Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 2084 – 2100
© TÜBİTAK
doi:10.3906/elk-2008-166

Bagging ensemble for deep learning based gender recognition using test-time augmentation on large-scale datasets

Taner DANIŞMAN[*]
Department of Computer Engineering, Faculty of Engineering, Akdeniz University, Antalya, Turkey

Received: 31.08.2020 • **Accepted/Published Online:** 05.04.2021 • **Final Version:** 26.07.2021

Abstract: We present a bagging ensemble of convolutional networks in combination with the test-time augmentation technique to improve performance on the cross-dataset gender recognition problem. The bagging ensemble combines the predictions from multiple homogeneous models into the ensemble prediction. Augmentation techniques are often used in the learning phase of the CNNs to improve the generalization ability. On the other hand, test-time augmentation is not a common method used in the testing phase of the learned model. We conducted experiments on models trained using different hyperparameters. We augmented the test data and combine the predictive outputs from these network models. Experiments performed on diverse gender datasets, including Adience, AFAD, CelebA, Gallagher, Genki-4K, IMDb, LFW, Morph, VGGFace2, and Wiki, showed that the use of bagging ensemble of convolutional networks and test-time augmentation outperforms standalone models. We obtained the highest cross-dataset accuracy in the literature on seven out of eleven datasets. For the remaining four datasets we reported the cross-dataset results for the first time. According to our experiments, VGGFace2, IMDb, and CelebA datasets provided the highest cross-dataset classification results for most of the test datasets in the gender recognition problem.

Key words: Cross-dataset gender recognition, bagging methods, deep learning, test-time augmentation

1. Introduction

Gender recognition is a two-class classification problem to classify given visual data as male or female classes. Visual data can be a portrait face, 3D volume, gait, body, or even just hands [1] or ears [2]. Face based gender recognition under unconstrained settings is getting more attention in recent years [3–5]. It contributes to the other vision problems and research fields such as biometrics, face recognition, age prediction, targeted advertising, recommendation systems, and human-computer interaction. Demographic studies showed that face-based gender recognition in the wild is a challenging classification problem due to variations in appearances such as age, head position, ethnicity, lighting, and facial expressions [6–8]. To overcome challenging situations, there is a need for large-scale training sets with diverse identities, well representing the problem space by proper feature selection method, using an optimized machine learning algorithm, and use of ensemble learning methods. Among these, the two most important factors affecting classification accuracy are the selections of an error-free dataset and the machine learning method.

Dataset selection plays an essential role in classification problems. Part of the datasets used in the gender recognition domain, derived from face recognition research. These datasets are not well suited for the gender recognition problems due to limited diversity and controlled scenes. The most obvious example for a controlled

[*] Correspondence: tdanisman@akdeniz.edu.tr

dataset is the Feret dataset [9], where the scene is completely controlled. The use of small-scale and scene-controlled datasets eliminates many challenging situations and does not reflect the success of the classification methods. According to [10], overoptimistic performances have been achieved for controlled datasets.

Having a large, error-free, and a diverse dataset is a must in supervised learning. Large-scale datasets and visual variations provide different representations of the classes. For example, they enrich the soft-biometric features, such as ethnicity, age, and pose distribution, within the classes. These datasets may reveal the actual performance of the classification methods. The variations in the training data improve the generalization ability of the classifiers. To increase the variations in the training samples, it is necessary to use augmentation techniques (random rotations, translations, random brightness changes, horizontal flipping, etc.). Thus, it covers the solution domain much more than a scene controlled dataset. As indicated by [11], ”There is a need to assemble datasets with a larger number of images than what is being used now”. However, this is not a trivial task due to the nature of the large-scale datasets. Automated systems and crawlers collect images from the web to construct large-scale datasets. Any wrong class labels also affect the overall performance of the classifiers. The fact that the ground truth labeling is correct is of great importance for supervised learning. An incorrect class label, when used in conjunction with augmentation techniques, reduces the success of the methods. For this reason, the training data must be as clean as possible.

A clean dataset, in combination with a powerful machine-learning method, is a key point for successful classification. Deep learning methods and their derivatives are successful examples that can be used for the gender recognition problem. These methods start solving the problem from random locations by using random weights (if transfer learning is not applied) and iteratively approach the solution domain using an optimization algorithm. Using random weights enables different training models to be obtained, where training models can also be called hypotheses. In supervised learning, there are many hypotheses existing in the solution space which can perform acceptable prediction, thus the learned model is subject to variability. However, finding the optimal hypothesis is a difficult optimization problem. To overcome part of this problem, ensemble methods can be used.

Ensemble classifiers offer more flexibility to represent the solution domain by combining multiple hypotheses. Therefore, they give better predictive performance than individual use of standalone classifiers. Ensembles can be implemented through bagging, boosting, and stacking methods. Bagging (bootstrap aggregating), produces an ensemble model that is more robust than the individual standalone models composing it. In bagging, the learning process is independent of all other learning processes thus provides more flexibility (e.g., parallel learning), while boosting learns in a sequential and dependent way. Bagging and boosting methods use homogeneous classifiers, while the stacking method use heterogeneous classifiers. Table 1 shows the main features of ensemble methods.

Table 1. Ensemble methods and their features.

Method	Dependency	Learning	Weight	Classification	Expected result
Bagging	Independent	Parallel	Equal	Deterministic averaging	Lower variance
Boosting	Dependent	Sequential	Model’s performance	Deterministic	Lower bias
Stacking	Independent	Parallel	Learning	Meta-classifier	Lower bias

A proper evaluation protocol is an important factor for evaluating the performance of different methods. When the size of the training set is small, researchers tend to use n-fold cross-validation based evaluation within the same dataset to measure the predictive performance of their method. In n-fold cross-validation, the data is randomly sampled into n equal size folds. $n-1$ folds are used for training and one fold is left for validation, where each fold is used once as validation data. Averaging the results of n validation tests provides a better estimation of the predictive performance of the method [12]. However, its joint application with oversampling on imbalanced datasets results in biased and overly-optimistic estimates [13, 14]. A better solution is to use cross-dataset evaluations to present the generalizability of the methods.

The main purpose of the study is to perform bagging ensembles of convolutional networks, in combination with test-time augmentation (TTA), for gender recognition on large-scale datasets in a cross-dataset evaluation scenario. The main advantages of the cross-dataset evaluation are that it provides results without any bias from soft-biometric features of the dataset and shows the generalization ability of the models. As indicated in [15], the predictive performance of the cross-dataset models deteriorated significantly compared to tests within the database. In our approach, we benefit from the advantages of both the bagging method and test-time augmentation to overcome challenges in cross-dataset tests. The main contribution of this study is three-fold:

- We provide state-of-the-art cross-dataset results for large-scale gender datasets using ensembles of deep networks with TTA. Using our method, we obtained the highest gender recognition accuracy reported on Genki-4K, IMDb, LFW, Morph, and Wiki datasets.

- We manually corrected dataset annotations for the large-scale gender datasets. This is an important step towards better evaluations of the classification methods.

- To the best of our knowledge, this is the first cross-dataset gender recognition experiment performed on manually corrected large-scale datasets in the literature. We performed 110 cross-dataset tests.

