A Convolutional Network-based Applied Approach to Frontal Facial Recognition and Solutions for Occluded Images

Iago Belarmino Lucena [*] **Lucas de Oliveira Santos** [*]
Matheus Araujo dos Santos [*] **Adriell Gomes Marques** [*]
João Matheus Lima Lucio [*] **Luís Fabrício de Freitas Souza** [**]
Paulo Antônio Leal Rego [***] **Pedro Pedrosa Rebouças Filho** [*]

[*] *Instituto Federal de Educação, Ciência e Tecnologia do Ceará (e-mails: iagolucena@lapisco.ifce.edu.br,lucas.santos@lapisco.ifce.edu.br, matheus.santos@lapisco.ifce.edu.br,adriell.gomes@lapisco.ifce.edu.br, joaomatheusllucio@lapisco.ifce.edu.br, pedrosarf@ifce.edu.br)*
[**] *Programa de Pós-Graduação em Engenharia de Teleinformática, Universidade Federal do Ceará (e-mail: fabricio.freitas@lapisco.ifce.edu.br)*
[***] *Departamento de Computação, Universidade Federal do Ceará (e-mail: pauloalr@ufc.br)*

Abstract:
Facial recognition technology is constantly being used in the most diverse sectors, from facial expression analysis to measure customer satisfaction in stores to a police instrument for identifying people. To summarize, from an image obtained by a camera, the technique identifies faces contained in that photo and compares them with a database of faces previously registered in the system. Based on advances in terms of algorithms and processing of *hardware* obtained in recent years, it was possible to provide solutions for verification and facial recognition through modern cell phones. The present work addresses the problem of facial recognition analysis that has the potential to scale for applications based on server processing. Using Histogram of Oriented Gradients to extract face features and the FaceNet Convolutional Neural Network this study brings results in different public (Labeled Face in the Wild and CelebFaces) and private databases. This study obtained satisfactory results with an accuracy of 90% in the best cases for private databases. As part of the work, this study also sought to evaluate the effect of partial occlusion on faces from the use of face masks because of the Covid-19 pandemic scenario, obtaining satisfactory results above 80%.

Keywords: face recognition; face detection; masked face detection; facenet; convolutional network;

1. INTRODUCTION

The availability of computers with high processing capacity and low cost, in addition to the emergence of embedded systems, contributed to the increased interest in various applications of digital image and video processing (Li et al., 2019). Among the applications on the rise, facial recognition has been a requirement for many commercial systems for a wide range of applications: law enforcement, residential and private security, banking, travel, education, etc. (Madhavan and Kumar, 2019; Kortli et al., 2020; Drozdowski et al., 2020) .

Facial recognition can be defined as a biometric identification technique that uses morphological and anthropometric information extracted from the human face in order to establish the degree of similarity between different facial images (da Silva Junior, 2020).

Although studies on facial recognition have been conducted since 1960, it is still a task with some challenges not yet overcome, mainly related to performance improvements, such as better time and greater accuracy in detection (Selitskaya et al., 2020).

The primary approach is divided into two basic steps: face detection and recognition. Face detection is a fundamental step because if the face is not isolated from the image background, image attributes extraction and face tracking are not possible (Wilson and Fernandez, 2006). Furthermore, detecting faces is a much more complex problem than detecting a static object since the face is a dynamic object that has different shapes and colors (Muller et al., 2004).

After detection, the extraction of features becomes necessary for the classification and recognition of the person to be performed. In this sense, the identification of face

specific points is fundamental, being used not only to identify of these points but also for the face tracking and position estimation (Zhu and Ramanan, 2012). However, problems in recognizing the essential face components for the its description as well as in predicting the correct face shape can make it difficult to calculate these (Kazemi and Sullivan, 2014) points.

Facial recognition still suffers from problems involving ambient lighting and the variation of the positions of the faces, as well as its occlusion by objects of everyday use such as glasses, caps, and scarves (Mi et al., 2019; Schroff et al., 2015). It becomes more critical in the context of surveillance, forensic analysis, and photo identification is a task whose accuracy is vital. Often, limitations in the overall quality of images reduce the ability to make decisions about a person's identity (Suma et al., 2019). As noted, although many advances have been made to solve the problem of facial recognition, many challenges remain open and need attention to create a system that can meet the needs of security and personal verification.

