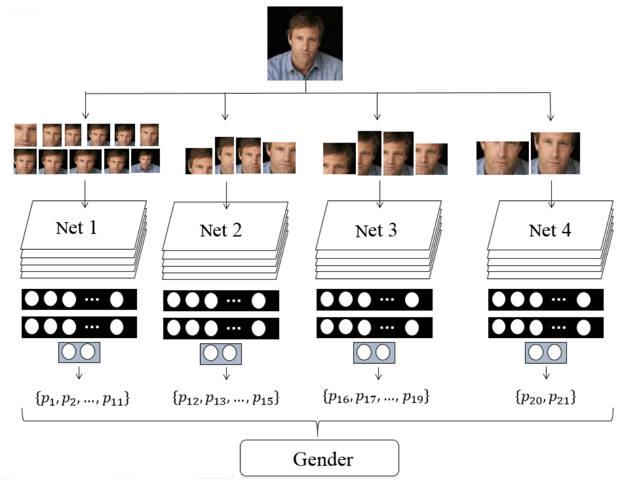

Net 1
Net 2
Net 3
Net 4
$\{p_1, p_2, \ldots, p_{11}\}$
$\{p_{12}, p_{13}, \ldots, p_{15}\}$
$\{p_{16}, p_{17}, \ldots, p_{19}\}$
$\{p_{20}, p_{21}\}$
Gender

Fig. 1 Framework of the proposed classifier. The candidate face image generates patches of different parts. The patches enter trained nets correspondingly. Nets 1, 2, 3, and 4 are designed for whole faces, right parts, left parts, and lower parts, respectively. The results for all patches vote for the final classification output.

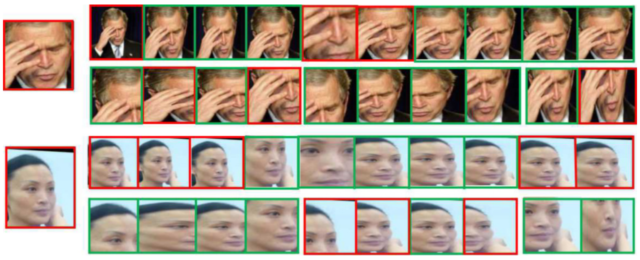


Fig. 2 Visual example of how the proposed method corrects the classification error caused by occlusion and perspective changes in a single image. Correct and incorrect predictions are shown in green and red boxes, respectively. Two automatically detected and aligned input test images are shown on the left-hand side of the image. These two images were predicted incorrectly by our baseline system, i.e., a single CNN model trained with the aligned images. The images of the right-hand side are the classification results of the 21 patches and the original face image. After majority voting, our system yielded the correct classification results.

- A set of complementary large patches that help boost gender classification accuracy by working cooperatively.
- A novel gender classification method based on multiple face patches and CNN that is robust to occlusion and view changes.
- A new gender classification database named AAF comprising the face images of individuals over a broad age range (from 2 years to 80 years). This database will be released for public research.

This paper is organized as follows. The related works

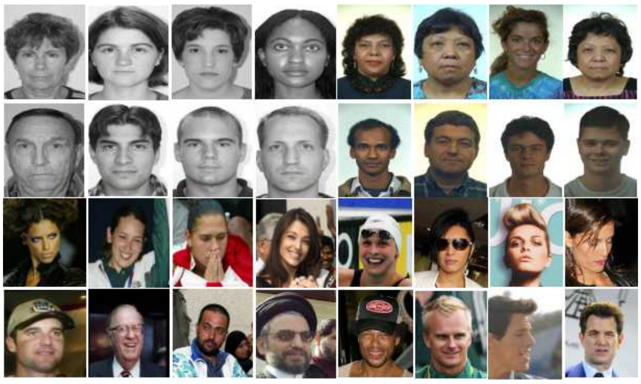


Fig. 3 Examples of face images captured under constrained (row 1 and row 2) and unconstrained conditions (row 3 and row 4). Constrained face images are typically frontally and similarly posed, whereas unconstrained images exhibit different lighting conditions, poses, and occlusions.

are introduced in Section 2. The proposed gender classification method is described in detail in Section 3. The experimental results are presented and summarized in Section 4. The conclusion is presented in Section 5.

2 Related Work

A number of works exist on gender classification. These works typically focused on the design of classifiers and face representations. Many classical classifiers, e.g., K-Nearest Neighbor (KNN)[11], Decision Tree (DT)[12], Support Vector Machine (SVM)[13], and softmax regression (Softmax)[14] are applied in gender classification. KNN[11] is a supervised learning algorithm wherein the classification result of a test sample is decided by the majority of the K-nearest neighbor category. KNN aims to classify a new sample on the basis of feature similarities and a training set. KNN does not require a training process given that its classification process is based on a training gallery. DT[12] is tree-structured classifier that consists of three kinds of nodes, i.e., decision node (root), chance node (internal node), and end node (leaf). The root of DT (decision node) takes in the features to be classified; the internal node (chance node) represents a *test* of an attribute (usually a certain dimension of an input feature), and its corresponding branch denotes the outcome of this node (usually a binary decision); and each leaf (end node) designates a class label. The paths from root to leaf represent the classification rules of DT learned in training, wherein internal nodes are created when the separation of different attributes considerably increases the likelihood of the training data. The SVM[13] classifier is based on statistical learning theory and the structural risk minimization principle. Its basic premise is to learn the optimal margin between classes. In training, data points are mapped to a new feature space with linear/nonlinear transformation (called kernel functions), such that points of different categories can be divided by a clear gap (margin) that is as wide as possible. Test points are then mapped in the same manner and classified in accordance with the learned between-class margin. Softmax[14] is a classifier that generalizes softmax regression (similar to logistic regression which can split data points by its learnt distribution) to multiclass categorization. By minimizing its cost function in training, Softmax learns some problem-specific parameters that can be used to determine the probability of each particular outcome of the dependent variable.

