adcaij.bib

Real-world human gender classification from oral region using convolutional neural netwrok

Mohamed Oulad-Kaddour[a], Hamid Haddadou[a], Cristina Conde[b], Daniel Palacios-Alonso[b], and Enrique Cabello[b]

[a]Laboratoire de la Communication dans les Systèmes Informatiques, Ecole naionale Supéerieure en Informatique , BP 68M, 16309, Oued-Smar, Algiers, Algeria

[b]Rey Juan Carlos University, C/Tulipan, s/n,28933, Mostoles, Madrid, Spain. m_ouled_kaddour@esi.dz

KEYWORD

Gender classification; face biometrics; oral region' biometrics; convolutional neural networks; deep learning

ABSTRACT

Gender classification is an important biometric task. It was widely studied in the literature. Face modality is the most studied for human-gender classification in the practice. Moreover, the task was also investigated from face components like as iris, ear, and periocular region. In this paper, we aim to investigate gender classification by exploiting the oral region based on the mouth. In the proposed approach, we adopt a convolutional neural network. For experimentation, we extracted the region of interest using the RetinaFace algorithm from the FFHQ dataset faces. We achieved acceptable results surpassing those using the mouth as a modality or facial sub-region in geometric approaches. Obtained results also show the importance of the oral region hidden in the Covid-19 context for which we suppose that the adaptation of existing solutions is indispensable.

1. Introduction

Human-gender classification is a widely studied task. It is one of the most active research areas in biometrics. This is due to the various fields of its usability such as security, video surveillance, robotic and demographic collection. Human-gender classification describes a binary classification problem. It's generally studied focalizing on face modality. However, it was also investigated from others modalities like as fingerprints (S. Tarare and Turkar, 2015), hand (Affifi, 2019), face (Makinen and Raisamo, 2008), ears (D. Yaman and Ekenel, 2018) and iris (Aravena, 2017).

In this paper, we aim to investigate an experimental study for human-gender classification from the oral region based on the mouth. Figure 1 shows the lower face' components and the targeted region. As motivation for our choice of the oral region as a region of interest in the proposed approach, we cite:

- Acceptability and privacy: Naturally, we consider that for any biometric modality targeting a region

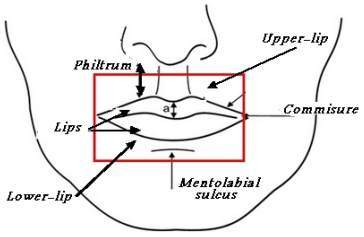

Figure 1: Lower face' components

of interest (ROI) each derived modality obtained by focusing on a sub-region of the original ROI will enhance acceptability. In the case of facial data, even if face modality is sufficiently acceptable by people, we suppose that the oral region is more acceptable in comparison with the whole face. Indeed, people' identities are highly hided and their privacy is well protected.

- Eventual alternative for face modality: Furthermore that oral region can be considered as a biometric modality (Choras, 2010), there are some scenarios where the lower face or moth describe the principal part available like in the case of combined occlusion (hat and black glasses-wearing) or offensive attack. Consequently, the oral region can be an alternative for the whole face' modality.
- Modality' characteristics: As a biometric modality, the oral region is rich in gender' related information and texture. Indeed, the difference in lips angle and format can be observed for both gender' classes subjects (Figure 2). In addition and for a considerable part of people, it contains gender reserved synthetics information's like lipstick for the female class and mustache presence for male class (Figure 2).

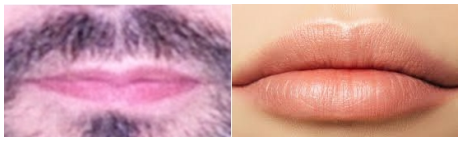

Figure 2: Samples showing some difference inter Male versus female mouth

The proposed approach is deployed by adopting deep learning techniques, namely: convolutional neural networks. Indeed, convolutional neural networks or CNNs describe the state-of-the-art for a lot of image classification problems because of achieved results (Alzubaidi, 2016). In addition, CNNs allow releasing researchers from the classical features engineering obstacle as they assured the deployment of powerful solutions in comparison with priors techniques' based ones.