In this context, *matching* on faces is a technique for facial recognition that compares two images and analyzes similarities between some face regions under a predefined metric. For example, assuming two face images as matrices or vectors, it is possible to measure the similarity between them once the larger the corresponding *matching* areas, the smaller the distances of the dimensional spaces of the compared images could be. These techniques are implemented to perform this comparison on different faces, and different environments with a flow of people or to search for an individual in the facial database from his photo (Vokhmintcev et al., 2016). Research on the use of the graphics processing unit (GPU) (Ouerhani et al., 2010; Li et al., 2012; Kong and Deng, 2010), in facial recognition applications show good results and provide a basis for this work.

This study proposes the problem of facial recognition;

- Face detection and recognition using CNN network.
- Detection and recognition of faces using face masks (Facial recognition in images with occlusion).
- Use of FaceNet network for facial recognition in different dataset bases.
- Use of HOG for frontal faces detection, by extraction face features (Dalal and Triggs, 2005) with and without a mask.

2. RELATED WORKS

Face detection is a contemporary challenge even with the advent of new techniques and technologies. The improvement of the face detection model results in faster and more effective systems, promoting reliability and security for applications that use this technology. Therefore, this section will discuss works that seek to improve facial detection using different current techniques.

Aiming at the development of a robust facial detection system, in their study Peixoto et al. (2020) proposed a new method which performs the adaptive compression of the attributes obtained from facial recognition systems to enable the integration of IoT systems and machine and deep learning techniques for facial recognition applications, reducing computational, network, and energy costs. Compression was applied to the extracted attributes through the FaceNet (Schroff et al., 2015) extractor, which generates in its output a vector of attributes of dimension 128, composed of discriminatory features of each face. Altogether, to evaluate the proposed method, the authors used four sets of data, obtaining significant results in the compression, with a reduction of 90% of the files concerning their original size.

There is also DeepID, which produces higher accuracy with a relatively small difference of 97.45% with a standard deviation of 0.26% (Sun et al., 2014). FaceNet, which had the highest accuracy of 99.63% (Schroff et al., 2015) was trained with a private dataset consisting of millions of images taken from social media (Amos et al., 2016). While Openface is a model developed from DeepFace and GoogleNet that has been redesigned with several changes that are trained using the CASIA dataset and FaceScrub (Amos et al., 2016).

Research on University Classroom Presence System studied CNN FaceNet and Support Vector Machine (SVM) for face detection. After obtaining the images, the results for face detection were extracted with the aid of FaceNet and then classified using the SVM. This study compares the performance of 3 deep learning model architectures for face detection; the models are FaceNet, VGG16 and Convolutional Neural Network (CNN). The best accuracy results were achieved with the FaceNet model, with an accuracy rate of 99.6%, just looking at the performance of the deep learning architecture model, lacking real-time recognition and threshold determination for face recognition (Nyein and Oo, 2019).

Finally, research on a real-time attendance system used facial recognition with the aid of deep learning techniques. The dataset used consisted of images of 28 students, with 10 photos each, constituting 280 images. Facial recognition used the Euclidean distance calculation with a limit lower than 0.6 to consider the hit; the accuracy obtained was 95%. With the deep learning architecture, the accuracy was excellent, achieved through the Haar Cascade (Kuang and Baul, 2020) detection method.

So, aware of the challenges arising from face detection, such as improving detection results, reducing facial data storage, and the occurrence of occlusion as a variable in the problem, the work will analyze the situation making use of FaceNet on different types of bases .

3. MATERIALS AND METHODS

3.1 Databases

Public Databases

Two public databases were selected: the *Labeled Face in The Wild*(LFW) database, containing 13,233 images of 5,749 people obtained from the internet, and the *CelebFacesA* database, containing 202,599 images of 10,117 people. Some examples of the bases are in Figure 1

Private database

We worked with a private database with around 7 million samples. The base photo patterns are illustrated in Figure

(a) LFW

(b) CelebFaces

Figure 1. Public databases used on this work

2, from photos generated by the authors themselves due to confidentiality. About 4.9 million pictures are in color and have a resolution similar to the first photo in Figure 2, with a minimum resolution of 1 megapixel. The rest has characteristics identical to those of the 3 remaining images. Four partitions of the original base were then assembled to evaluate the problem of facial recognition in the face of different levels of photo quality. The method's accuracy will be calculated based on the characteristics of the targets.

Figure 2. Illustration of different photos from the private base.