Various feature representations are used in gender classification[1–3, 8–10, 15–17]. These representations include traditional image descriptors (such as Histogram of Oriented Gradient (HOG), Local Binary Patterns (LBP), and Gabor wavelet feature), bioinspired features (like Biologically Inspired Features (BIF), Eigen Face, and Attributes), and neural network features. These representations are sometimes followed by mapping (for example, Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA)) and have been proven to be effective in the gender recognition task. For example, Guo et al.[17] showed that the LBP feature with nonlinear SVM classifier yields an accuracy of 90.53% in the gender classification of the face images of individuals over a broad age range. Valentin et al.[2] proposed the Eigen Face representation, which was able to accurately classify the gender of approximately 94% of frontal face images. Golomb et al.[1] proposed the extraction of ANN features with a three-layer network. This approach achieved a precision of 93% on 8 × 8 face images collected by the researchers.

The currently most popular methods are predominantly based on CNN, an end-to-end structure that learns feature representation and classification. By feature learning with several long-tested structures, such as LeNet[18], AlexNet[19], and GoogleNet[20], CNN has been widely studied, and its performance in numerous visual applications, including image classification, image retrieval[21], and object detection[22], has been greatly improved. CNN has also improved the accuracy of gender classification[9, 10]. For example, Fan et al.[9] proposed a CNN structure that can jointly optimize the compactness and discriminative ability of representation for face images. The learned compact representation achieves a classification accuracy of 96.8% in a gender classification task on the LFW database. Liu et al.[10] cascaded two CNNs (AlexNet and LeNet) to train a network specifically for gender recognition. The two nets are pretrained differently: AlexNet is pretrained with massive face identities for attribute prediction, whereas LeNet is pretrained with massive general object categories for face localization. The two nets are then jointly fine-tuned with gender tags and achieve a state-of-the-art accuracy of 98% on CelebA database. Most existing methods use a single input, such as an aligned face image, to recognize gender. This approach, however, renders the methods vulnerable to occlusion or rotation. In this work, we propose a CNN-based gender classification framework with large face patches to alleviate this problem. We use the AlexNet structure to train a number of gender classification CNNs. Various large face patches are selected to provide complementary information to the recognition process. We show that the proposed method exhibits competitive performance on several benchmark databases.

3 Proposed Approach

This work employs CNN to build a gender classifier. Facial patches are extracted from preprocessed face images and subsequently fed into different CNNs to enable cooperative gender recognition. Individual steps, such as face preprocessing (Section 3.1), facial patch definition (Section 3.2), network training (Section 3.3), and classifier building (Section 3.4), are described in this section.

3.1 Face preprocessing

Our system begins by preprocessing face images in accordance with a previously described method[10]. As illustrated in Fig. 4, this process involves three steps. Given a face image without landmarks, we employ the Active Shape Model (ASM)[23,24] to locate the facial landmark points that lie on the boundary of the eyebrows, eyes, nose, and lips (Fig. 4b). We then rotate the image to align the eyes horizontally (Fig. 4c). Face patches are cropped on the basis of the landmark points with prior knowledge (Fig. 4d) as described in Section 3.2.

3.2 Facial patch definition

Given the aligned face image and landmarks, we divide the face area in accordance with the Three Chambers and Five Eyes principle: the distances between the forehead and the eyes, the eyes and the nasal floor, and the nasal floor and the jaw are the same (Three Chamber Heights), and the widths among the medial canthi, lateral canthus, and its neighboring sideburn are the same (Five Eye Widths). The Chamber Height and Eye Width are denoted as h and w, respectively. We then impose an axis in units of h and w on the face image for patch cropping. An illustration of images (*imgs*) and the large patches we explored, namely, Faces, Parts, and Patches, are presented Section 4.2.1. The details of patch extraction are explained as follows:

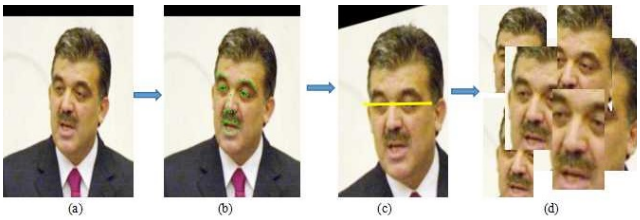


Fig. 4 Face preprocessing pipeline. (a) Input image; (b) result of face landmark detection; (c) image rotation for the horizontal alignment of eyes; and (d) cropping of various face patches for network training or testing.

(1) *Imgs* are uncropped images, i.e., the original image *img 1* and aligned image *img 2*. Most methods use the aligned image directly for gender classification (e.g., Refs. [10, 13, 16]). By contrast, in this work, we use the result of *img 2* as our baseline and demonstrate that gender classification accuracy can be further boosted through complementation among large patches in the classification process.

(2) *Faces*, referred to as **whole-face patches**, are patches that contain whole-face information cropped in accordance with facial landmark points from the aligned image (*img 2*). *Face 1* is a face-only patch (a cropped image of the Three Chambers and Five Eyes area) that merely contains facial features and lacks contextual information. Contextual information, e.g., facial contour, facial hair, and hair style, which may be helpful in gender recognition in some cases, is added gradually to *face 1* to yield patches *faces 2* to *10*. These nine patches are defined to retain the multiscale information of the whole-face image. Each patch is an observation of the whole-face from a different scale and contains different contextual combinations. For example, *face 10* has the information of bangs, *face 9* has the information of the entire hair style, and *face 8* has the information of the hair style and its surrounding area.

(3) *Parts*, referred to as **organ patches**, are areas cropped from the aligned image (*img 2*) designed to cover organ areas, i.e., the eyes and eyebrows, nose, and mouth, that may contribute to the human cognition of gender.