1.1 Related works

Darryl Stewart et al. (Darryl Stewart, 2013) studied automatic gender classification using static and dynamic features. After face and mouth localization, they followed a standard-DCT (discrete cosine transform) based process for features extraction. They reported reasonable results for speaker-independent gender classification using the publicly available XM2VTS dataset. Best score' values of 82.19% and 18.36% were respectively achieved for both accuracy and EER(Equal Error Rate) metrics.

In (T.X. Wu, 2012), T.X. Wua et al. proposed a multi-view gender classification approach using facial components symmetry. They segmented the face on four facial components, namely: eyes, nose, mouth, and chin. They used the SVM (support vector machine) technique as a sub-classifier for each component. Sum, maximum, product, and fuzzy integral were studied for sub-classifer combinations. Accuracy of 82.23% was got for the mouth region.

Bing Li et al. (B. Li, 2012) proposed a framework for human-gender classification by combining facial and external information. They considered five facial regions: forehead, eyes, nose, mouth, and chin with two additional parts: hair and clothing. They exploited LBP (local binary pattern) operator for feature extraction and SVM for classification. The sub-regions scores were combined by using various strategies including sum, product, maximum, and majority voting rules. FERET and BCMI dataset faces were exploited for experiments. Respective best classification rates of 82.9% and 83.6% were obtained on the region (mouth) for FERET and BCMI datasets.

Rai and Khana (Preeti Rai, 2014) investigated the application of artificial occlusions on the face to perform an occlusion robust system. They defined various occlusion generation strategies by blacking face parts. They used Gabor filters and PCA (principal component analysis). The best classification rate of 85.3% was achieved on FERET faces by keeping the lower via upper face' blacking.

Affifi and Abdelhamid (Afifi and Abdelhamed, 2019) used five isolated facial components including the oral region (mouth) for deep gender classification. They used four convectional neural networks and Adaboost algorithms for final classification for respective sub-facial images: fuzzy face, both eyes, mouth, and nose. The final classification decision was performed by using a linear discriminant classifier. They achieved an accuracy rate of 89.09% for the mouth region.

Approach	Year	Principle	best score on oral region
T.X (T.X. Wu, 2012)	2012	SVM	82.8%
Darry (Darryl Stewart, 2013)	2013	DCT	82.19%
Bing Li (B. Li, 2012)	2012	LBP+SVM	83.6%
Rai (Preeti Rai, 2014)	2014	Gabor filters + PCA	85.3%
Affifi (Afifi and Abdelhamed, 2019)	2019	CNNs + Adaboost	89.05%

Table 1: State of the art analysis

Table 1 recapitulates a comparative analysis of the state-of-the-art' reference existing works previously discussed by making their principles and best gender classification rate. The oral region was also used for automatizing of other real-world classification problems. Mouth status estimation (Jie Cao and He, 2016), person identification (Darryl Stewart, 2013), and lip-reading (Shrestha, 2018) are examples of artificial intelligence tasks already studied from the mouth.

Jie Cao et al. (Jie Cao and He, 2016) propose the use of a deep convolutional neural network for mouth status' estimation in the wild by taking into consideration various types of attacks. In experimentation, they performed subject-dependent (SI) corresponding to data repletion inter train and test sets, and subject-independent (SI) to avoid data repletion inter the train and test sets. Respective accuracy of 90.5% and 84.4% were achieved for both SD and SI experiments.

Karan (Shrestha, 2018) suppose lip-reading as a difficult task. He proposes the training of two deeper separate CNN architectures for real-time word prediction. He used the Haar classifier for face and mouth localization. He applied batch normalization to speed up the training process and better stability. A dropout rate of 40% was empirically fixed for the network generalization and overfitting reduction. Best validation accuracy of 77.14% was obtained.

Yannis et al. (Yannis M. Assael, 2016)proposed LipNet, a recurrent neural network exploiting spatiotemporal convolutions for lip-reading. LipNet is the first model trained with an end-to-end strategy for lip movement in videos frames translation to text. LipNet surpass priors works and achieved new stat-of-the-art accuracy of 95.2% in sentence-level.