Table 1. Characteristics of the assembled datasets and their respective amount of images.

dataset	**Sample to be searched Feature**	**Aim Feature**	**Quantity**
V1.1	colorful	colorful	12.475
V2.1	black and white	colorful	11.676
V3.1	colorful	black and white	15.963
V-mask	colorful	colorful	11.954

3.2 Deep Learning Network - FaceNet

FaceNet uses a trained convolutional network to optimize feature vectors instead of using intermediate layers to do the same task as previous implementations. This network generates an attribute vector with only 128 features, having a low computational cost compared to other deep learning networks with thousands of attributes. Figure 3 presents the architecture used by FaceNet.

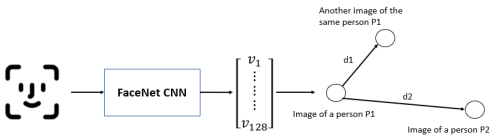

Figure 3. The structure used by FaceNet. After passing the set of images through the network architecture, the attributes pass through an L2 normalization layer to form the vector of 128 features. The cost function (*Triplet Loss Function*) is then used to update the weights of the network, based on the logic that images of the same person must have smaller distances than images of different people

FaceNet is trained so that the Euclidean distance directly represents the similarity between the faces; that is, images of the same face have attribute vectors with small distances, and images of different faces have greater distances. This optimization is part of the cost function (*Triplet Loss Function*) used to update the weights of the network. The authors evaluated the model developed through face verification, which compares two faces to define whether the faces belong to the same people or different people. The metric is defined by the validation rate *VAL(d)*:

$$VAL(d) = \frac{|TA(d)|}{|P_{same}|}, \tag{1}$$

P_{same} are face pair which belong to the same person and the set of all true accepts $TA(d)$ is defined by

$$TA(d) = (i,j)\ \epsilon\ P_{same}\ with\ D(x_i,x_j) \leq d, \tag{2}$$

on that (i,j) is the face pair to be compared, $D(x_i,x_j)$ is the L2 distance between the 128 characteristics vectors extracted by FaceNet and d is a threshold set by the authors.

One of the important results presented by the authors is the result obtained with different sizes of face images. This result is shown in Table 2. The results obtained with images with up to 120x120 pixels present results close to and VAL above 80%, and images with 40x40 pixels presented VAL of 37.8%, well below the other values, not being a good option to perform facial recognition.

Table 2. Effect of image size to the validation rate

Pixel	VAL(%)
1600	37,8
6400	79,5
14400	84,5
25600	85,7
65536	86,4

The method still reached 99.63% accuracy on the LFW Huang et al. (2008) base, surpassing the results obtained by DeepID2+.

3.3 Histogram of Oriented Gradients - HOG

The method of extracting features from HOG images (*Histogram of Oriented Gradients*) was proposed by Dalal and Triggs (2005) whose objective was pedestrian detection. In practice, this method works by extracting information regarding the orientation of edges contained in an image, and these edges are calculated using edge detection methods such as Sobel(Han et al., 2020).

The method takes an image as input and returns a n dimensional feature vector. The method is based on the distribution (histograms) of gradient directions (oriented gradients). Gradients (derived at x and y) of an image are useful because the magnitude of gradients is large around edges and corners (regions with sharp changes in intensity), and edges and corners are known to contain much more information about the shape of the object than flat regions.

Assuming that the images are grayscale, the method is divided into four phases: calculation of the orientation and magnitude of edges in the photo; division of the image into blocks and cells; analysis of the orientation histogram of the gradients by cells, later grouped in blocks; finally, the concatenation of these histograms, thus forming the descriptor vector HOG (Cruz et al., 2012).

Figure 4 presents the image resulting from the use of the HOG attribute extraction algorithm. To find the HOG descriptor, it is initially necessary to calculate the horizontal and vertical gradients to obtain the gradient histogram on which the Sobel filter can be applied. With this, it is possible to calculate the orientation and magnitude of the edges of the image in shades of gray. The gradient image (image on the right in Figure 4) removed a lot of excess information, leaving highlighted contours.

Figure 4. The image on the right is the result after using HOG.

Subsequently, two structures are defined called cells and blocks. At each pixel, the gradient has a magnitude and a direction. The gradient histogram is calculated over blocks of dimensions $v \times v$ that group together several cells to provide a more compact and noise-robust representation.

The next step is to create a histogram of gradients over each $v \times v$ block. In order to remove the light varition effect, it necessary to normalize the histogram, making the descriptor less invariant to lighting and shadows. Each block can now be represented by a histogram of the gradient orientation of each cell.