(4) *Patches*, referred to as **partial-face patches**, are patches cropped from *img 2* and contain partial-face information that may remain unchanged when occlusion or rotation occurs. Specifically, *patches 1* and *5* contain the upper face, which remain unchanged when the lower face is occluded by a mouth-muffle or hand; *patches 2* and *6* contain the lower face, which are uninfluenced by headpieces or a different hairstyles; *patches 3, 7, 9*, and *11* contain the right face, which

are less distorted than the left part of the face when the face is turning to the left; and *patches 4, 8, 10*, and *12* contain the left-face, which are less distorted when the face is turning to the right.

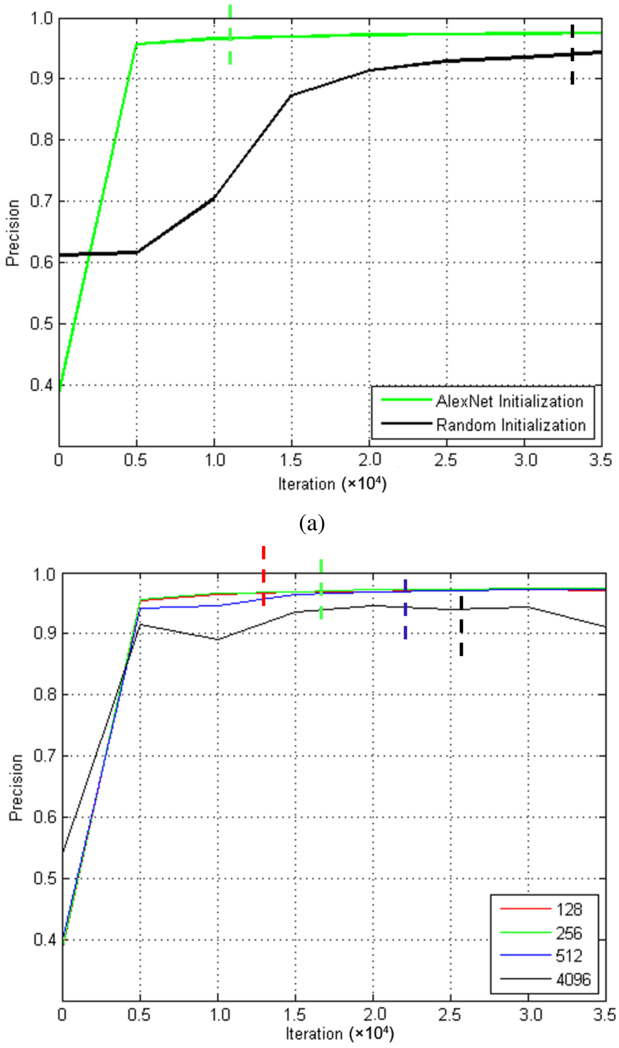
In summary, *img 2* is the aligned image from which all patches are drawn. The basic principles underlying the definition of large patches (*faces 1–10*, *parts 1–3*, and *patches 1–12*) are (1) the exploration of different combinations of contextual information (for *faces 1–10*); (2) human cognition (for the organ patches *parts 1–3*); and (3) the identification of face patches that are most likely to survive a variety of interference in real life (for *patches 1–12*). As described in Section 4.2.1, we test the independent and cooperative ability of our large patches on a validation set (a subset of the AAF database) and select the most powerful set of patches for use in our proposed method.

3.3 Network training

The proposed method is based on AlexNet[19], which achieves a winning top-5 test error rate on LSVRC2012[25]. AlexNet consists of five convolutional layers and two fully connected layers and exhibits satisfactory performance in a myriad of tasks[10, 19, 22].

Our proposed method leverages AlexNet[19] by maintaining its structure while modifying the last two layers to form our network. Specifically, the length of the last two fully connected layers (also referred to as the feature length) is changed from 4096 to 256. This modification corresponds to the experimental results presented in Section 4.2.3 showing that 256-D features outperform 512-D and 128-D features. This behavior demonstrates that the features of the proper-dim and not those of the highest-dim perform best for a specific task, e.g., gender classification. The precision vs. iteration (number of training iterations) curve for our validation set is shown in Fig. 5. The red, green, blue, and black curves in Fig. 5b represent feature lengths of 128, 256, 512, and 4096, respectively. Vertical dotted lines mark the convergence point of each curve and show that short lengths are indicative of fast convergence and that the highest accuracy is not achieved by the shortest or longest feature length. This image also shows that downsizing the last two layers reduces computational time cost and boosts classification accuracy.

The fine-tuning technique is used in network training. In fine-tuning, parameters in the first five layers of AlexNet are reserved and used to initialize our net during training. Such an initialization strategy saves time and considerably reduces the number of images needed in training because AlexNet can explore the main features of all objects after being trained on an extremely large image set (e.g., ImageNet). Fine-tuning from AlexNet parameters means that the parameters of the first five layers are already close to a converging stage. Thus, a small number of images can cause the new network to converge with considerably fewer iterations than starting from random parameters. The green and black curves in Fig. 5a



(a)

(b)

Fig. 5 Validation precision vs. iteration on AAF subset. The vertical dashed lines denote convergence points for each curve. (a) We compare the convergence between two starting models, i.e., pre-trained on ImageNet and random initialization. (b) We vary the dimension of the two Fully Connected (FC) layers in AlexNet as 128, 256, 512, and 4096. It can be observed that using pre-trained model converges faster and yields higher validation precision, and that FC layers with dimension of 256 perform better than others.

illustrate the performance of the network in training with AlexNet initialization and random initialization and demonstrates the effectiveness and necessity of effective initialization in network training.

In total, four same-structured nets are trained separately with different patches: whole faces, left faces, right faces, and lower faces. The learning rates of the output layer (2-dim) and the two fully connected layers are set 10 times of those of the five convolutional layers because they have fresh starts. As expected, all four networks converge quickly in practice (specifically, within 15 000 iterations).