We present the related works in both gender recognition and ensemble methods in Section 2. Materials and methods, including the datasets, normalization, augmentation, and ensembles of the deep network models, are in Section 3. Experimental results on cross-dataset tests and discussion are provided in Section 4. Finally, Section 5 concludes the study.

2. Related works

In this section, we discuss the standalone models and bagging based models in the literature for gender recognition. Much of the existing research on gender recognition is based on standalone base classifiers. There are few studies in the literature using bagging based ensembles for gender recognition. Earlier methods use SVM [16–18], Neural networks [19] and AdaBoost [20, 21] algorithms for gender recognition based on pixels and handcrafted features such as PCA, LBP and Gabor features. Handcrafted feature extraction methods are applied on the localized face to reduce dimensionality and highlight important information in image data. Since the feature extraction methods used in visual gender recognition problems show similarities with methods used in other visual classification problems (e.g., face recognition, object classification), they are also used in the gender recognition problem.

Feature extraction methods are divided into two groups, geometry-based and appearance-based methods. With the widespread use of convolution-based deep learning methods in the field of visual classification, where the feature extraction is automatically realized by the convolutions, the use of handcrafted feature extraction methods has decreased. Recent studies in gender recognition focus more on Deep Convolutional Networks that made successful progress on visual classification tasks [3, 22–25].

Earlier studies showed that LBP and Gabor jets offer better performance than raw pixels when used with SVM and variants. For this reason, these two features are extensively used in the literature [15, 26–31]. In [28], appearance-based raw pixels and feature-based (Gabor and LBPs) descriptors are used with linear SVMs and linear discriminant analysis (LDA). VJ based face detector and an eye detector are used to detect the face and eyes. They considered both internal and external facial features by selecting different face crops. They used face crop size of 105 × 90 and 120 × 105 and performed their experiments using pixels, Gabor jets, and LBPs on Linear SVM and LDA with PCA. They obtained higher results when they use Gabor Jets and LBP on the LFW dataset. According to their experiments, PCA+SVM and PCA+LDA schemes have similar performances. In [15], the authors also studied the effect of different cropping factors on the gender recognition problem. For training, they use LBP and LBPHS features on SVM with RBF kernel. They performed cross-dataset experiments on original and gender-balanced Feret, LFW, and Morph datasets. Similar to [28], authors in [26] used LBP features obtained from different face crops on linear SVM classifier. They performed cross-dataset experiments. In [31], low-level fusion of intensity, shape, and texture features are used in combination with minimum redundancy and maximal relevance (mRMR) feature selection for gender recognition on LFW, Feret, and Und datasets. Eidinger et al. [29] studied gender recognition on Gallagher's and Adience datasets. They experiment with dropout learning techniques and Linear Support Vector Machines. They process the images using a robust face alignment technique, then use LBP and four-patch LBP (FPLBP) as a feature representation model.

The bagging method for gender recognition is studied in [27, 30]. In [27], multiple SVM based linear models are used to create bagging of classifiers and stacking for gender recognition. They use histograms obtained from LBP and HOG descriptors. In [30], the weighted bagging method is used in combination with LBP histograms and Gabor wavelets. Using a dynamic weighting method they compared the majority voting with the weighted bagging model. They found that the weighted bagging model provides higher accuracy for both female and male classes. They also collect their dataset having 28,235 images from FaceBook.

More recent studies covering cross-dataset experiments focus on convolutional neural networks [4, 32, 33]. In these studies, convolutions are used for feature extraction and SoftMax is used for the prediction. A commercial age, gender, and emotion recognition system is developed in [33]. They created several deep convolutional networks trained on their large dataset having over 4 million images over 40,000 identities. For gender recognition, they considered different ethnicity and age groups. They use augmentation techniques (e.g., horizontal flip and random crop) for training and for prediction of age. However, they did not use TTA for gender recognition.

A DCNN architecture based on MobileNet [34] for gender recognition is proposed in [4], at a reduced cost. They experiment with nontrainable parameters of the MobileNet architecture including input resolutions, width multipliers, and the number of layers. They experiment with the influence of the changes in the network architecture on the performance of the model. For the training, they do not align the images, but used augmentation techniques instead. Their cross-dataset experiments on VGGFace2, LFW, MIVIA-Gender, Wiki, and Adience datasets are competitive with our results. However, their results on IMDb are very poor because they did not clean the dataset.

Gender is also considered as a face attribute in face attribute estimation problems. In [32], a heterogeneous face attribute estimation method based on deep multitask learning is proposed for the recognition of numerous face attributes. According to their cross-dataset experiments, they obtained 77.40% and 89.00% accuracy for

Morph-LFW and LFW-Morph tests.

3. Materials and methods

In the case of supervised learning, complex problems need both large-scale training data and well-established network architecture with optimized hyperparameters. We first collected publicly-available datasets in gender recognition field. Before using them, we preprocessed the raw information and filtered it for annotation errors to avoid multiple augmentations of the incorrect sample during training epochs in supervised learning. Then, we experimentally adjusted the hyper-parameters of the problem including learning rate, learning decay, number of layers, number of fully connected neurons, convolution kernel size, and maximum number of iterations. Besides these, we applied augmentation policies to improve the generalization ability of the classifier and to prevent overfitting. To reduce the variance, we applied the bagging method on the SoftMax layer of the homogeneous base classifiers. Bagging of classifiers is an improved method compared to the basic majority voting approach, in which the output of the SoftMax layer is considered binary. In bagging, numerical values are added together to produce the final ensemble network output. Averaging of the base classifier predictions is useful when there is no correlation between them. Otherwise, the final prediction is the same as the base classifier's prediction. Therefore, we need to provide a way to increase the correlation among different base classifier outputs for the same input $D = \left\{(x_i, y_i)\right\}$.

We used $A = 3$ base classifier to create the ensemble model. Given a set of observations $\chi = \left\{x_i \in R^M\right\}$ and a set of true labels $Y = \left\{y_i \in N\right\}$ and a training set $D = \left\{(x_i, y_i)\right\}$ as an input, our aim is to learn a model M based on D using supervised learning. For testing, each test sample is augmented T times for each model M. In our experiments, both the number of models A and T value is empirically selected as 3.

Theoretically, we expect more accuracy with an increasing number of layers. However, having too many parameters with the increasing training epochs generally results in memorizing the input. Therefore, we implement a simple 6-layer convolutional network having 382,626 trainable parameters and 896 nontrainable parameters (used in 5 batch normalization layer) for the gender recognition problem. We select the dropout value between 0.05 and 0.15 for each base model to randomly drop part of the connection from the network. As a result, the model becomes robust and insensitive to the weights of the other nodes, thus it can generate a more generalized model. The first dropout layer was applied after the fourth batch normalization layer. The other is applied before the dense layer. For each base classifier, the learning architecture is finalized by the SoftMax layer that computes the likelihood that the input image belongs to a particular class. Figure 1 shows our deep learning architecture for creating the base models used in the ensemble experiments.

The input data is a normalized face image with a resolution of 64×64 in RGB color space, obtained from the VJ face detector [35]. Then using eye detection, we estimate the roll angle θ, from the eye locations. We rotate the face with respect to the center of the face and then scaled it down to 64×64 pixels to reduce computational complexity.