3.4 Metrics

Euclidean Distance

The Euclidean distance, shown in Equation 3 is defined as a measure of similarity, a technique generally linked to computer vision algorithms because it is better applied to non-standard data, such as samples without scale adaptation treatment; in this way, it can be said that its final result is insensitive to *outliers* (Cardarilli et al., 2019). As a disadvantage, this similarity measure is usually not recommended when there is a difference in scale or magnitude between dimensions. The distance equation for two points A and B in n-dimensional space can be seen in the equation. For analysis at work, vectors whose distance is less than 0.6 are equal or represent the same individual.

$$d_{AB} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{3}$$

4. METHODOLOGY

The evaluation method detects faces contained in the bases, uses the FaceNet network to obtain vectors with dimensions reduced to 128x1, and calculates the Euclidean distance.

4.1 Feature Extraction

Feature extraction takes into account relevant facial points that are identified through the HOG method, using the *dlib* library. Thus, the vectorization step can be exemplified by Figure 5.

An application is proposed to capture and process frontal faces, looking for the face in a pre-registered database. After verification, a list of possible candidates is displayed on the base, with a ranking score.

4.2 Obtaining masked faces

Due to the Covid-19 pandemic, several governments have adopted recommendations and impositions of new habits, especially regarding health measures. One of the most popular is precisely the use of face masks to avoid contagion of the disease. As a result, the challenges of developing facial recognition systems have increased. The occlusion effect caused by masks will be added to the face recognition problem on public and private bases. This is done through the technique *MaskTheFace* described in (Anwar and Raychowdhury, 2020) and exemplified in Figure 6. From face detection with the HOG method, the geometric coordinates of the faces were marked to explain which parts must be covered by artificial masks.

5. RESULTS

This section is subdivided into two stages: the first stage consists of detecting faces and recognition of frontal faces. The second stage is based on facial recognition using masks on the same faces (occlusion type).

Summary of used bases

Table 3. Accuracy results were obtained by the KNN method with datasets V1.1, V2.1, V3.1, and V-mask using the dataset with 7M samples as a target.

Recognition method	Consult dataset	Top-1	Top-2	Top-5	Top-10	Top-20	Top-50	Top-100
KNN	V1.1	0.61355	0.91230	0.93162	0.94701	0.95864	0.97154	0.98076
	V2.1	0.18979	0.30892	0.35055	0.41050	0.46617	0.54128	0.60192
	V3.1	0.10675	0.36785	0.41678	0.47554	0.53079	0.60609	0.65646
	V-mask	0.01422	0.02819	0.03764	0.05530	0.07679	0.11059	0.14681

Table 4. Accuracy results were obtained by the KNN method with the V1.1, V2.1, and V-mask datasets using the dataset with 4.9M samples as the target.

Recognition method	Consult dataset	Top-1	Top-2	Top-5	Top-10	Top-20	Top-50	Top-100
KNN	V1.1	0.61756	0.91695	0.93587	0.95030	0.96160	0.97547	0.98317
	V2.1	0.27706	0.43722	0.48133	0.53717	0.59687	0.66855	0.72157
	V-mask	0.02242	0.04350	0.05822	0.08064	0.10473	0.15083	0.19634

Table 5. Accuracy results obtained by the KNN method from the search of *matches* of faces with masks in public databases.

Recognition method	Consult dataset	Top-1	Top-2	Top-5	Top-10	Top-20	Top-50	Top-100
KNN	CelebFaces	0.3837	0.4785	0.5907	0.6667	0.7356	0.8124	0.8614
	LFW	0.5257	0.6187	0.7208	0.7854	0.8381	0.8916	0.9153

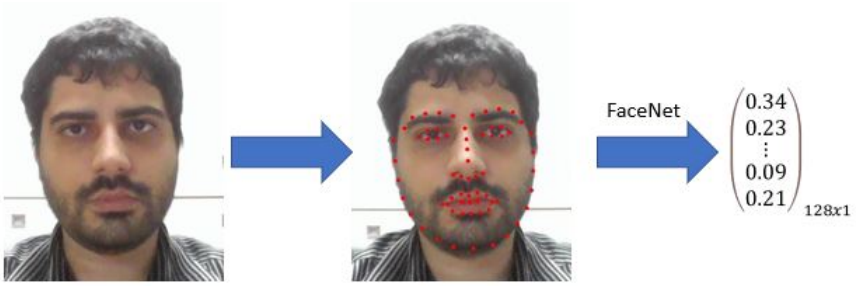

Figure 5. Vectorization process from facial point detection using HOG and FaceNet.