3.4 Classifier building

When training is completed, we aggregate the four nets to construct the final gender classifier. In the general framework (Fig. 1), four nets work in parallel: patches extracted from the aligned image individually enter a corresponding net and are output with a label (intermediate result). In the back-end, these intermediate results are subjected to majority voting to obtain the final decision. Compared to a single network, running four nets costs four times memory and the same run-time. In general, the proposed algorithm can run at 0.3 frame per second with a single Titan X GPU with 12 GB memory. Given that nets in this framework are specifically designed for different parts of faces, the proposed method can overcome the local disturbance caused by pose variations and occlusion. As discussed in Section 4, we verify that our proposed method can effectively reduce the detection error caused by pose variations and occlusion. We will make our source code, models, and the AAF dataset available to the public.

4 Experiments

In this section, we present the experimental results of the proposed method on real-world face databases. We show that the classification accuracy of our method is comparable with that of state-of-the-art gender classification methods. Thus, our method has general application prospects.

4.1 Databases

We carry out our experiments on LFW, CelebA, ColorFeret, and AAF (Fig. 6) databases. The specific contents of these databases are described in Table 1.

4.1.1 AAF database

The AAF database provides a large set of the face images of individuals of different ages which were captured under unconstrained conditions. Some examples of the images in the AAF database are shown in Fig. 6. We are interested in age range because most existing face databases (e.g., LFW and ColorFeret) consist mainly of images of middle-aged individuals and lack fine age annotation; the age annotations for face images in existing databases are either absent or coarse (e.g., merely annotated as young, middle-aged, or old). The lack of large-scale unconstrained face images with fine age labels complicates the construction of age prediction methods or the demonstration of the robustness of gender classification methods over all ages.

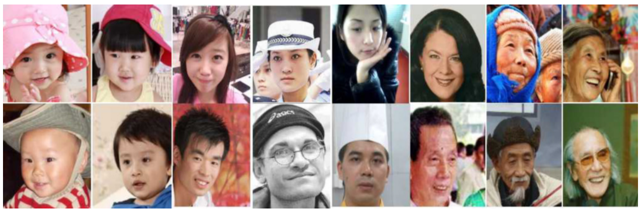


Fig. 6 An illustration of images in AAF database. Examples of images in AAF, with female ones in the first row, male ones in the second row. Images in AAF are collected from the web, and they are under very different conditions and cover all ages.

Table 1 Databases used in the experiments. *ImgNum* denotes the number of images in the database; *M/F Ratio* denotes the male-female ratio of the database; and *ImgNum (Train+Test)* denotes the number of images used in training and testing process in this work.

Database	ImgNum	M/F Ratio	ImgNum (Train + Test)
AAF	13 298	0.81	6649 + 6649
LFW[26]	13 233	1.85	11 910 + 1323
CelebA[10]	202 599	0.72	182 637 + 19 962
ColorFeret[5, 27]	2722	1.44	0 + 2722

Therefore, we established the AAF database, which shares the characteristic of wide variability (such as illumination, orientation, and expression) with in-the-wild databases and provides a broad distribution over all ages. We believe that AAF represents an important contribution to existing databases and is especially important for studying the problems of age prediction and cross-age gender recognition.

The AAF database has the following statistics and properties:

- The database contains 13 298 face images. Each image has one face located in the center of the image.
- Each image contains a different individual and is given a unique name that indicates the individual's age

and serial number (e.g., “00000A02” is the face image of a 2-year-old child, and “13297A80” is the face image of an 80-year-old individual).

• The images are available in 300 × 340 pixels JPEG format. A few images are in gray-scale, and others are in color.

• All images show facial regions that have been manually cropped from raw images downloaded from the Internet and annotated with two labels, i.e., age (2–80 years) and gender (male, female).

Database construction involves the following steps: (1) collection of raw images that contains the faces of all ages from the Internet; (2) manual cropping of face images from raw images; (3) labeling of face images by age and gender; (4) rescaling of labeled images into 300 × 340 pixels; and (5) renaming of labeled images according to a standardized format by concatenating serial number and age label (e.g., 00000A02, 00001A02, ..., 13297A80).

After the AAF database is established, we use it as a validation set to test the effectiveness of different classifiers and facial patches to construct our proposed method and to test the general applicability of our method, i.e., whether our proposed method can or cannot handle images of individuals of various ages. We fine-tune our network with half of the images in the database and test it on the remaining half. The classification result shows that our method works well when applied to the face images of individuals of different ages.

4.1.2 Existing databases

The LFW database[26] is an in-the-wild database of face images designed for the unconstrained face recognition[15] task. It contains more than 13 000 face images captured under various environmental conditions collected from the Web. This database can be used to represent real-world cases given its rich variation and diversity. The gender label of LFW is annotated preliminarily by official Application Programming Interface (API)[28] and artificially adjusted later on. We randomly select 90% images for training and use the other 10% for testing.

The CelebA database[10] is a large face database with 10 000 identities, each of which has approximately 20 images. It contains 202 599 images (118 162 female images and 84 427 male images). Experiments carried out on this database use the proposed classifier with network fine-tuned on the first 9000 identities. Testing is conducted with the remaining 1000 identities (training and testing sets are same as Ref. [10]).

The ColorFeret database[5,27], which we use to test the general applicability of our proposed method, consists of 11 338 gray and color images of 994 identities (403 females and 591 males). Each identity has at most 13 images taken in different poses. We apply the classifier trained on CelebA directly to a subset of ColorFeret (the same subset used in Refs. [29, 30]), and find out that our classifier maintains competitive performance on this different database without modification.

4.2 Algorithm construction

In this section, we detail how our classification algorithm is constructed. The selection of patches and rectification of false detection caused by side faces and occlusion are explained in Section 4.2.1. Section 4.2.2 provides the comparison of different classification methods on CNN features in gender recognition. Section 4.2.3 presents our justification for the selection of the current net structure. The experiments described in this section are carried out on the validation set (a subset of AAF), which consists of 4000 middle-aged persons (2000 for training and 2000 for testing), with a male-to-female ratio of 1.

