ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 11 N. 3 (2022), 249-261
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.27797

Real-world Human Gender Classification from Oral Region using Convolutional Neural Network

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

KEYWORDS

gender classification; face biometrics; oral region biometrics; convolutional neural networks; deep learning

ABSTRACT

Gender classification is an important biometric task. It has been widely studied in the literature. Face modality is the most studied aspect of human-gender classification. Moreover, the task has also been investigated in terms of different face components such as irises, ears, and the periocular region. In this paper, we aim to investigate gender classification based on the oral region. 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 faces dataset. We achieved acceptable results, surpassing those that use the mouth as a modality or facial sub-region in geometric approaches. The obtained results also proclaim the importance of the oral region as a facial part lost in the Covid-19 context when people wear facial mask. We suppose that the adaptation of existing facial data analysis solutions from the whole face is indispensable to keep-up their robustness.

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. Its study is generally focused on face modality. However, it was also investigated from others modalities, such as fingerprints (Tarare et al., 2015), hands (Affifi, 2019), the periocular region (Busch, 2019), ears (Yaman et al., 2018) and irises (Tapia and Carlos, 2017).

In this paper, an experimental study has been carried out on human-gender classification based on the oral region, specifically on the mouth. Figure 1 shows the parts of the lower face and the targeted region. The scientific literature cited hereafter has been the motivation for our choice of the oral region and of the proposed approach:

- Acceptability and privacy: Naturally, we consider that for any biometric modality targeting a region 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's identities are well hidden and their privacy is protected.
- Eventual alternative for face modality: The oral region can be considered as a biometric modality (Choras´, 2010), there are some scenarios where the lower face or mouth describe the principal part available as in the case of combined occlusion (wearing a hat and black glasses) or an 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 lip angle and format can be observed for the subjects of both gender classes (Figure 2). In addition and for a considerable part of people, it contains gender-reserved synthetic information, such as lipstick for the female class and mustache presence for the male class (Figure 2).

The proposed approach is deployed by adopting deep learning techniques, namely: convolutional neural networks. Indeed, convolutional neural networks or CNNs have been employed in state-of-the-art literature to solve image classification problems because of achieved results (Alzubaidi et al., 2021). In addition, CNNs enable researchers to overcome classical features engineering challenges, as they assure the deployment of powerful solutions in comparison to prior techniques.

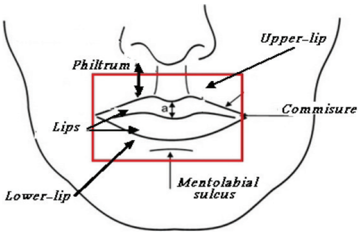

Figure 1. Lower face components

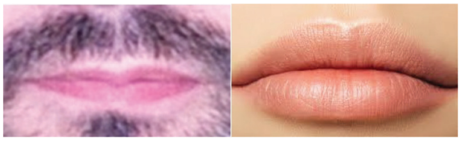

Figure 2. Samples showing some differences between male and female mouth

1.1. Related works

In the literature, the whole human face is the most used modality for facial data analysis and recognition. However, facial components and sub-regions were also used for related tasks. For the gender attribute from the oral region, there are few studies targeting it as a biometric modality (Darryl et al., 2013). In other studies, the oral region was used as a facial sub-region in geometric approaches (Afifi and Abdelhamed, 2019; B. Li, 2012) combining different intermediaries scores.

Darryl Stewart et al. (2013) studied automatic gender classification using static and dynamic features. After face and mouth localization, the authors followed a standard-DCT (discrete cosine transform) based process for feature extraction. Reasonable results were reported 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 Wu et al., 2012, a multi-view gender classification approach was proposed using facial component symmetry. The authors segmented the face according to 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. An accuracy of 82.23% was obtained for the mouth region.

Bing Li et al. (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-region scores were combined by using various strategies including sum, product, maximum, and majority voting rules. FERET and BCMI face datasets were exploited for experiments. Best classification rates obtained for the region (mouth) were of 82.9% and 83.6% for FERET and BCMI datasets, respectively.

Rai and Khanna (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 face dataset by keeping the lower part of the face via upper face blacking.

Afifi and Abdelhamid (2019) used five isolated facial components including the oral region (mouth) for deep gender classification. The authors used four convolutional 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. An accuracy of 89.09% was achieved for the mouth region.

Table 1 recapitulates the comparative analysis of the reviewed state-of-the-art literature, by making their principles and best gender classification rate. The oral region was also used to automatize other real-world classification problems. Mouth status estimation (Jie et al., 2016), person identification (Darryl et al., 2013), and lip-reading (Shrestha, 2018) are examples of artificial intelligence tasks already studied from the mouth.

Jie Cao et al. (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 overlapping inter train and test sets, and subject-independent (SI) to avoid data overlapping inter the train and test sets. Respective accuracies of 90.5% and 84.4% were achieved for both SD and SI experiments.

