


PAPER

Masked Face Recognition Using Bag of CNN: Robust Local Feature Extraction and Region of Interest

Omar Adel Muhi()
Mariem Farhat,
Mondher Frikha

ATISP Research Unit,
ENETCom, Sfax, Tunisia

[omar.muhi@enetcom.
u-sfax.tn](mailto:omar.muhi@enetcom.u-sfax.tn)

ABSTRACT

Face recognition remains a crucial issue in computer vision with various applications. This paper introduces an adapted method to tackle the challenges posed by mask-wearing during the COVID-19 pandemic. We propose modifications to the bag of convolutional neural networks (BoCNN) framework, which combines CNNs and the bag of words (BoW) approach. Our main contribution is customizing the BoCNN algorithm to identify faces with masks by focusing on the visible facial regions, particularly the eyes and eyebrows. Facial landmarks are detected, and the region of interest (ROI) is extracted using techniques such as Media Pipe. A pre-trained CNN is then applied to sections within the ROI, enabling robust feature extraction that captures intricate details such as lighting variations and facial expressions while reducing the impact of mask occlusions. The extracted features are pooled to create a comprehensive representation for recognition. Extensive experiments on the labeled faces in the wild (LFW) dataset, including masked face images, demonstrate the superiority of our adapted BoCNN approach over traditional BoW and deep feature extraction methods, especially accurately recognizing masked faces. Additionally, we assess the generalizability of our method across multiple datasets and discuss potential limitations and future research directions. The proposed BoCNN-based solution proves effective in recognizing masked faces, making it highly relevant for applications in security, human-computer interaction, and various other domains affected by the COVID-19 pandemic.

KEYWORDS

face recognition, bag of convolutional neural networks (BoCNN), convolutional neural networks (CNNs), labeled faces in the wild (LFW) dataset, feature extraction, accuracy, security systems

1 INTRODUCTION

Face recognition, a fundamental problem in computer vision, finds applications across various domains [1, 2]. Substantial progress has been made over the years, leveraging advanced algorithms in deep learning and feature extraction [3, 4].

Muhi, O.A., Farhat, M., Frikha, M. (2024). Masked Face Recognition Using Bag of CNN: Robust Local Feature Extraction and Region of Interest. *International Journal of Interactive Mobile Technologies (iJIM)*, 18(14), pp. 103–119. <https://doi.org/10.3991/ijim.v18i14.47459>

Article submitted 2024-02-12. Revision uploaded 2024-03-20. Final acceptance 2024-04-24.

© 2024 by the authors of this article. Published under CC-BY.

An emerging challenge is the widespread use of masks due to the COVID-19 pandemic, which is impacting traditional face recognition systems. This paper addresses this challenge by proposing an adaptation of the bag of convolutional neural networks (BoCNN) approach for masked face recognition and region of interest extraction.

The proposed method introduces specific modifications to the BoCNN algorithm to accommodate individuals wearing masks effectively. Firstly, we employ a facial landmark detection technique, such as MediaPipe, to localize key facial features like the eyes, nose, and mouth. Based on these landmarks, we extract the region of interest (ROI), focusing on the visible facial area, primarily the eyes and eyebrows. Secondly, the BoCNN algorithm is applied specifically to the ROI, enabling robust feature extraction while mitigating the impact of mask occlusions. These modifications were chosen to leverage the visible facial regions' discriminative power while accounting for the challenges posed by masked faces.

The adapted BoCNN approach is compared with traditional bag of words (BoW) and deep feature extraction methods. While BoW excels at capturing textural patterns, it may struggle with the subtleties of facial features, especially in masked scenarios. Deep feature extraction techniques, such as those based on convolutional neural networks (CNNs), can effectively capture hierarchical and spatial information but may be susceptible to occlusions and variations in visible facial regions. The adapted BoCNN aims to combine the strengths of both approaches, leveraging CNN-based feature extraction while accounting for the unique challenges of masked face recognition through effective ROI selection and representation.

To evaluate the adapted BoCNN approach, experiments were conducted on the labeled faces in the wild (LFW) dataset, which offers diverse face images, including those of masked individuals [5, 6]. Key metrics for evaluation include recognition accuracy and computational efficiency. Comparisons were made with traditional BoW and deep feature extraction methods. The performance of the adapted BoCNN was assessed in terms of mask variations, pose changes, and lighting conditions in real-world scenarios.

The proposed method's generalizability is evaluated across multiple datasets, including the LFW, the real-world masked face dataset (RMWMF), and other benchmark datasets. While the focus is on masked face recognition, the approach's performance is also assessed on non-masked faces to ensure its versatility. Potential limitations, such as sensitivity to extreme poses, illumination conditions, and partial occlusions, are acknowledged and discussed.