We used augmentation techniques to increase the diversity of the training samples in order to improve the generalization ability of base classifiers. Furthermore, we randomly rotate every sample up to ± 3 degrees, since our eye detection method used for in-plane roll normalization has ± 2 degrees MAE in detecting the roll angle θ. By rotating the face more than the MAE error in θ, our network will learn to handle the possible eye detection errors as well. As a result, we do not need precise eye detection. We also performed random zooming range, width and height shift range, random horizontal flip, and brightness changes. The same random effects are used in the TTA step as well.

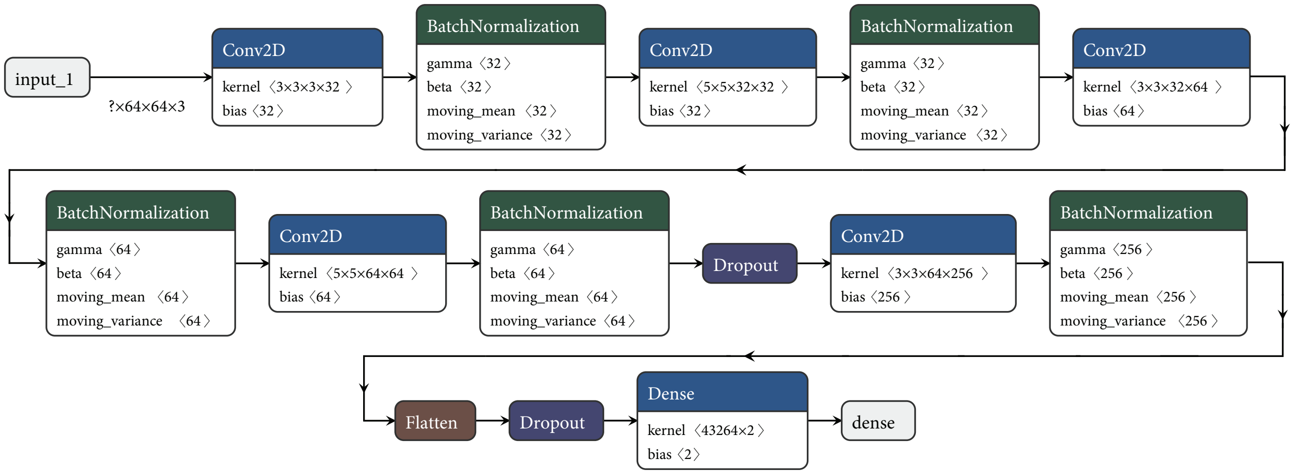


Figure 1. Deep learning architecture for each base model where the dropout values are dynamically updated.

Due to the limited capacity of the GPU memory, we used the mini-batch method to feed our deep network. We applied these policies to each mini-batch during one epoch that will provide a randomly augmented subset of original mini-batch to each epoch. Thus, it helps to prevent overfitting problems during training. The network weights are updated after each mini-batch. Since the image samples in the mini-batch determine the weight updates, they should locally represent the main classification problem. Thus, we may expect an equal distribution of classes in each mini-batch, otherwise, there will be a bias towards a class.

Gender datasets usually have more than one sample image per identity. For example, there are 530 sample images for George W. Bush in the LFW dataset. A mini-batch size of 64 can consist of 64 George W. Bush photos if shuffling is not used. As a result, the model learned by the network will be semantically person identification other than gender recognition. A solution to this problem is to use shuffling so that mini-batches contain representative samples from each class. In our experiments, 32 samples from the female class and 32 samples from the male class are used in a mini-batch size of 64.

3.1. Datasets

We performed our experiments on a wide variety of publicly available gender datasets, including Adience, AFAD, CelebA, Gallagher, Genki-4K, IMDb, LFW, Academic Morph, UTKFace, VGGFace2, and Wiki. Table 2 shows details of the datasets used in the experiments. These datasets mostly have a single face per image. While human annotators manually determined gender information in some datasets, in others it was determined by using semiautomatic methods like attribute classifiers. As indicated in [36], a clean dataset without labeling error is a required step towards higher performance. When used with training time augmentation, noisy labels also augmented and decreased the classification accuracy. For this reason, we manually checked and annotated the gender labels before the supervised learning process. We corrected the wrong labels and removed nonface photos. To do that we used available metadata (where extant) provided by the datasets. Otherwise, we manually annotated the dataset. Manual annotations can be accessed through the GitLab link[1].

Adience dataset is an age and gender dataset collected from Flickr albums. Gender is almost balanced in this dataset. The Asian face age dataset (AFAD) proposed for age estimation. It contains cropped face photos

[1] Dataset annotations https://gitlab.com/danisman.taner/manual-annotations-for-common-gender-recognition-datasets.

Table 2. Details of the datasets before and after the label corrections.

Dataset	Samples	Identities	Original		Annotated		Preprocessed	
			Female	Male	Female	Male	Female	Male
Adience [29]	26,580	2284	10,154	9216	10,346	9013	7653	6258
AFAD [39]	164,432	N/A	63,680	100,752	64,016	101,313	26,319	64,269
CelebA [40]	202,599	10,177	118,165	202,599	118,923	83,676	97,620	65,889
Gallagher [41]	5080	N/A	14,559	13,672	0	0	11,015	9859
Genki-4K[2]	4000	4000	N/A	N/A	1966	2024	1,459	1509
IMDb [42]	178,600	20,283	80,660	97,940	79,813	97,667	65,922	84,082
LFW [43]	13,233	5748	2966	10,268	2950	10,255	2,542	8605
Morph [44]	55,134	13,618	8489	46,645	8249	46,885	8,064	45,479
UTKFace [45]	24,104	N/A	11,523	12,581	11,572	12,531	8,109	9082
VGGFace2 [46]	3,156,872	9131	1,299,574	1,842,316	1,306,612	1,834,960	762,804	927,044
Wiki [42]	62,328	N/A	10,262	31,912	10,246	31,926	6,650	17,589

[2] Genki-4K (2009) [online]. Website http://mplab.ucsd.edu, [accessed 13 April 2020].

obtained from 'selfie' images on a social network. We manually annotated this dataset against labeling errors. Large-scale CelebFaces Attributes dataset (CelebA) contains celebrity images covering large pose variations and background clutter. Gallagher dataset involves wide range of illumination, ethnicity, ages, in-plane and out-of-plane poses. Genki-4K dataset contains 4,000 images with expression and head-pose labels. IMDb dataset contains faces of the most popular 100,000 actors listed on the IMDb website. This is an automatically crawled dataset. It assumes that the single face images on the actor's IMDb web page are likely to belong to the actor. However, this assumption is not correct for all images. Besides, there is a need to eliminate images that do not contain any face by filtering the $face\_score$ parameter contained in the dataset. The Labeled faces in the wild dataset (LFW) is a partially labeled face dataset having photographs of individuals collected from the web, mainly actors, politicians, and athletes. The Academic Morph dataset has mugshot images of 13,618 identities. It provides both age and gender features. UTKFace dataset contains images with a long age span. Images are labeled by age, gender, and ethnicity. The VGGFace2 dataset is the extended version of the VGGFace dataset that contains more than 3 million images. More than 14,000 gender labels were corrected on this dataset. The Wiki dataset and IMDb dataset share the same meta information. It contains profile images from pages of the people from Wikipedia with the same identity as the IMDb dataset.