Table 1. State of the art analysis

Approach	**Year**	**Principle**	**Best score for the oral region**
T.X (Wu et al., 2012)	2012	SVM	82.8%
Darryl (Darryl et al., 2013)	2013	DCT	82.19%
Bing Li (B. Li, 2012)	2012	LBP+SVM	83.6%
Rai (Rai and Khanna, 2014)	2014	Gabor filters + PCA	85.3%
Afifi (Afifi and Abdelhamed, 2019)	2019	CNNs + Adaboost	89.05%

Karan Shrestha (2018) pointed to the complexity of lip reading. The author proposed the training of two deeper separate CNN architectures for real-time word prediction. The Haar classifier was used for face and mouth localization. Batch normalization was applied to speed up the training process and improve 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. (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 translation to text in videos frames. LipNet surpassed prior works and achieved new state-of-the-art accuracy of 95.2% in sentence-level.

In Carrie and Darryl (2020), LipAuth was proposed a system for lip-based authentication for mobile devices. The authors trained convolutional neural networks inspired by LipNet (Yannis et al., 2016). They used the XM2VTS dataset and collected new datasets, qFace and FAVLIPS. An equal error rate of 1.65% was obtained by benchemarrking the system on the XM2VTS.

2. Methodology

In this section, the proposed methodology is presented along with the applied convolutional neural network.

2.1. Overview

The overview of our approach is illustrated in Figure 3. At first we extracted the facial annotations corresponding to left and right eyes, nose, left and right commissure of the mouth. We computed the five facial annotations with the RetinaFace algorithm (Jiankang et al., 2019). Secondly, we performed the localization 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.

2.2. Convolutional neural network

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

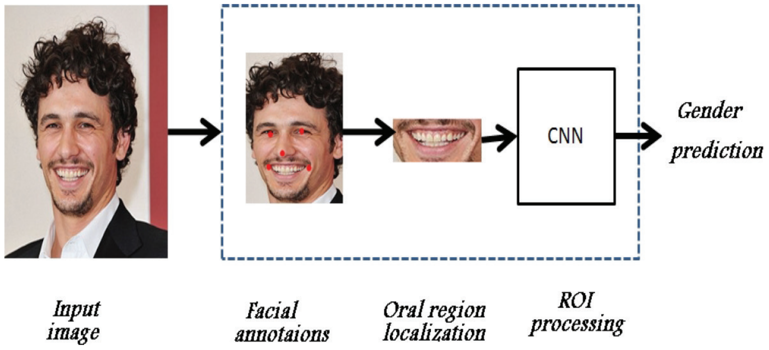

Figure 3. Overview of the proposed methodology

- Convolutional layer: It is a fundamental layer type in a CNN. At this stage, convolutional operator is applied on the input with the goal of learning features to enable detection in future subjects. A convolutional operator is mainly characterized by its kernel size.
- Pooling layer: In this type of layer, pooling operators are applied with the goal of reducing the data- dimensionality. They are generally used following convolutional layers, to generate maps of high resolution. Max-pooling and average-pooling are examples of the most used pooling methods.
- Batch normalization: It is an artificial intelligence technique that aims to speed up the training of very deep neural networks, to maintain stability during training phases.
- Fully connected layers: In a fully connected layer (FC), neurons are connected with all activation's in the previous layers. 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. A dropout layer simply works by forgetting a part of its input. Dropout layers are generally placed following FCs layers.

In our case, we adopted a sequential convolutional neural network by exploiting various layer 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 carry out the experimentation, we used the FFHQ (Flickr faces high quality) dataset. it is a recent real- world dataset collected for training the StyleGAN (Karras et al., 2019); an adversarial neural network generating realistic artificial fake corpus. FFHQ is a rich dataset in terms of facial variations,

Table 2. Details of the proposed CNN architecture

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 3. FFHQ dataset details

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

such as age, ethnicity, background, facial expressions, and occlusion. It contains 70K faces extracted from images acquired in the wild under high resolution. The dataset was collected from the well-known Flickr website (www.flickr.com/photos/) by selecting facial images under permissive licenses.

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

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 weights initialization, we firstly assigned 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 performed a real-time batch augmentation by using the horizontal flip.

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

Figure 4. FFHQ samples

Table 4. Used GPU memory characteristics

GPU size	GPU type	Specs
12 GO	Nvidia	Telsa K80

3.3. Evaluation metrics

To evaluate the trained network experimentally, we used various metrics based on the following parameters: TP (true positive), TN (true negative), FP (false positive) and FN (false negative). The 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 describes the rate of correctly classified subjects in the whole test-set. It is defined as follows:

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

The rate of true positive describes the rate of correctly classified intra-class subjects. 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 formula is:

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

The receiver operating characteristic is a curve that allows a graphical interpretation for the CNN comportment in front of test-set. It is addressed by displaying the evolution of FAR and FRR values during the test phase.