4.2.1 Patch selection

We test the gender classification ability of different face patches and aggregations (Fig. 7). Linear SVM classifiers are trained separately for different patches using 4096-D CNN features from the last fully connected layer of AlexNet[19].

An illustration of each patch is given in Fig. 8: *Imgs 1* and *2* denote the original image and the aligned image; *faces 1–10* denote the different crops of whole-face patches; *parts 1, 2*, and *3* denote the organ patches of the human face (specifically, the eyes and eyebrow area, nose, and mouth); and *patches 1–12* denote partial-face patches. The patches are extracted in accordance with the position of feature points in the face image, where the face-only patch (*face 1*) is extracted in accordance with the aesthetic principle of the Three Chambers and Five Eyes and other patches of complete faces are expanded from *face 1* (see details in Section 3.2).

We compare the results of different aggregations and select the most effective aggregation (*agg 5*) to develop our proposed method. The results of some typical aggregations are shown in Fig. 7. The components of the aggregations are listed in Table 2: *agg 1* is the aggregation of an aligned image and an organ patch

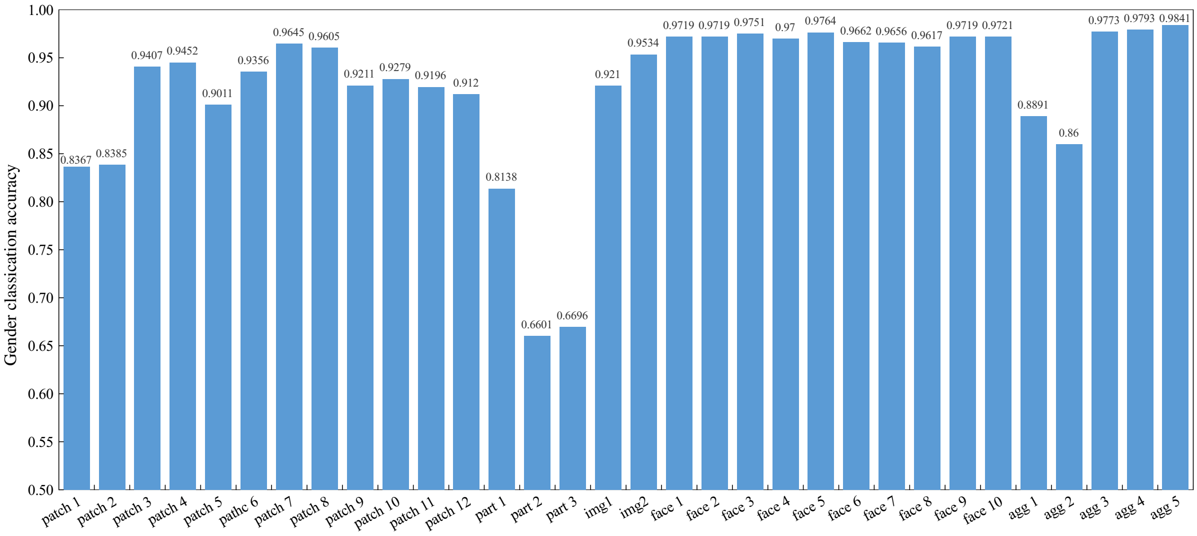


Fig. 7 Gender classification ability of different patches (Fig. 8) and aggregations (Table 2). The gender recognition ability of each patch or aggregation is represented by its classification accuracy on the AAF subset. We have observed the following: (1) Aligned image (*img 2*) performed 3.24% better than the original image (*img 1*). (2) Organ patches (*parts 1, 2,* and 3) lacked good ability for gender classification and harmed aggregation performance. For example, *img 2* had an accuracy of 95.34% without organ patch aggregation of organ patches. Its accuracy, however, decreased by 6.43% and 9.34% when aggregated with organ patches (*agg 1*, *agg 2*). (3) Whole-face patches (*faces 1–10*) performed best independently (e.g., *face 5* had the highest accuracy of 97.64%). Aggregation performance improved when the number of patches increased. Specifically, aggregation with a single whole-face patch was outperformed by that with whole-face patches (*agg 3*), and aggregation with whole-face patches (*agg 3*) was outperformed by that with whole-face and partial-face patches (*agg 5*). (4) Aggregation with aligned images and patches other than the upper ones (*agg 5*) had the highest performance of 98.41%, which is higher than that of *agg 4* that consisted of aligned images and all whole-face and partial-face patches. The accuracy of *agg 4* was 97.93%.

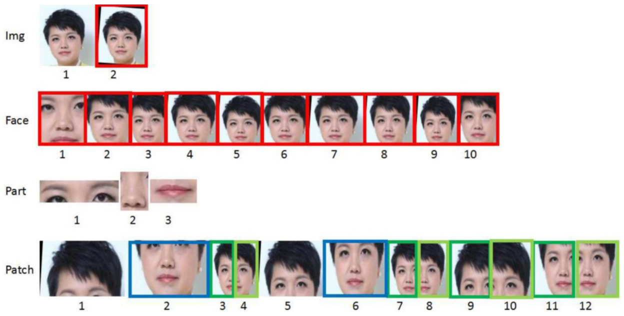


Fig. 8 Illustration of patches. The original image (*img 1*) was aligned (*img 2*) in accordance with facial landmarks. The organ patches for eyes, nose, and mouth (*parts 1–3* in row 3) were extracted directly on the basis of landmarks. Whole-face patches (*faces 1–10* in row 2) and partial-face patches (*patches 1–12* in row 4) were extracted from the aligned image (*img 2*) in accordance with the process described in Section 3.2. In the training process of our method, whole faces (red), right parts (shallow green), left parts (dark green), and lower parts (blue) were used to train nets 1, 2, 3, and 4, respectively.

Table 2 Components of different aggregations. *Agg 5* is the aggregation used in our proposed method.