Large scale datasets are usually created by crawling the web. It is difficult to restrict identities existing in both training and test datasets. Thus, we used these datasets as-is. However, IMDb, LFW and VGGFace2 datasets provide identity information. For example, 5.74% of the identities in VGGFace2 are the same as the identities in LFW. We selected an equal number of training samples from female and male classes to reduce bias.

3.2. Environment

We performed our experiments using Tensorflow and Keras deep learning framework on an Ubuntu OS (20.04) with CUDA 10.1 (Nvidia GTX 1060 6GB RAM). OpenCV implementation of Viola–Jones face detector [35] and neural network-based eye detector [37] available in STASM library [38] is used in the normalization step.

3.3. Evaluation metrics

The results obtained from different studies must be carefully compared, since the implementation details and evaluation protocols are all different in these studies, even for the same training sets. We used BeFIT evaluation metrics to measure the performance of the proposed method. Since we care about true negatives as much as true positives, we considered ROC area under curve (ROC AUC) to determine the predictive power of the ensemble model. To compute the accuracy, we consider the correctly classified images divided by the total number of images as shown in Eq. (1).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

To compare different experiments, we used macro accuracy, also known as mean unweighted accuracy (UA) metric, as shown in Eq. (2). This metric is independent of the size of the test sets. It gives equal weight to each result.

$$UA = \frac{1}{N} \sum_{i=1}^{N} Acc_i \tag{2}$$

In Eq. (2), N represents the total number of cross-dataset tests and Acc_i represents the accuracy obtained from a train-test couple using Eq. (1).

4. Experiments

We performed four different experiments to see the effect of bagging and TTA. The first two experiments focus on the effect of bagging without TTA NoBAG_NoTTA and YesBAG_NoTTA. Similarly, the last two experiments focus on the effect of bagging with TTA NoBAG_YesTTA and YesBAG_YesTTA. Detailed results are presented in subsections 4.2 and 4.3.

4.1. Experimental setup

In all experiments, we followed the cross-dataset evaluation protocol, where we select a dataset as a training set and another dataset as a test set. 5% of the datasets were used for the validation. For each mini-batch, we provide an equal number of samples for female and male classes. Models with the highest validation accuracy during 20 epochs were stored for each data set. Since we used three models in our experiments, we generate three different models from the same network architecture with different dropout values.

The choice of the initial learning rate and the changes in its lifespan is one of the most crucial decisions for neural networks. The learning rate is determined by the $\lambda = 1e - 3 * 0.95^x$ where x represents the epoch number. For the 20 epochs, it operates in the range $9.5 * 10^{-4} - 3.58 * 10^{-4}$. The dropout value is set to 0.05, 0.10, and 0.15 for the three homogeneous models. Table 3 shows details of the training and augmentation parameters.

For each dataset, the input image size is set to be 64×64 pixels and the maximum epoch is set to 20 epochs. 64×64 input image size was empirically selected as a result of initial experiments on the Genki-4K dataset. We experimented 16×16, 32×32, 64×64, and 128×128 input resolutions. The main objective here is to select a dimension to get the maximum benefit from the convolutions. Among others, 64×64 and 128×128 provides the highest accuracy with the proposed network architecture. Due to computational constraints and negligible performance differences, we selected the input size of 64×64.

Table 3. Training and augmentation parameters.

Validation split	Input size	Batch size	Epochs	# of models	Learning rate	Optimizer
5%	64×64	64	20	3	$9.5 * 10^{-4} - 3.58 * 10^{-4}$	Adam
Rotation range	Zoom range	Width shift range	Height shift range	Horizontal flip	Brightness range	Drop factor
3	0.06	0.06	0.06	True	[0.8, 1.1]	[0.05, 0.10, 0.15]

4.2. Bagging experiments without TTA

Table 4 shows the baseline experiments NoBAG_NoTTA and YesBAG_NoTTA without TTA. We obtained a mean UA of 90.20% and 91.45% respectively. When TTA is not used, for all cases, bagging provides better results than baseline scores. In this case, the use of bagging increased the mean UA for each of the test sets.

4.3. Bagging experiments with TTA

Table 5 shows the results of bagging experiments (NoBAG_YesTTA) and (YesBAG_YesTTA) using TTA. We obtained a mean UA of 89.84% and 91.58%, respectively. In NoBAG_YesTTA experiment we obtained the worst mean UA (89.84%) score. According to the results of the four experiments, standalone use of TTA did not increase performance. On the other hand, when it is used with the bagging method, it further improves the performance by a small amount. This may be because our TTA pipeline makes minor changes to the input images.

According to Tables 4 and 5, VGGFace2, CelebA, and IMDb datasets as training sets provide the best cross-dataset accuracy for the gender recognition problem. These datasets are the top three datasets in terms of dataset size and also contain the highest number of different identities. Besides, the UTKFACE data set used as the training data set is in the first five datasets according to the average accuracy value obtained in test datasets. Unfortunately, the number of identities in this dataset is not known. However, according to the experimental results, we can say that it contains enough identities for gender recognition. The use of content-rich (multiple identities, age variances, etc.) datasets in the training will generate a model that can be successful with similar contents. However, as seen in the results of the Morph dataset in Table 4 and 5, having a dataset with a larger number of different identities does not always result in higher accuracy. The Morph dataset has more identities than CelebA and VGGFace2 datasets (13,618 vs 10,177 and 9,131). Due to the ethnic characteristics of the Morph data set, lower accuracy rates were obtained in test datasets that do not have similar characteristics. In general, datasets that have similar characteristics are expected to show similar test performance.

We obtained the lowest accuracy values on Adience, Gallagher, and UTKFace datasets. When we investigate the false-positive (FP) and false-negative (FN) results on these test datasets, we see that majority of the false predictions are for children under six years old. This is an expected result, as there are not enough sample images for children under the age of six in any training set other than the Adience, Gallager, and UTKFace datasets. As a result of the experiments, we also achieved the highest accuracy in the Adience dataset using UTKFace as a training dataset, as it contains some child photos.

To understand the main reason behind the model's prediction, we visualized the filters and output of the convolutional layers. First, we checked the contents of the convolutional filters. Figure 2 shows the visualization of the filters obtained from the VGGFace2 dataset via gradient ascent in input space where deeper layers contain

Table 4. Cross-dataset accuracy results for the NoBAG_NoTTA and YesBAG_NoTTA experiments. Results obtained by averaging the result of five runs.