4. Results and discussion

In this section, the obtained experimental results for the proposed approach, using the cited metrics, are presented and discussed. Feature visualisation is performed and a baseline comparison with prior works is established.

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 the images correspond to the extracted oral region from frontal and semi-profile faces. The facial variations of the used sets are recapitulated in Table 6 in terms of facial expression, face pose, image quality, age, and ethnicity. Training and test set details are resumed in Table 5.

The obtained confusion matrix is shown in Table 7. Table 8 resumes the achieved accuracy (ACC) and RTPs (FRP and MRTP) describe the rate of correctly classified subject intra-class. An accuracy of 92.70% has been achieved for the whole test-set. For female and male genders, we got respective values of 94.00% ad 91.40% for the FTP and MTP parameters. It has been noted that the male gender easier to detect than the female gender. It can be justified by the fact that the upper-lip texture for the male gender, is easily detected by the trained convolutional neural network, especially in the case of subjects who have a mustache.

4.2. ROC-AUC, EER

Figure 5 shows the ROC curve traced for the tested network. Table 9 shows 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. Moreover, a quantitative value of 0.966 closer to 1 was obtained for the AUC parameter and a high area was covered. In

Table 5. Train and test size

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

Table 6. Train and test set's facial variations

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

Table 7. Confusion matrix

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

Table 8. Obtained ACC, FTP and MTP

	ACC	FTP	MTP
FFHQ	92.70%	91.40%	94.00%

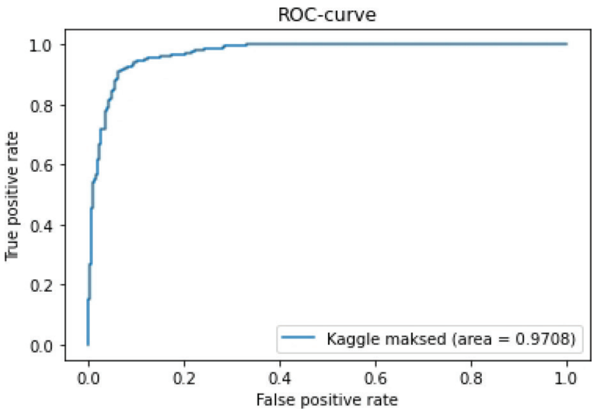

Figure 5. ROC curve

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

Parameter	Value
EER	0.0798
AUC	0.9708

Addition, a low EER error of 0.103 has been achieved. This allows to assert that the human gender can normally be predicted from the oral region.

Figure 6 shows samples of correctly classified and misclassified subjects from the test set. Some factors affecting the classification results have been noted 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 misclassified subjects are smiling. For the second factor of upper lip texture, it has been noted that almost all of the misclassified subjects from the male class are mustacheless, where the visual textures between both gender classes are close. For the last factor, which was

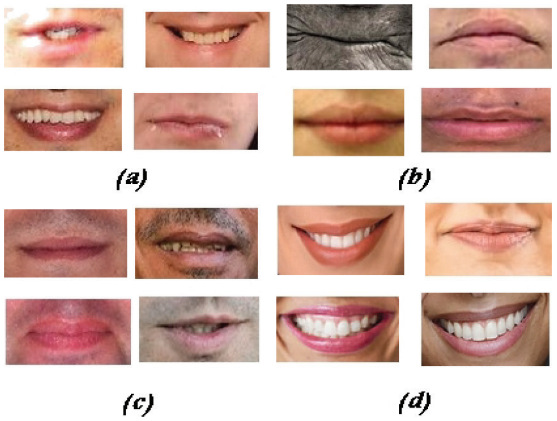

Figure 6. Example of correctly classified and misclassified subjects (a): misclassified males (b): misclassified females (c): correctly classified males (d): correctly classified females

age, it was found that elderly persons were misclassified and this can be justified by the fact that lip format and texture are lost with age. In addition, we observed that the face pose and ethnic variations do have any significant influence over the prediction rates.

4.3. Features visualisation

With the goal of detecting the most discriminative parts of the trained convolutional neural network in the task of gender classification, we used the Grad-Cam technique (Selvaraju et al., 2017). We computed activation maps by applying the Grad-Cam (Selvaraju et al., 2017) technique on the last convolutional operator of the network.

We computed the Grad-Cam class activation maps on several subjects from both female and male classes. Figure 7 shows samples of the obtained results. By looking at the obtained maps, it was found that lip texture and commissure were the parts of the oral region that strongly determined the networks' classification in the case of females. In turn, it was noted that upper and lower lip textures strongly determined the networks' classification in the case of males.