In (Wright, 2020) Wright et al. proposed LipAuth, a system for lib-based authentication for mobile devices. They trained convolutional neural networks inppisred by LipNet (Yannis M. Assael, 2016). They used the XM2VTS dataset and collected new datasets, qFace and FAVLIPS. An equal error rates of 1.65% was obtained by benchemarrking the system on the XM2VTS.

2. Methodology

In this section the proposed methodology is presented by presenting its overview and the used convolutional neural network.

2.1 Overview

The overview of our approach is illustrated in the Figure 3. At first we extract the facial annotations corresponding to left and right eyes, nose, left and right mouth's commissure. We computed the five facial annotations with the RetinaFace algorithm (Jiankang Deng, 2019). Secondly, we perform the localisation of the oral region by exploiting the left and right commissure. Finally, we passed the extracted region of interest to the convolutional neural network (CNN) for processing and decision making.

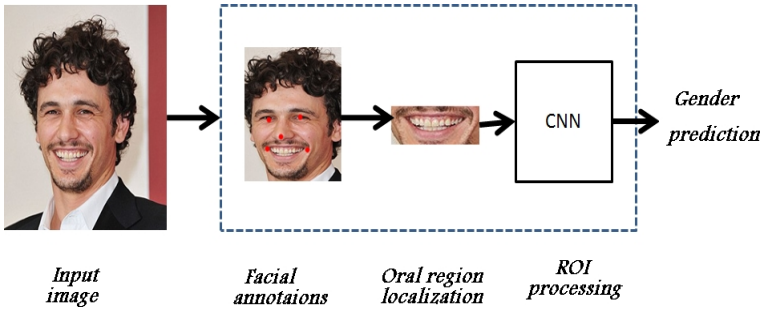

Figure 3: Overview of the proposed methodology

2.2 Convolutional neural network

Convolutional neural networks or CNNs are powerful learning based technique for image classification. A CNN integrates principally following layer types: convolutional, pooling, batch-normalization, dropout, fully connected.

- Convolutional layer: It is one of fundamental layer' types in a CNN. In this stage, convolutional operator is applied on the input in the goal to learn features that will be detected in future subjects. A convolutional

Layer (type)	Output Shape	Param
conv2d_1 (Conv2D)	(40, 85, 32)	896
batch_normalization (BatchNo)	(40, 85, 32)	128
max_pooling2d (MaxPooling2D)	(20, 42, 32)	0
conv2d_2 (Conv2D)	(18, 40, 64)	18496
batch_normalization_2	(18, 40, 64)	256
max_pooling2d_2	(9, 20, 64)	0
conv2d_3 (Conv2D)	(7, 18, 128)	73856
max_pooling2d_3	(3, 9, 128)	0
flatten_3 (Flatten)	(3456)	0
batch_normalization	(3456)	13824
FC (Dense)	(256)	884992
batch_normalization	(256)	1024
dropout (Dropout)	(256)	0
batch_normalization	(256)	1024
dense (SoftMax)	(None, 2)	514

Table 2: Proposed CNN architecture' details

operator is mainly characterized by its kernel size.

- Pooling layer: In this type of layers, pooling operators are applied in the goal to reduce the data-dimensionality. They are generally used following convolutionals layers generating maps of high resolution. Max-pooling and average-pooling are example of most used pooling methods.
- Batch normalization: It is an artificial intelligence technique that aims to speed up the training of very deeper neural networks, to keep stability during training phases.
- Fully connected layers: In a fully connected layers (FC), neurons are connected with with all activation's in the previous layer. FC layers are generally placed at the end of CNNs.
- Dropout layers: In deep learning, a dropout layer is a layer that aims to avoid the overfitting of trained CNN. The simple way how works a dropout layer is the performing the forgetting of a part of their input. Dropout layers are generally placed following FCs ones.

In our case, we adopted a sequential convolutional neural network by exploiting various layers types, namely: convolutional (conv2d), batch normalization (BatchNo), maximum pooling (MaxPooling2D), fully connected (Dense), and dropout layers. We used a binary softmax layer for the final decision. The detail of the proposed CNN is shown in Table 2.