		Test datasets (NoBAG_NoTTA)											
		Adience	AFAD	CelebA	Gallagher	Genki-4K	IMDb	LFW	Morph	UtkFace	VGGFace2	Wiki	UA
Train datasets	Adience	-	87.34	93.51	82.39	89.31	94.22	92.38	92.00	89.28	93.36	91.15	90.49
	AFAD	79.55	-	90.77	80.87	88.58	91.62	92.36	90.68	87.30	91.64	89.64	88.30
	CelebA	83.63	92.33	-	88.26	97.22	98.97	98.17	95.34	91.84	98.51	97.19	94.15
	Gallagher	79.66	77.60	92.43	-	91.02	92.68	91.28	88.89	88.45	92.38	89.80	88.42
	Genki-4K	76.04	76.55	92.94	81.10	-	93.77	91.79	90.16	87.49	92.74	89.75	87.23
	IMDb	84.23	93.26	98.26	88.30	97.00	-	97.71	95.11	91.80	98.22	97.11	94.10
	LFW	78.31	83.36	95.83	81.93	92.53	96.76	-	91.66	88.86	95.52	93.69	89.85
	Morph	72.16	74.29	84.80	76.15	80.91	82.94	81.65	-	81.26	86.00	80.02	80.02
	UtkFace	82.39	89.66	96.30	82.39	93.68	96.80	96.13	94.16	-	96.29	94.54	92.23
	VGGFace2	83.95	93.28	98.87	89.30	97.40	99.14	98.74	96.39	92.54	-	97.98	94.76
	Wiki	80.89	87.53	97.33	86.37	96.49	97.99	97.30	94.48	90.31	97.34	-	92.60
UA		80.08	85.52	94.10	83.71	92.41	94.49	93.75	92.89	88.91	94.20	92.09	90.20

		Test Datasets (YesBAG_NoTTA)											
		Adience	AFAD	CelebA	Gallagher	Genki-4K	IMDb	LFW	Morph	UtkFace	VGGFace2	Wiki	UA
	Adience	-	89.82	94.81	84.11	91.10	95.72	93.86	93.11	90.88	94.86	92.92	92.12
	AFAD	81.21	-	92.53	83.11	90.32	93.46	93.76	91.63	88.88	93.32	91.39	89.96
	CelebA	84.31	94.42	-	89.22	97.84	99.21	98.54	95.82	92.17	98.85	97.80	94.82
	Gallagher	81.65	80.91	94.40	-	93.16	94.75	93.52	91.17	90.35	94.47	92.41	90.68
	Genki-4K	77.04	78.62	94.06	83.26	-	95.17	93.15	91.14	88.84	94.08	91.67	88.70
	IMDb	84.89	94.55	98.49	88.97	97.54	-	98.07	95.42	92.18	98.50	97.52	94.61
	LFW	79.34	85.78	96.69	83.51	94.23	97.70	-	93.17	89.76	95.52	95.14	91.08
	Morph	73.91	76.80	86.25	78.44	82.20	84.52	83.43	-	82.70	87.73	81.88	81.79
	UtkFace	84.13	91.85	97.01	83.56	95.28	97.57	96.72	94.51	-	97.04	95.55	93.32
	VGGFace2	84.73	94.52	99.09	90.17	97.98	99.33	98.91	96.74	92.80	-	98.42	95.27
	Wiki	82.01	90.54	97.79	87.45	97.47	98.55	98.11	95.75	90.85	97.95	-	93.65
UA		81.32	87.78	95.11	85.18	93.71	95.60	94.81	93.85	89.94	95.23	93.47	91.45

fine details and Score-CAM based activation maps to visualize the pixels that contribute to the prediction of the trained model. The visuals given in Figure 2 belong to the output of the convolutional layer for the given input images that have been correctly classified. According to the superimposed images, gender recognition results mostly depends on pixels around the eyes, mouth, chin, and hair boundaries. For male examples we also see that shirt collar contributes to the final prediction.

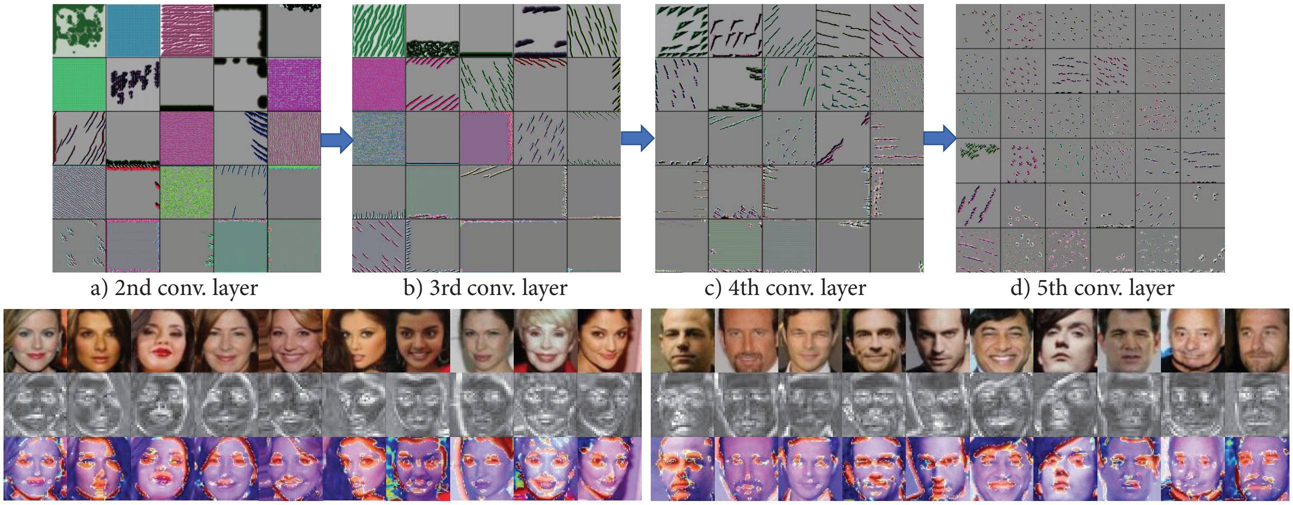
Table 5. Cross-dataset accuracy results for the NoBAG_YesTTA and YesBAG_YesTTA experiments. Results obtained by averaging the result of five runs.

		Test datasets (NoBAG_YesTTA)											
		Adience	AFAD	CelebA	Gallagher	Genki-4K	IMDb	LFW	Morph	UtkFace	VGGFace2	Wiki	UA
Train datasets	Adience	-	86.31	93.51	82.34	89.27	94.06	91.92	91.81	89.16	93.20	91.07	90.27
	AFAD	79.05	-	90.65	80.68	88.47	91.12	91.94	90.51	86.84	91.37	89.02	87.97
	CelebA	83.23	91.79	-	88.03	97.00	98.88	98.07	95.45	91.71	98.40	97.08	93.96
	Gallagher	79.04	74.75	91.44	-	90.22	91.57	89.71	88.79	87.59	91.56	88.83	87.35
	Genki-4K	75.91	76.31	92.69	80.72	-	93.68	91.39	90.27	87.21	92.48	89.74	87.04
	IMDb	83.73	92.68	98.18	87.91	96.87	-	97.71	95.17	91.77	98.11	97.03	93.92
	LFW	77.47	82.18	95.60	80.70	91.99	96.52	-	91.37	88.54	95.51	93.39	89.33
	Morph	71.91	73.50	84.22	75.11	79.90	82.49	80.76	-	80.89	85.06	79.94	79.38
	UtkFace	81.99	88.81	96.08	82.17	93.75	96.59	95.95	94.10	-	96.15	94.34	91.99
	VGGFace2	83.70	93.07	98.84	89.34	97.39	99.12	98.72	96.47	92.35	-	98.01	94.70
	Wiki	80.89	87.28	97.15	85.87	96.16	97.79	97.09	94.18	90.10	97.13	-	92.36
UA		79.69	84.67	93.84	83.29	92.10	94.18	93.33	92.81	88.62	93.90	91.85	89.84