Name	Component
Img with Part I (*agg 1*)	*img 2, part 1*
Img with Part II (*agg 2*)	*img 2, parts 1, 2,* and *3*
Whole face only (*agg 3*)	*img 2, faces 1–10*
All patches (*agg 4*)	*patches 1–12, img 2, faces 1–9*
All patches but upper parts (*agg 5*)	*patches 2–4, 6–12, img 2, faces 1–9*

(Part I), *agg 2* is the aggregation of an aligned image and all three organ patches (Part II), *agg 3* is the aggregation of an aligned image and all ten whole-face patches, *agg 4* is the aggregation of an aligned image and all whole-face and partial-face patches, and *agg 5* is the aggregation of an aligned image and all whole-face and partial-face patches except for partial-face patches for the upper part of the face (*patches 1* and *5*). As presented earlier, the classification results for *aggs 1–5* are obtained by majority voting of their component patches.

The classification results obtained by patches working independently or cooperatively (*aggs 1–5*) are shown in Fig. 7, from which we have the following observations:

(1) Alignment is important. The accuracy for aligned image *img 2* is 3.24% higher than that for the original image *img 1*.

(2) Independently, facial parts important for human

cognition (the organ patches *parts 1, 2,* and *3*) perform poorly on the gender classification task (the accuracies of organ patches *parts 1, 2,* and *3* are 81.38%, 66.01%, and 66.96%, respectively). By contrast, patches that contain additional information exhibit considerably improved performance (e.g., *patches 1–12, faces 1–10* all perform at least 10% higher than *parts 1–3*). Furthermore, the participation of organ patches in aggregation harms rather than boosts performance as shown by comparing *agg 1* and *agg 2* vs. *img 2*: *img 2* has an accuracy of 95.34% when it is not aggregated with organ patches (*parts 1–3*). Its accuracy, however, drops to 88.91% and 86% when aggregated with organ patches (*agg 1, agg 2*).

(3) Whole-face patches perform best independently (e.g., *face 5* has the highest accuracy of 97.64%). The improvement in the performance of aggregations with additional patches, however, may be attributed to the complementarity of different patches as demonstrated by the outperformance of a single whole-face patch by the aggregation of whole-face patches (*agg 3*), and the outperformance of the aggregation of whole-face patches (*agg 3*) by aggregation with whole-face and partial-face patches (*agg 5*).

(4) The aggregation of aligned images and patches other than the upper ones (*agg 5*) has the highest performance of 98.41%. The accuracy of *agg 4*, which consists of an aligned image and all whole-face and partial-face patches, is 97.93%. The patches of the upper parts are highly likely to exhibit distortions, such as hats, scarves, and hairpins. This characteristic may account for the superior performance of *agg 5*.

In this experiment, we demonstrate that patches help boost gender classification performance and our selection of the set of patches used in our proposed method. We choose the aggregation with the highest accuracy (*agg 5*). An example of the way large patches help rectify false detection is provided in Fig. 2: occlusion or a drastic perspective change exerts strong influence on a local area, thus distorting holistic information and causing the aligned image (*img 2*) yield an erroneous recognition result. However, the majority of large patches retain sufficient effective information because they are designed to ultimately produce correct classification results. This characteristic accounts for the better robustness and performance of the proposed patch aggregation than those of a single input.

4.2.2 Classifier selection

We compare the suitability of different classifiers for CNN features. We directly use AlexNet[19] to extract the 4096-D features of aligned images (*img 2* in Fig. 8) and train different classifiers, i.e., KNN, Decision Tree, SVM, and Softmax classifier (introduced in Section 2), to recognize gender.

(1) KNN: Given a test image, find its five nearest neighbors in the training set and vote labels of these neighbors to obtain the label of the test image;

(2) DT: Learn a decision tree from training images; then, the label of a test image is given by the decision of this learned tree;

(3) SVM: Train a linear SVM classifier on the training set to classify test images;

(4) Softmax: Train a Softmax classifier on the training set to classify test images.

The results of these classifiers are shown in Table 3, where we can see that the accuracies of KNN, DT, SVM, and Softmax are 75.32%, 92.28%, and 95.34%, and 96.05%, respectively. These results illustrate that the Softmax classifier is the most suitable classifier for CNN features in this task. Therefore, in our proposed structure, we set the loss function of CNN output layer as SOFTMAX LOSS and take the net predictions as labels for each patch.

4.2.3 Network structure selection

The highest-dim features are not the best feature for a specific task. We randomly select 80% images for training and 20% for testing (all images are automatically aligned) to identify the most appropriate feature length in this task. We modify the lengths of the last two fully connected layers of AlexNet (i.e.,

Table 3 Comparison of classifiers for the AAF subset.

Method	Accuracy (%)
KNN	75.32
DT	92.28
SVM	95.34
Softmax	**96.05**

Table 4 Gender classification with different feature lengths on the AAF subset.

Feature length	Accuracy (%)
4096	97.35
512	97.81
256	**98.04**
128	97.96

4096, 512, 256, and 128) and fine-tune each layer. We then compare the ability of different feature lengths by observing the classification accuracy of each net. As shown in Fig. 5 and Table 4, the feature length of 256 is the best for this task. Figure 5 and Table 4 show the precision-iteration curve and specific data of the training process of the nets, respectively. Given the feature length of 256, the network has a gender classification precision of 98.04%, which is higher than the precision obtained by features shorter or longer than 256 (e.g., 97.81% for the 512-D feature and 97.96% for the 128-D feature). Therefore, we choose the AlexNet structure with fully connected 256-D layers in the proposed method.

4.3 Gender classification

We randomly select 90% of images from the LFW database to train our nets. The four networks are trained with face patches of the whole, left, right, and lower face parts. Patches for LFW are generated in accordance with facial landmark points. In evaluation, we test the accuracy of these four nets separately, and then compare aggregation results. As shown in Table 5, the performance of our gender classification method is 98.55%, which is 2.11% higher than the performance of our baseline and 1.75% higher than that of the former state-of-the-art method[9].