3. Experimental setup

The experimental setup is presented in this section by presenting the used dataset, the CNN' training conditions and the evaluation metrics.

3.1 **Dataset**

To realize experimentation, we used the FFHQ (flickr faces high quality) dataset. it is a recent real-world dataset collected for training the StyleGAN (T. Karras and Aila., 2010), an adversarial neural network generating realistic artificial fake corpus. FFHQ is a rich dataset in terms of facial variations, like age, ethnicity, background, facial expressions, and occlusion. It contains 70K faces extracted from images acquired in the wild under high resolution.

	size in faces	size in subjects	Gender-labeling
FFHQ	∼70k	∼70k	Labeled

Table 3: FFHQ dataset' details

The Table 3 resumes the FFHQ dataset details by giving its size in faces, its size in subjects and the gender labeling information. The figure 4 shows samples from the dataset.

Figure 4: FFHQ samples

3.2 CNN' pre-training and data augmentation

As previously illustrated in table 2, we fixed the input size at 40*85. For better CNN's weights initialization, we firstly affected them small random values. Then, we pre-trained the network by using the dog-cat dataset publicly available on the Kaggle framework. During the network' training, we realized a real-time batch augmentation by using the horizontal flip.

GPU size	GPU type	Specs
12 GO	Nvidia	Telsa K80

Table 4: Used GPU memory characteristics

According to deep learning techniques material requirements, all experiments were performed on machine integrating GPU (graphics processing unit) in the goal to speed up the training time (Table 4). We used 75 epochs in training and we called back the best settings got in intermediary epochs. We used a small batch size of 10 and we experimentally fixed the dropout rate to a potion of 0.25.

3.2.1 Evaluation metrics

To evaluate experimentally the trained network, we used various metrics based on next parameters: TP (true positive) , TN (true negative), FP (false positive) and FN (false negative). Following metrics were used as performance indicators: accuracy (ACC), the rate of true positive (RTP) in an objective class, equal error rate (EER) and receiver operating characteristic (ROC). The accuracy, describe the rate of correctly classified subjects in the whole test-set. It is defined as follow:

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

The rate of true positive describe the rate of correctly classified subjects intra-class. It is computed as following for a given targeted class:

$$RTP = \frac{TP}{TP + FP} \tag{2}$$

The equal error rate is determined by computing medium value for optimal values FAR_opt and FRR_opt where FAR=FP/(FP+TN) and FRR=FN/(TP+FN). The EER formul is :

$$EER = \frac{FAR_{opt} + FRR_{opt}}{2} \tag{3}$$

The receiver operating characteristic is a curve allowing a graphical interpretation for the CNN' comportment in front of test-set. It is dressed by monitoring the evolution of FAR and FRR values' evaluation during test.

4. Results and discussion

In this section, the obtained experimental results for the cited metrics are presented and discussed.

4.1 ACC, RTP

For the training of the network, we used a gender-balanced set of 5k images from the FFHQ dataset. For the test, we randomly selected a gender-balanced set of 1000 images. All images correspond to the extracted oral region from frontal and semi-profile faces. The facial variations of the used sets are recapitulated in Table 5 in terms of facial expression, face pose, image quality, age, and ethnicity. Training and test' set details are resumed in Table 6.

	Female	Male	totale
Train (in kilo images)	5k	5k	10k
Test	500	500	1000

Table 5: Train and test size

Variation	Observation
Facial expression	Random
Face pose	Frontal and semi-profile
Image Qality	Acceptable
Age	Random, except young
Ethnic	Random: Black, White, Asiatic,...

Table 6: Train and test set's facial variations

The obtained confusion matrix is shown in table 7. Table 8 resumes the achieved accuracy (ACC) and RTPs (FRP and MRTP) describing the rate of correctly classified subjects intra-classes. Well-accuracy of 92.70% is achieved for the whole test-set. For female and male genders we got respective values of 94.00% ad 91.40% for

		Predicted class	
		Female	Male
Real class	Female	457	43
	Male	30	470

Table 7: Confusion matrix

	ACC	FTP	MTP
FFHQ	92.70%	91.40%	94.00%

Table 8: Obtained ACC, FTP and MTP

the FTP and MTP parameters. We note that the male gender is well detectable in comparison with the female gender. It can be justified by the fact that for the male gender, the upper-lip texture is easily detectable by the trained convolutional neural network, especially in the case of mustached subjects' faces.