		Test datasets (YesBAG_YesTTA)											
		Adience	AFAD	CelebA	Gallagher	Genki-4K	IMDb	LFW	Morph	UtkFace	VGGFace2	Wiki	UA
Train datasets	Adience	-	89.84	95.20	84.83	91.57	95.97	94.01	93.08	91.05	95.14	93.32	92.40
	AFAD	80.88	-	92.91	83.24	90.86	93.58	94.19	91.92	88.75	93.59	91.35	90.13
	CelebA	84.25	94.63	-	89.38	97.61	99.26	98.56	96.17	92.34	98.86	97.82	94.89
	Gallagher	81.75	78.49	94.12	-	93.86	94.71	93.13	91.74	90.52	94.40	92.26	90.50
	Genki-4K	77.56	79.93	94.20	83.42	-	95.70	93.36	91.51	88.81	94.37	92.16	89.10
	IMDb	84.74	94.71	98.50	89.14	97.77	-	98.21	95.71	92.23	98.50	97.64	94.72
	LFW	79.10	86.15	96.75	82.86	92.68	97.78	-	93.39	89.67	96.82	95.42	91.06
	Morph	73.79	77.14	86.32	78.33	82.30	84.92	83.24	-	82.85	87.46	82.53	81.89
	UtkFace	84.32	92.50	97.11	84.20	95.45	97.66	96.86	94.65	-	97.17	95.69	93.56
	VGGFace2	84.51	94.83	99.12	90.40	98.04	99.36	99.02	96.94	92.76	-	98.51	95.35
	Wiki	82.13	91.40	97.82	87.75	97.44	98.57	98.21	95.63	90.87	97.96	-	93.78
UA		81.30	87.96	95.21	85.36	93.76	95.75	94.88	94.07	89.99	95.43	93.67	91.58

4.4. Effect of bagging

To see the effect of the bagging method over the base models we provide ROC curves and AUC scores obtained from base models and the bagging model. Since we conducted 110 cross-dataset experiments, we only show some ROC curves obtained from AFAD tests as shown in Figure 3. According to Figure 3, it is clear that the bagging method provides a higher AUC score than the corresponding base models.



a) 2nd conv. layer b) 3rd conv. layer c) 4th conv. layer d) 5th conv. layer

Figure 2. Visualization of the filters via gradient ascent to stochastically explore intermediate feature maps. 64×64 cropped female and male examples from CelebA dataset. The left and right eye locations are fixed at $(20, 25)$ and $(45, 25)$ pixel locations respectively. The output from 2^{nd} convolutional layer is superimposed on the input image. The pixels that contribute to the final result most are the highlighted areas on the image.

We performed another experiment with 30 different models using Genki4K-LFW cross-dataset test. Figure 4 shows the standard model's accuracy and bagging accuracy. As the number of models increases, the bagging accuracy is almost steady in the range 93.88%–94.62%. However, the standard model's accuracy varies from 81.77% to 93.64%. The sudden accuracy changes in the standard model do not affect the bagging result. The use of 90 models in bagging gives 94.15% accuracy which is higher than our best experimental result using 9 models (93.36%). According to the results, we can conclude that increasing the number of models yields a positive effect on the prediction performance, with a cost of computational complexity.

We also compare our results with the state-of-the-art cross-dataset studies on the literature. Table 6 summarizes state-of-the-art results of different studies grouped by the test dataset. According to Table 6, our method provides the best accuracy on cross-dataset tests.

4.5. Computation time

We performed our experiments on an isolated machine. At 64×64 resolution, time to test one sample on Nvidia GTX 1060 GPU is 1.43 ms (694 images per second) for images loaded by mini-batch. The measured time does not include the time to load images from the disk. The complexity of the proposed approach is linear with the number of models that contribute to the final ensemble prediction. However, there is no dependency among the base models. Therefore, predictions can be performed using data and task parallel methods.

5. Conclusion

Deep learning methods are extensively used for the gender recognition problem due to their higher gender recognition rates. In general, model performances are measured within the same dataset using cross-validation techniques, which may produce results biasing towards the internal distribution of data. For this reason, we used cross-dataset evaluation instead of the traditional cross-validation based evaluation within the same dataset. Due to the challenging differences in human faces, we employed the bagging method. We further support our

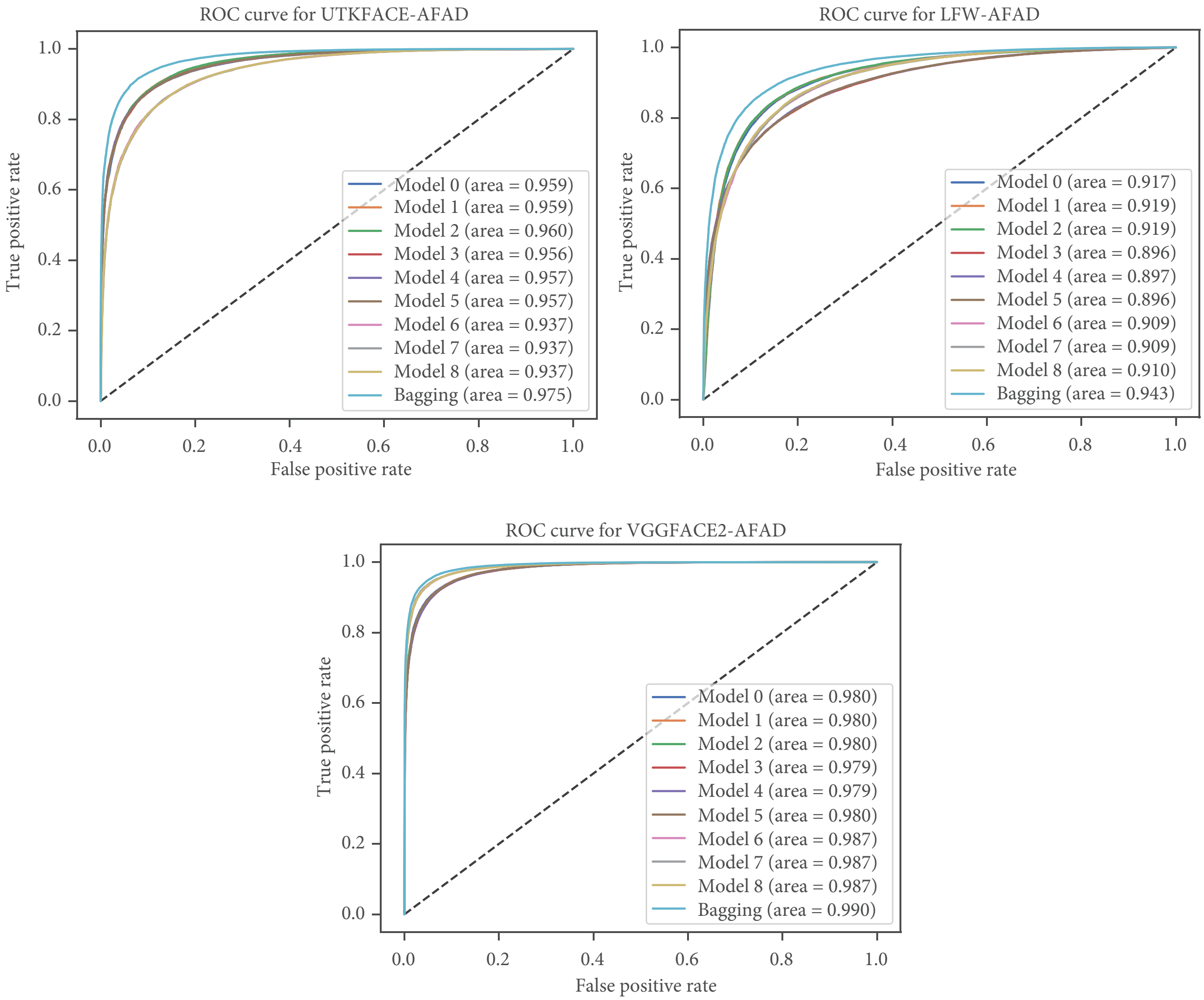


Figure 3. ROC curves for tests on the AFAD dataset. AUC scores of bagging method are higher than base model scores.