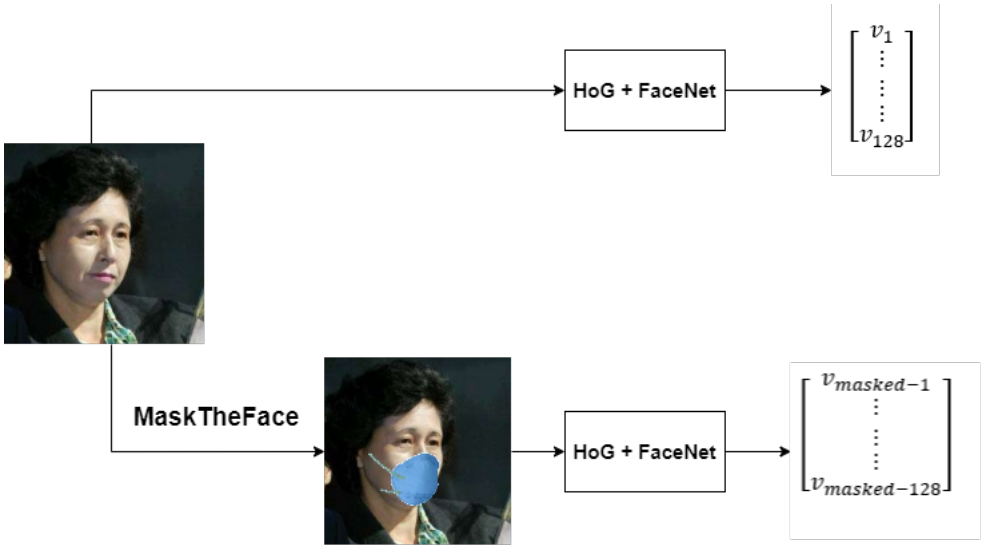

Figure 6. Embedding process from face landmarks detection by using HOG and FaceNet

- Dataset 1 with 7 million images in grayscale and RGB.
- Dataset 2 with 4.9 million images, except for the grayscale ones.
- LFW and CelebFaces bases with approximately 13 thousand and 202 thousand images, respectively.

5.1 Experiments with Dataset 1

From the reading of Table 3, it is possible to notice that the results are worse as the studied partitions have low-quality targets or query images, which is verified in the V2.1 databases. And V3.1. On the other hand, except for

the top-1 result, the V1.1 base presented accuracy above 90% with the use of FaceNet. The V-mask showed only 14.68% of accuracy in its best result.

5.2 Experiments with Dataset 2

Compared with the dataset 1 experiment, there is a slight increase in the accuracy percentages in all partitions. One of the factors is precisely that, by reducing the final amount of target base, the false positives that obey the established Euclidean distance threshold are also reduced.

5.3 Experiments with Public Databases

The same procedure was repeated with the LFW and CelebFaces bases to verify the FaceNet application with artificial masks faces case. An essential factor is that, unlike private databases, these public databases generally have several copies of frontal photos of the same individual. While there are cases of up to 20 different images for the same person, the private database rarely had more than three other images of the same face. Through the Table 5, it is possible to notice percentages of accuracy much higher than those resulting from the analyzes with the private base from the V-mask partition.

6. CONCLUSION

This study brought an innovative technique of detecting and recognizing human frontal faces through the use of CNN FaceNet and the HOG technique. The proposal also addressed the problem of occlusion in images using face masks.

The proposed model obtained 98.07% of accuracy on private bases, considering the case of search with images within the top-100 candidates for the V1.1 partition, with the best data quality image, and 61.35% in the case of the top-1. In all other analyses, the results are above 90%, demonstrating the robustness of the approach.

On the other hand, it is noticeable that the mixed quality of the query and target images for V2.1 and V3.1 affected the result, showing that the technique has lower performance for cases where image samples do not similar features and standards. This was because the superior quality in the capture was obtained from the general base and the difference in coloration, such as images in gray levels, image size captures, and different color images. For tests on images with synthetic masks on people, the use of FaceNet proved to be promising, obtaining different values of accuracy, with the best results being 86.14% and 91.53% of accuracy in the top-100 candidates for LFW and CelebFaces bases, respectively. An influencing factor that leads to results in this order is the number of face samples available per person, which is much higher than individuals from the private base.

For future work, the evaluation of the same problem by other forms of facial detection, such as specific convolutional networks, may influence the result after image vectorization by FaceNet. Also, it would be interesting to measure the effect of changing the threshold and how it could influence the number of true and false positives.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001". Also Pedro Pedrosa Rebouças Filho acknowledges the sponsorship from the Brazilian National Council for Research and Development (CNPq) via Grants Nos. 431709/2018-1 and 311973/2018-3.