4.2 ROC-AUC, EER

Figure 5 shows the ROC curve traced for the tested network. Table 9, resumes the EER error and the AUC (area under the curve) describing the area covered by the ROC curve. By looking at the ROC curve we note clearly the convergence of the classifier as a good discriminator. At the same, a quantitative value of 0.966 closer to 1 is obtained for the AUC parameter and a high area is covered. In Addition, a low EER error of 0.103 is achieved. It allows asserting that the human gender can normally predicted from the oral region.

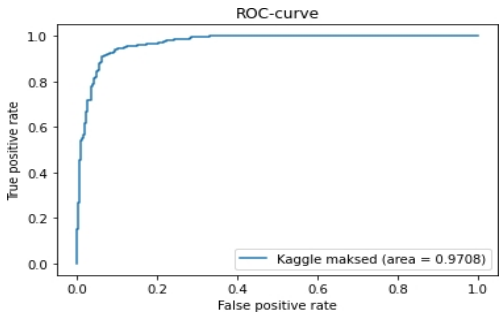

Figure 5: ROC curve

Figure 6 shows samples of correctly classified and misclassified subjects from the test set. We note some factors affecting the classification results by looking at the whole set of misclassified subjects: facial expressions and upper lip texture. Indeed, we observed that the classifier is more sensitive for smiling faces as a lot of

Parameter	Value
EER	0.0798
AUC	0.9708

Table 9: Obtained AUC, EER (in probabilistic info.)

Figure 6: Example of correctly classified and classified subject
(a): misclassified males (b): misclassified females
(c): correctly classified males (b): correctly classified females

misclassified subjects are with a smiling faces. For the second factor of upper lip' texture, we noted that almost of misclassified subjects from the male class are mustacheless where the visual textures inter both gender' classes are close. The last factor of age, we cited it for old persons and it can be justified by the way that for old age a considerable discriminative texture is lost as well as the lip format. In addition, we observed that the face pose and ethnic variations do not affect remarkably the prediction rates.

4.3 Features visualisation

In the goal to detect the most discriminative parts where the trained convolutional neural network is watching for gender classification decision making, we used the Grad-Cam technique. We computed activation maps by applying the Grad-Cam technique on the last convolutional operator of the networks.

We computed the Grad-Cam class activation maps on several subjects from both female and male classes. The Figure shows samples of obtained results. By looking at the obtained maps, we asserted that for the female gender, lip texture and commissure are oral region' parts where networks look to make decision. For the male gender, we noted that lower-lip and upper-lip textures describe the most important parts for which the network makes more attention to make decision.

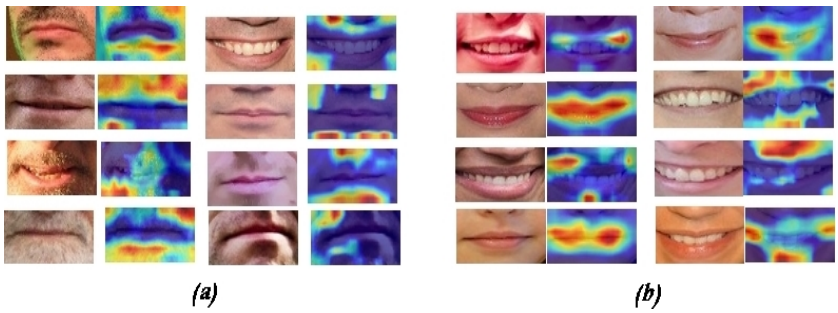

Figure 7: Example of features visualisation for both gender' classes
(a): male samples (b):females sample

4.4 Baseline comparison

Finally, after evaluating the proposed approach, we realized a baseline comparison with priors works using oral region' related part as a biometric modality or facial region. The comparative Table 10 is made by making, for each work, the best achieved accuracy.