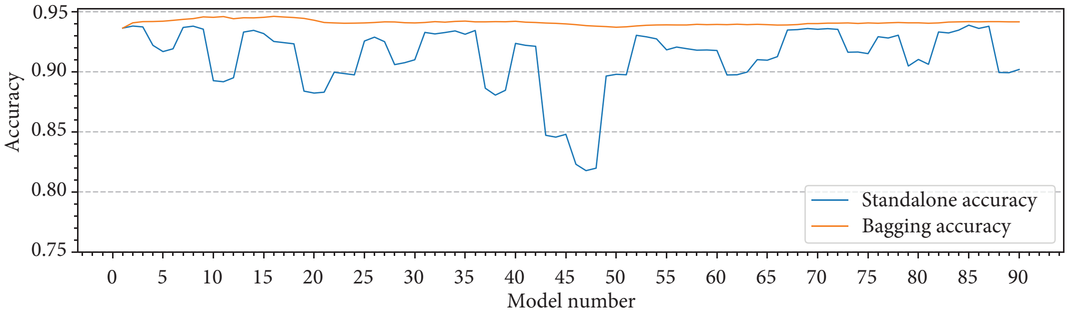


Figure 4. Effect of bagging using 30 models with TTA (3 TTA sample per model).

Table 6. Cross-dataset accuracy compared with the state-of-the-art. Bold numbers show the highest accuracy rate in the literature for the specified test data set.

Method	Year	Train set	Test set	Features	Accuracy%
SVM [26]	2012	Morph	Gallagher	LBP	76.74
Bagging+TTA CNN Ours	2020	Morph	Gallagher	Pixels	**78.33**
CNN [29]	2014	Gallagher	Adience	LBP+FLBP	77.80
MobileNet variant [4]	2020	400K	Adience	Pixels	84.48
Bagging+TTA CNN Ours	2020	VGGFace2	Adience	Pixels	84.51
CNN [33]	2017	Sighthound	Adience	Pixels	**91.00**
SVM [26]	2012	Morph	LFW	LBP	75.10
SVM+RBF [15]	2014	Morph	LFW	LBPHS	76.64
AlexNet variant [32]	2018	Morph	LFW	Pixels	77.40
Bagging+TTA CNN Ours	2020	Morph	LFW	Pixels	**83.24**
PCA+LDA [28]	2011	Gallagher	LFW	Pixels	81.07
PCA+SVM [28]	2011	Gallagher	LFW	Pixels	81.40
PCA+LDA [28]	2011	Gallagher	LFW	LBP	89.15
PCA+LDA [28]	2011	Gallagher	LFW	Gabor jets	89.27
PCA+SVM [28]	2011	Gallagher	LFW	Gabor jets	89.64
PCA+SVM [28]	2011	Gallagher	LFW	LBP	89.77
Bagging+TTA CNN Ours	2020	Gallagher	LFW	Pixels	**93.13**
MobileNet variant [4]	2020	VGGFace2	LFW	Pixels	98.73
Bagging+TTA CNN Ours	2020	VGGFace2	LFW	Pixels	**99.02**
SVM+RBF [15]	2014	LFW	Morph	LBPHS	88.43
AlexNet variant [32]	2018	LFW	Morph	Pixels	89.00
Bagging+TTA CNN Ours	2020	LFW	Morph	Pixels	**93.39**
Bagging+TTA CNN Ours	2020	VGGFace2	Morph	Pixels	**96.94**
MobileNet variant [4]	2020	VGGFace2	IMDb	Pixels	80.74
Bagging+TTA CNN Ours	2020	VGGFace2	IMDb	Pixels	**99.36**
MobileNet variant [4]	2020	VGGFace2	Wiki	Pixels	95.78
Bagging+TTA CNN Ours	2020	VGGFace2	Wiki	Pixels	**98.51**
SVM [18]	2014	Gallagher	Genki4-K	Pixels	91.07
Bagging+TTA CNN Ours	2020	Gallagher	Genki-4K	Pixels	**93.86**
Bagging+TTA CNN Ours	2020	VGGFace2	Genki-4K	Pixels	**98.15**
Bagging+TTA CNN Ours	2020	VGGFace2	CelebA	Pixels	**99.12**
Bagging+TTA CNN Ours	2020	VGGFace2	AFAD	Pixels	**94.83**
Bagging+TTA CNN Ours	2020	VGGFace2	UTKFace	Pixels	**92.76**
Bagging+TTA CNN Ours	2020	CelebA	VGGFace2	Pixels	**98.85**

model by test-time augmentation. According to the experiments on a wide variety of datasets, we showed that the proposed method provides state-of-the-art results for cross-dataset gender recognition. Our experiments showed that low ethnic diversity data sets like Morph are not appropriate for gender recognition. We conclude that VGGFace2, CelebA, and IMDb datasets provide better average accuracy than other datasets.

References

[1] Afifi M. 11K Hands: gender recognition and biometric identification using a large dataset of hand images. Multimedia Tools and Applications 2019; 78 (15): 20835-20854. doi: 10.1007/s11042-019-7424-8

[2] Yaman D, Eyiokur FI, Sezgin N, Ekenel HK. Age and gender classification from ear images. In: International Workshop on Biometrics and Forensics; Sassari, Italy; 2018. pp. 1-7. doi: 10.1109/IWBF.2018.8401568

[3] Rodríguez P, Cucurull G, Gonfaus JM, Roca FX, Gonzàlez J. Age and gender recognition in the wild with deep attention. Pattern Recognition 2017; 72: 563-571. doi: 10.1016/j.patcog.2017.06.028

[4] Greco A, Saggese A, Vento M, Vigilante V. A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. IEEE Access 2020; 8: 130771-130781.

[5] Lee J, Chan Y, Chen T, Chen C. Joint Estimation of Age and Gender from Unconstrained Face Images Using Lightweight Multi-Task CNN for Mobile Applications. In: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR); Miami, FL, USA; 2018. pp. 162-165. doi: 10.1109/MIPR.2018.00036

[6] Abdurrahim SH, Samad SA, Huddin AB. Review on the effects of age, gender, and race demographics on automatic face recognition. The Visual Computer 2018; 34 (11): 1617-1630.