4.4. Baseline comparison

Finally, after evaluating the proposed approach, we made a baseline comparison with prior works using the oral region related part as a biometric modality or facial region. The comparative study presented in Table 10, shows the best-achieved accuracy in each research. It must be noted that in almost all of the existing works the authors experimented on a limited size test set.

As can be seen in Table 10, the gender classification rate obtained in our study surpassed those achieved in the literature. Indeed, we achieved a greater classification rate for the human gender attribute prediction from the oral region using a large, balanced test set of 1000 images derived from a real-world dataset.

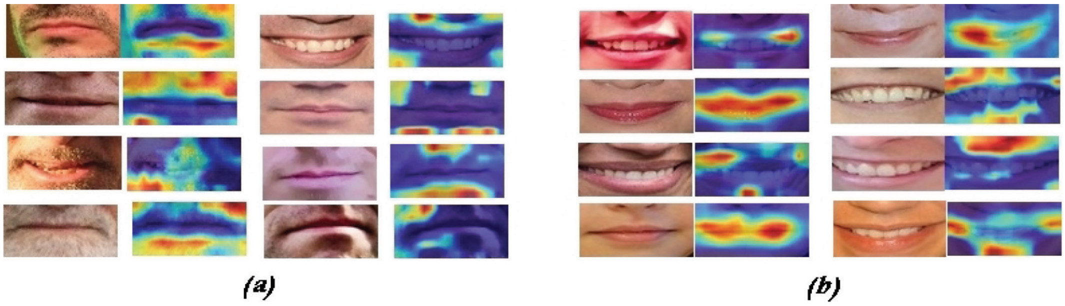

Figure 7. Example of features visualisation for both gender classes (a): male samples (b):female samples

Table 10. Baseline comparison

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

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, the RetinaFace algorithm has been used. For feature extraction and classification, we adopted a convolutional neural network (CNN). CNNs are a powerful tools for image processing and the present implementation is in line with this field of research. To perform our experimentation we used facial images of frontal regular and semi-profile from the real-world FFHQ dataset. The use of semi- profile images widens the field of application. In the present system not only frontal images can be used, but also face variations are incorporated from the beginning, allowing the user to pose in a more natural way.

To assess the effectiveness of our proposed approach, we exploited relevant metrics. Good results were achieved as a low EER was obtained and a 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. We also observed that from the oral region, male gender is easier to detect than the female gender. In relation to the oral features that strongly determined the CNN's classification of gender, lip textures and commissure were important in the female class while upper lip and lower lip textures were more important in the male class. Finally, we evidenced the robustness of the proposed approach by performing a comparative study of prior existing solutions for gender classification from the oral region.

Finally, considering the results of this study, future lines of research should focus on:

- The preparation of a facial dataset related to upper-occlusions with the goal of comparing, experimentally, the results of the proposed approach with the scenario of using the whole human face under upper occlusions.
- The adaptation of the approach for automatic upper-face occlusion detection by integrating a pre-processing step based on deep learning techniques with the goal of enhancing results.
- The generalisation of the approach for the investigation of others categorisation tasks from the oral region, such as: age estimation, race classification and prediction of facial expression.

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, L., Zhang, J., Humaidi, A., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M., Al-Amidie, M., and Laith, F., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *J Big Data*. Springer.

Carrie, W., and Darryl, W. S., 2020. Understanding visual lip-based biometric authentication for mobile devices. In *EURASIP J. on Info. Security*.

Choraś , M., 2010. The lip as a biometric. Pattern Anal Applic 13, 105–112. https://doi.org/10.1007/s10044-008-0144-8.

Darryl, S., Adrian, P., and Jianguo, Z., 2013. Gender classification via lips: static and dynamic features. In *IET Biometrics*.

Jiankang, D., Jia, G., Yuxiang, Z., Jinke, Y., Irene, K., and Stefanos, Z., 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. https://doi.org/10.48550/arXiv.1905.06641.

Jie, C., Haiqing, L., Zhenan, S., and Ran, H., 2016. Accurate mouth state estimation via convolutional neural networks. In *IEEE International Conference on Digital Signal Processing (DSP)*. IEEE.

Karras, T., Laine, S., and Aila, T., 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410. IEEE/CVF.

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

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

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE.

Shrestha, K., 2018. Lip Reading using Neural Network and Deep learning. https://doi.org/10.48550/arXiv.1802.05521.

Tapia, J., and Carlos, A., 2017. Gender Classification from NIR Iris Images Using Deep Learning. In *Deep Learning for Biometrics*. Springer.

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

Viedma, I., Tapia, J., Iturriaga, A., and Busch, C., 2019. Relevant features for gender classification in NIR periocular images. In *IET Biom.*, page 340–350.

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

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

Yannis, M. A., Brendan, S., Shimon, W., and Nando, d. F., 2016. LipNet: End-to-End Sentence-level Lipreading. https://doi.org/10.48550/arXiv.1611.01599.