CelebA is subjected to the same procedure, except that feature points are obtained through database annotation rather than automatically detected through ASM. We use the same training and testing set as that used by Ref. [10] and obtain the classification accuracy (Table 5) of 97.83%, which is 1.53% higher than the classification accuracy of the baseline (classification result of the aligned image) and is comparable with that of the state-of-the-art method (98% in Ref. [10]). As shown in Table 6, the classification accuracy of our proposed structure decreases in the absence of the four nets. This result demonstrates the necessity of each net in our classification structure. Figure 9 presents some images from the four databases that had been misclassified by our proposed method. Most of these images are severely occluded or are poor quality. These characteristics likely account for their incorrect recognition.

Table 5 Gender classification results. The comparison of gender classification accuracy on three databases (LFW, CelebA, and AAF) is shown in this table. Ours (nets 1–4) denote the net for whole, left parts, right parts, and lower parts of faces; “aligned image only” denotes the classification result by net1 for the aligned image (*img 2*), which represents the baseline of our method.

Method	LFW (%)	CelebA (%)	AAF (%)
Facial attrubutes[7]	79.52	–	–
Face tracker[29]	–	91.00	–
PANDA-l[30]	–	97.00	–
Deep face attributes[10]	94.00	**98.00**	–
LBP adaboost[8]	94.81	–	–
Compact face representation[9]	96.80	–	–
Aligned image only (baseline)	96.44	96.30	95.78
Ours (net1)	97.93	97.72	96.05
Ours (net2)	92.58	96.82	91.20
Ours (net3)	91.81	96.88	91.96
Ours (net4)	94.80	97.26	93.56
Ours (final)	**98.55**	97.83	**98.41**

4.4 General applicability

We test our network on the AAF database and use 50%

Table 6 Gender classification results of the proposed method with different aggregations. Comparison of gender classification accuracy on CelebA with different net aggregations is shown in this table: *nets 1–4* denote the net for whole, left parts, right parts, and lower parts of faces, respectively; *final* denotes our proposed method; *final-net1* denotes the aggregation without net1, *final-net2* denotes the aggregation without net2, and so on.

Method	Accuracy (%)
Final-net1	97.70
Final-net2	97.74
Final-net3	97.76
Final-net4	97.80
Final	**97.83**

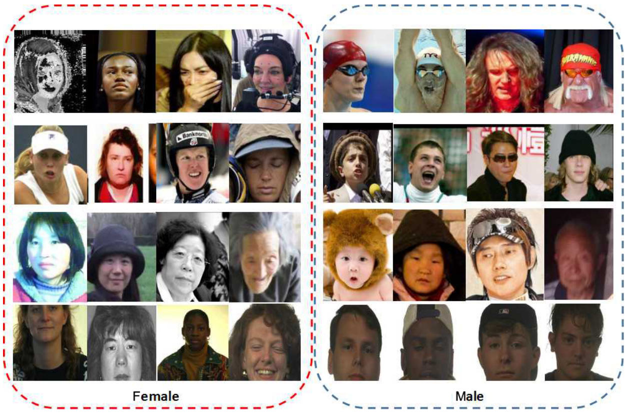


Fig. 9 Examples of misclassified images. Face images from the CelebA, LFW, AAF, and ColorFeret databases that are misclassified by the proposed method are shown in rows 1, 2, 3, and 4, respectively. Left: female images recognized as male; Right: male images recognized as female.

of images for training and the rest for testing to show that our method can be used for the gender classification of individuals of all ages. The classification accuracy of the proposed method is shown in Table 5. With fine-tuning of the network, our method can achieve a high gender classification accuracy of 98.41% (2.63% higher than baseline) on face images under unconstrained conditions and over a wide age range. This result shows that our method is robust and has strong adaptability.

We demonstrate the transferability of our method over different databases by directly applying the classifier trained on CelebA to the ColorFeret database without any modification. As shown in Table 7, the proposed method can perform well on a different database even without modification, and large patches continue to improve classification accuracy in transfer usage (the precision of the proposed method without patches is 91.37% and that of the proposed method with large patches is 93.38%). This shows that our method is generally applicable and has considerable practical value.

5 Conclusion

We proposed a multipatch gender classification method that is based on CNN. Experimental results showed that our proposed method outperforms existing single input methods by reducing error detection stemming from face rotation or occlusion. Our major contributions are threefold: (1) evaluating the ability of large patches for gender classification and identifying the effective aggregation of complementary patches; (2) proposing a gender classification method that is robust to occlusion and view change and that achieves the state-of-the-art gender classification accuracy of 98.55% on the LFW database (defeating the second-best by 1.75%) and competitive performance on the CelebA and ColorFeret databases; and (3) providing AAF, a new gender classification database comprising face images of individuals of all ages, and validating the robustness of our proposed method across ages.

Our proposed method relies on facial landmark points to extract large patches. Nevertheless, in this approach, gender classification accuracy is reduced by malposed landmark points. In the future, we will improve the performance of our proposed method with an approach that determines face landmark points with increased accuracy.

Table 7 Gender classification accuracy on ColorFeret.

Method	Experimental setting	Accuracy (%)
SVM + RBF[31]	80% traning + 20% testing	93.50
CoNN[32]	80% traning + 20% testing	**97.10**
Ours (single-image)	100% testing	91.37
Ours (proposed method)	100% testing	93.38

Acknowledgment

This work was supported by the National High-Tech Research and Development (863) Program of China (No. 2012AA011004), the National Science and Technology Support Program (No. 2013BAK02B04), and the National Key Research and Development Plan (No. 2016YFB0801301).

References

[1] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, SEXNET: A neural network identifies sex from human faces, in *Proc. 1990 Conf. on Advances in Neural Information Processing Systems*, Denver, CO, USA, 1990.