[7] Sun Y, Zhang M, Sun Z, Tan T. Demographic analysis from biometric data: achievements, challenges, and new frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018; 40 (2): 332-351. doi: 10.1109/TPAMI.2017.2669035

[8] Bekios-Calfa J, Buenaposada JM, Baumela L. Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters 2014; 36: 228-234. doi: 10.1016/j.patrec.2013.04.028

[9] Phillips PJ, Wechsler H, Huang J, Rauss PJ. The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 1998; 16 (5): 295-306. doi: 10.1016/S0262-8856(97)00070-X

[10] Ng CB, Tay YH, Goi BM. Recognizing human gender in computer vision: a survey. In: Anthony P, Ishizuka M, Lukose D (editors). PRICAI 2012: Trends in Artificial Intelligence. PRICAI 2012. Lecture Notes in Computer Science, Vol. 7458. Springer, Berlin, Germant: Springer, 2012, pp. 335-346. doi: 10.1007/978-3-642-32695-0_31

[11] Ng CB, Tay YH, Goi BM. A review of facial gender recognition. Pattern Analysis and Applications 2015; 18 (4): 739-755. doi: 10.1007/s10044-015-0499-6

[12] Seni G, Elder J. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. San Rafael, CA, USA: Morgan & Claypool, 2010. doi: 10.2200/S00240ED1V01Y200912DMK002

[13] Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. IEEE Computational Intelligence Magazine 2018; 13 (4): 59-76. doi: 10.1109/MCI.2018.2866730

[14] Rao RB, Fung G. On the dangers of cross-validation. an experimental evaluation. In: Proceedings of the SIAM International Conference on Data Mining; Atlanta, GA, USA; 2008. pp. 588-596. doi:10.1137/1.9781611972788.54

[15] Erdoğmuş N, Vanoni M, Marcel S. Within- and cross- database evaluations for face gender classification via befit protocols. In: 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP); Jakarta, Indonesia; 2014. pp. 1-6.

[16] Jain A, Huang J, Fang S. Gender identification using frontal facial images. In: 2005 IEEE International Conference on Multimedia and Expo; Amsterdam, Netherlands; 2005. p. 4. doi:10.1109/ICME.2005.1521613

[17] Bin Xia, He Sun, Bao-Liang Lu. Multi-view gender classification based on local Gabor binary mapping pattern and support vector machines. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); Hong Kong, Hong Kong; 2008. pp. 3388-3395. doi: 10.1109/IJCNN.2008.4634279

[18] Danisman T, Bilasco IM, Djeraba C. Cross-Database Evaluation of Normalized Raw Pixels for Gender Recognition under Unconstrained Settings. In: 22nd International Conference on Pattern Recognition; Stockholm, Sweden; 2014. pp. 3144-3149. doi: 10.1109/ICPR.2014.542

[19] Phung SL, Bouzerdoum A. A pyramidal neural network for visual pattern recognition. IEEE Transactions on Neural Networks 2007; 18 (2): 329-343. doi:10.1109/TNN.2006.884677

[20] Baluja S, Rowley HA. Boosting sex identification performance. International Journal of Computer Vision 2007; 71 (1): 111-119. doi: 10.1007/s11263-006-8910-9

[21] Mäkinen E, Raisamo R. Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008; 30 (3): 541-547. doi: 10.1109/TPAMI.2007.70800

[22] Jia S, Lansdall-Welfare T, Cristianini N. Gender classification by deep learning on millions of weakly labelled images. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW); Los Alamitos, CA, USA; 2016. pp. 462-467. doi: 10.1109/ICDMW.2016.0072

[23] Levi G, Hassncer T. Age and gender classification using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Boston, MA, USA; 2015. pp. 34-42. doi:10.1109/CVPRW.2015.7301352

[24] Mansanet J, Albiol A, Paredes R. Local Deep Neural Networks for gender recognition. Pattern Recognition Letters 2016; 70: 80-86. doi: 10.1016/j.patrec.2015.11.015

[25] Ranjan R, Patel VM, Chellappa R. HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019; 41 (1): 121-135. doi: 10.1109/TPAMI.2017.2781233

[26] Ramón-Balmaseda E, Lorenzo-Navarro J, Castrillón-Santana M. Gender Classification in Large Databases. In: Alvarez L, Mejail M, Gomez L, Jacobo J (editors). Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Berlin, Germany: Springer, 2012, pp. 74-81.

[27] Castrillón-Santana M, Lorenzo-Navarro J, Ramón-Balmaseda E. Improving Gender Classification Accuracy in the Wild. In: Ruiz-Shulcloper J, Di Baja G (editors). Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Berlin, Germany: Springer, 2013, pp. 270-277.

[28] Dago-Casas P, González-Jiménez D, Long Long Yu, Alba-Castro JL. Single- and cross- database benchmarks for gender classification under unconstrained settings. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops); Barcelona, Spain; 2011. pp. 2152–2159.

[29] Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security 2014; 9 (12): 2170-2179.

[30] Kang MW, Kim Y, Kim YS. Collecting large training dataset of actual facial images from facebook for developing a weighted bagging gender classifier. Cluster Computing 2017; 20 (3): 2157-2165. doi: 10.1007/s10586-017-0958-5

[31] Tapia JE, Perez CA. Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape. IEEE Transactions on Information Forensics and Security 2013; 8 (3): 488-499.

[32] Han H, Jain AK, Wang F, Shan S, Chen X. Heterogeneous face attribute estimation: a deep multi-task learning approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018; 40 (11): 2597-2609.

[33] Dehghan A, Ortiz E, Shu G, Masood S. DAGER: deep age, gender and emotion recognition using convolutional neural network. arXiv 2017. arXiv:1702.04280.

[34] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv 2017. arXiv:1704.04861.

[35] Viola P, Jones MJ. Robust real-time face detection. International Journal of Computer Vision 2004; 57 (2): 137-154. doi: 10.1023/B:VISI.0000013087.49260.fb

[36] Wang F, Chen L, Li C, Huang S, Chen Y et al. The devil of face recognition is in the noise. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham, Switzerland: Springer International Publishing, 2018, pp. 780-795.

[37] Rowley HA, Baluja S, Kanade T. Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998; 20 (1): 23-38. doi: 10.1109/34.655647

[38] Milborrow S, Nicolls F. Locating Facial Features with an Extended Active Shape Model. In: Forsyth D, Torr P, Zisserman A (editors). Computer Vision – ECCV 2008. Berlin, Germany: Springer, 2008, pp. 504-513.

[39] Niu Z, Zhou M, Wang L, Gao X, Hua G. Ordinal regression with multiple output CNN for age estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. pp. 4920-4928.

[40] Liu Z, Luo P, Wang X, Tang X. Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV); Santiago, Chile; 2015.

[41] Gallagher AC, Chen T. Understanding images of groups of people. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops); Los Alamitos, CA, USA; 2009. pp. 256-263. doi: 10.1109/CVPR.2009.5206828

[42] Rothe R, Timofte R, Gool L Van. DEX: deep expectation of apparent age from a single image. In: IEEE International Conference on Computer Vision Workshops (ICCVW); Santiago, Chile; 2015.

[43] Huang GB, Ramesh M, Berg T, Learned-Miller E. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[44] Ricanek K, Tesafaye T. MORPH: a longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06); Southampton, UK; 2006. pp. 341-345.

[45] Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. pp. 4352-4360.

[46] Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. VGGFace2: a dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition; Xi’an, China; 2018.