REFERENCES

Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 20.

Anwar, A. and Raychowdhury, A. (2020). Masked face recognition for secure authentication.

Cardarilli, G.C., Di Nunzio, L., Fazzolari, R., Nannarelli, A., Re, M., and Spanò, S. (2019). *n*-dimensional approximation of euclidean distance. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(3), 565–569.

Cruz, J.E., Guimarães, L.N., and Shiguemori, E.H. (2012). Um estudo da detecção automática de campos de futebol deimagens aéreas e orbitais utilizando svm e descritores hog. In *XII Workshop de Computação Aplicada. São José dos Campos:[sn]*.

da Silva Junior, J.J. (2020). *Redes neurais profundas para reconhecimento facial no contexto de segurança pública*. Master's thesis, Universidade Federal de Goiás, Goiânia.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection.

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., and Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *arXiv preprint arXiv:2003.02488*.

Han, L., Tian, Y., and Qi, Q. (2020). Research on edge detection algorithm based on improved sobel operator. doi:10.1051/matecconf/202030903031. URL `https://doi.org/10.1051/matecconf/202030903031`.

Huang, G.B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1867–1874.

Kong, J. and Deng, Y. (2010). Gpu accelerated face detection. In *2010 International Conference on Intelligent Control and Information Processing*, 584–588. IEEE.

Kortli, Y., Jridi, M., Falou, A.A., and Atri, M. (2020). Face recognition systems: A survey. *Sensors*, 20(2), 342.

Kuang, W. and Baul, A. (2020). A real-time attendance system using deep-learning face recognition.

Li, D., Deng, L., Gupta, B.B., Wang, H., and Choi, C. (2019). A novel cnn based security guaranteed image watermarking generation scenario for smart city applications. *Information Sciences*, 479, 432–447.

Li, E., Wang, B., Yang, L., Peng, Y.t., Du, Y., Zhang, Y., and Chiu, Y.J. (2012). Gpu and cpu cooperative accelaration for face detection on modern processors. In *2012 IEEE International Conference on Multimedia and Expo*, 769–775. IEEE.

Madhavan, S. and Kumar, N. (2019). Incremental methods in face recognition: a survey. *Artificial Intelligence Review*, 1–51.

Mi, J.X., Luo, Z., Zhou, L.F., and Zhong, F. (2019). Bilateral structure based matrix regression classification for face recognition. *Neurocomputing*, 348, 107–119.

Muller, N., Magaia, L., and Herbst, B.M. (2004). Singular value decomposition, eigenfaces, and 3d reconstructions. *SIAM review*, 46(3), 518–545.

Nyein, T. and Oo, A.N. (2019). University classroom attendance system using facenet and support vector machine. In *2019 International conference on advanced information technologies (ICAIT)*, 171–176. IEEE.

Ouerhani, Y., Jridi, M., and Alfalou, A. (2010). Fast face recognition approach using a graphical processing unit "gpu". In *2010 IEEE International Conference on Imaging Systems and Techniques*, 80–84. IEEE.

Peixoto, S.A., Vasconcelos, F.F., Guimarães, M.T., Medeiros, A.G., Rego, P.A., Neto, A.V.L., de Albuquerque, V.H.C., and Reboucas Filho, P.P. (2020). A high-efficiency energy and storage approach for iot applications of facial recognition. *Image and Vision Computing*, 96, 103899.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Selitskaya, N., Sielicki, S., and Christou, N. (2020). Challenges in face recognition using machine learning algorithms: case of makeup and occlusions. In *Proceedings of SAI Intelligent Systems Conference*, 86–102. Springer.

Suma, R., Debattista, K., Watson, D., Blagrove, E., and Chalmers, A. (2019). Subjective evaluation of high dynamic range imaging for face matching. *IEEE Transactions on Emerging Topics in Computing*.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, 1988–1996.

Vokhmintcev, A., Sochenkov, I., Kuznetsov, V., and Tikhonkikh, D. (2016). Face recognition based on a matching algorithm with recursive calculation of oriented gradient histograms. In *Doklady Mathematics*, volume 93, 37–41. Springer.

Wilson, P.I. and Fernandez, J. (2006). Facial feature detection using haar classifiers. *Journal of Computing Sciences in Colleges*, 21(4), 127–133.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, 2879–2886. IEEE.