[2] D. Valentin, H. Abdi, B. Edelman, and A. J. O'Toole, Principal component and neural network analyses of face images: What can be generalized in gender classification? *J. Mathem. Psychol.*, vol. 41, no. 4, pp. 398–413, 1997.

[3] M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu, Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis, in *Proc. 4^{th} IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 202–207.

[4] S. Buchala, N. Davey, T. M. Gale, and R. J. Frank, Principal component analysis of gender, ethnicity, age, and identity of face images, in *P*roc. *IEEE ICMI*, Tozeur, Tunisia, 2005.

[5] P. J. Phillips, H. Wechsler, J. Huang, and R. J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[6] T. Sim, S. Baker, and M. Bsat, The CMU Pose, Illumination, and Expression (PIE) database, in *Proc. 5^{th} IEEE Int. Conf. on Automatic Face Gesture Recognition*, Washington, DC, USA, 2002, pp. 53–58.

[7] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, Robust gender recognition by exploiting facial attributes dependencies, *Pattern Recognit. Lett.*, vol. 36, pp. 228–234, 2014.

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

[9] H. Q. Fan, M. Yang, Z. M. Cao, Y. J. Jiang, and Q. Yin, Learning compact face representation: Packing a face into an int32, in *Proc. 22^{nd} ACM Int. Conf. on Multimedia*, Orlando, FL, USA, 2014, pp. 933–936.

[10] Z. W. Liu, P. Luo, X. G. Wang, and X. O. Tang, Deep learning face attributes in the wild, in *Proc. 2015 IEEE*

Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 3730–3738.

[11] P. Rai and P. Khanna, Gender classification using radon and wavelet transforms, in *Proc. 2010 5th Int. Conf. on Industrial and Information Systems*, Mangalore, India, 2010, pp. 448–451.

[12] W. Reichl and W. Chou, A unified approach of incorporating general features in decision tree based acoustic modeling, in *Proc. 1999 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, USA, 1999, pp. 573–576.

[13] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, Revisiting linear discriminant techniques in gender recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 858–864, 2011.

[14] H. Zhang and Q. Zhu, Gender classification in face images based on stacked-autoencoders method, in *Proc. 2014 7th Int. Congress on Image and Signal Processing*, Dalian, China, 2014, pp. 486–491.

[15] Y. Sun, Y. H. Chen, X. Q. Wang, and X. O. Tang, Deep learning face representation by joint identification-verification, in *Proc. 27th Int. Conf. on Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 1988–1996.

[16] P. Rai and P. Khanna, A gender classification system robust to occlusion using Gabor features based (2D)^{2}PCA, *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1118–1129, 2014.

[17] G. D. Guo, C. R. Dyer, Y. Fu, and T. S. Huang, Is gender recognition affected by age? in *Proc. 2009 IEEE 12th Int. Conf. on Computer Vision Workshops*, Kyoto, Japan, 2009, pp. 2032–2039.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proc. 25th Int. Conf. on Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[20] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1–9.

[21] L. Zheng, Y. Yang, and Q. Tian, SIFT meets CNN: A decade survey of instance retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, 2018.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587.

[23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, Active shape models-their training and application, *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1995.

[24] S. Meller, E. Nkenke, and W. A. Kalender, Statistical face models for the prediction of soft-tissue deformations after orthognathic osteotomies, in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2005*, J. S. Duncan and G. Gerig, eds. Springer, 2005, pp. 443–450.

[25] J. Deng, Imagenet large scale visual recognition, PhD dissertation, Princeton University, Princeton, NJ, USA, 2012.

[26] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, http: // www.tamaraberg.com/papers/Huang_eccv2008-lfw.pdf, 2007.

[27] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, The FERET evaluation methodology for face-recognition algorithms, *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 22, no. 10, pp. 1090–1104, 2000.

[28] M. Thulin and P. Masek, Software quality evaluation of face recognition APIS and libraries, https://gupea.ub.gu.se/bitstream/2077/38856/1/gupea_2077_38856_1.pdf, 2015.

[29] N. Kumar, P. Belhumeur, and S. Nayar, Facetracer: A search engine for large collections of images with faces, in *Computer Vision-ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, eds. Springer, 2008, pp. 340–353.

[30] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1637–1644.

[31] S. Baluja and H. A. Rowley, Boosting sex identification performance, *Int. J. Comput. Vis.*, vol. 71, no. 1, pp. 111–119, 2007.

[32] F. Hing, C. Tivive, and A. Bouzerdoum, A shunting inhibitory convolutional neural network for gender classification, in *Proc. 18th Int. Conf. on Pattern Recognition*, Hong Kong, China, 2006, pp. 421–424.

Yali Li received the BEng degree from Nanjing University, China, in 2007 and the PhD degree from Tsinghua University, Beijing, China, in 2013. Currently, she is a research assistant in Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, video analysis, etc.

Jingchun Cheng received the BEng degree from Tsinghua University, China, in 2014. She is a PhD candidate in Tsinghua University. Her research interests include image classification and object detection.

Shengjin Wang received the BEng degree from Tsinghua University and the PhD degree from Tokyo Institute of Technology, in 1985 and 1997, respectively. From 1997 to 2003, he was a member of the research staff with the Internet System Research Laboratories, NEC Corporation, Japan. Since 2003, he has been a professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the director of the Research Institute of Image and Graphics. His current research interests include image processing, computer vision, video surveillance, and pattern recognition.

Le Yu received the PhD degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2014. She is a research fellow with the China Mobile Information Security Center. Her research interests include multimedia and big data security, cloud computing, etc.

Jilong Wang received the PhD degree from Tsinghua University in 2000. Since 2010, he has been a full professor with Institute of Network Science and Cyberspace, Tsinghua University, China. He also served as an NOC director of several large scale Internet infrastructures including CERNET2, TEIN, and Tsinghua campus network since 2004. His research focuses on network measurement and security. He is the founder of DragonLab federation testbed, and Internet Innovation Union which is based on DragonLab.