Approach	Observation	Best score on oral region
Bing Li (B. Li, 2012)	Use mouth as facial part in geometric approach	83.6%
T.X (T.X. Wu, 2012)	Use mouth as facial part in geometric approach	82.8%
Darrry (Darryl Stewart, 2013)	Gender classification from mouth	82.19%
Rai (Preeti Rai, 2014)	Use lower face based on mouth	85.3%
Affifi (Afifi and Abdelhamed, 2019)	Use mouth as facial part _in geometric approach	89.05%
Proposed	Gender classification from mouth	92.70%

Table 10: Baseline comparison

As can be seen in Table 10, we got a gender classification rate surpassing those achieved in the literature. Indeed, we achieved a greater classification rate for the real-world human gender attribute from the oral region on a large test set of 1000.

5. Conclusion

In this paper, we proposed a deep learning-based approach for human gender classification from the oral region. For the extraction of the region of interest, we used the RetinaFace algorithm. As a classifier, we adopted a convolutional neural network. To perform our experimentation's we used facial images of frontal regular and semi-profile from the real-world FFHQ dataset. In tests, well results were achieved as low EER was obtained and global accuracy of 92.70% was returned for a test of 1000 images. Experimental results proclaim the feasibility of using the oral region as a biometric modality and show its importance as a facial part for human gender prediction. Finally, some factors affecting the classification were discussed and a comparison with prior approaches using oral region was performed.

6. References

Affifi, M., 2019. 11K hands: Gender recognition and biometric identification using a large dataset of hand images. In *Multimedia Tools and Applications*.

Afifi, M. and Abdelhamed, A., 2019. AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces. In *Journal of Visual Communication and Image Representation*. Elsivier.

Alzubaidi, Z. J. H. A. e. a., L., 2016. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *J Big Data*. Springer.

Aravena, J. T. C., 2017. Gender Classification from NIR Iris Images Using Deep Learning. In *Deep Learning for Biometrics*. Springer.

B. Li, B. L., X.C. Lian, 2012. Gender classification by combining clothing, hair and facial component classifiers. In *Neurocomputing*. Elsivier.

Choras, 2010. The lip as a biometric. In *Pattern Anal Applic.* Springer.

D. Yaman, N. S., F. I. Eyiokur and Ekenel, H. K., 2018. Age and gender classification from ear images. In *In International Workshop on Biometrics and Forensics*. IEEE.

Darryl Stewart, J. Z., Adrian Pass, 2013. Gender classification via lips: static and dynamic features. In *IET Biometrics*.

Jiankang Deng, Y. Z. J. Y. I. K. S. Z., Jia Guo, 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. arXiv:1905.00641.

Jie Cao, Z. S., Haiqing Li and He, R., 2016. Accurate mouth state estimation via convolutional neural networks. In *IEEE International Conference on Digital Signal Processing (DSP)*. IEEE.

Makinen, E. and Raisamo, R., 2008. Evaluation of gender classification methods with automatically detected and aligned faces. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE.

Preeti Rai, P. K., 2014. A gender classification system robust to occlusion using Gabor features based (2D)2PCA. In *J. Vis. Commun. Image R.*

S. Tarare, A. A. and Turkar, H., 2015. Fingerprint Based Gender Classification Using DWT Transform. In *International Conference on Computing Communication Control and Automation*.

Shrestha, K., 2018. Lip Reading using Neural Network and Deep learning. arXiv:1802.05521v1.

T. Karras, S. L. and Aila., T., 2010. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

T.X. Wu, B. L., X.C. Lian, 2012. Multi-view gender classification using symmetry of facial images. In *Neural Comput. Appl.*

Wright, S. D., C., 2020. Understanding visual lip-based biometric authentication for mobile devices. In *EURASIP J. on Info. Security*.

Yannis M. Assael, S. W. N. d. F., Brendan Shillingford, 2016. LipNet: End-to-End Sentence-level Lipreading. arXiv:1611.01599v2.