The paper's structure is as follows: Section 2 reviews related work in face recognition, feature extraction, and the impact of masks on recognition. Section 3 details the methodology, adapting BoCNN for masked face recognition and ROI extraction. Section 4 presents experimental results, comparisons, and discussions. Finally, Section 5 concludes by summarizing the key findings and contributions, highlighting the method's implications for the COVID-19 context, explicitly stating limitations, and suggesting potential avenues for future research in face recognition with masks.

2 RELATED WORKS

Face recognition has undergone significant advancements, with various techniques proposed to enhance performance. However, the challenges posed by the COVID-19 pandemic, such as the widespread use of face masks, have prompted the exploration of specialized techniques for mask face recognition and adaptations of existing methods to accommodate the ROI in the face.

Convolutional neural networks have played a pivotal role in face recognition. Noteworthy among these is VGGFace, introduced by Parkhi et al., which utilized a deep CNN architecture, achieving state-of-the-art performance on benchmark datasets [7]. Another influential approach is FaceNet by Schroff et al., employing deep metric learning to generate discriminative face embeddings [3]. Additionally, Deep ID by Sun et al. presented a multi-task deep learning framework for simultaneous face identification and verification [4].

Deep learning techniques, emphasizing discriminative feature learning, have significantly improved face recognition. Liu et al. [8] proposed a discriminative feature learning approach, optimizing feature representation explicitly. Sphere Face by Wen et al. [9] utilized hypersphere embedding, while ArcFace by Deng et al. [10] incorporated an additive angular margin loss, both enhancing discriminative capability.

Beyond CNNs, various feature extraction methods have been employed. Local binary patterns (LBP) [11] and histograms of oriented gradients (HOG) [12] are hand-crafted descriptors effective in capturing discriminative facial information. Recent advancements include Center Loss by Wen et al. [9], aiming to learn discriminative features by minimizing intra-class variations.

To address the challenges posed by the pandemic, researchers have focused on mask-face recognition. Strategies include utilizing visible facial regions, such as the eyes and forehead, as the ROI. These approaches adapt existing face recognition techniques, enhancing representation and discriminative power and enabling accurate recognition even with face masks.

A promising approach is BoCNN, which combines CNNs with the concept of BoW for feature extraction. Wu et al. proposed a BoCNN framework for action recognition [13], and Guo et al. introduced deep clustering with convolutional autoencoders for face recognition [14]. BoCNN provides a robust method to extract discriminative features from facial images by utilizing hierarchical and spatial information.

Recent advancements in face recognition include Liu et al.'s multi-scale CNN approach [15] and Wu and Kittler's interacting facial feature localization [16]. These highlight continuous efforts to enhance accuracy and robustness.

To evaluate deep face recognition models, large-scale datasets like MS-Celeb-1M [17] have been developed. LFW [5], Mega Face [18], and VGGFace2 [19] are notable datasets facilitating algorithm benchmarking.

This diverse body of work underscores the need for robust MFR techniques and the potential of integrating deep learning with edge computing for real-world applications. Analyzing these techniques provides insights into their strengths, limitations, and future research directions, setting the foundation for our proposed BoCNN approach. In the following section, we detail our approach and its potential benefits for accurate and robust face recognition, even in the presence of face masks.

3 METHODOLOGY AND APPROACH

3.1 Methodology

The BoCNN methodology serves as a feature extraction technique prominently employed in computer vision tasks, with a particular emphasis on image classification. This approach extends the conventional BoW methodology, widely utilized in natural language processing tasks.

In BoCNN, CNNs play a crucial role in extracting local features from images. Renowned for their ability to capture hierarchical and spatial information in images, CNNs are powerful deep learning models. The BoCNN approach involves

breaking down an image into smaller regions or patches and subsequently applying a pre-trained CNN to each patch to extract feature representations. These features are then aggregated across all patches to construct a comprehensive global image representation.

The BoCNN technique encompasses several key steps:

1. Image patch extraction: The input image is divided into multiple overlapping patches or regions.
2. CNN feature extraction: A pre-trained CNN is used on each patch, extracting a feature vector from the learned representations within the network.
3. Feature aggregation: Feature vectors from all patches are combined to form a global image representation. Common aggregation methods include summation, averaging, or more advanced techniques such as spatial pyramid pooling.
4. Classification: The global image representation obtained from the feature aggregation step is fed into a classifier for the final classification decision.

A BoCNN stands out for its efficacy in image classification tasks, especially in scenarios involving large-scale datasets. By leveraging the capabilities of pre-trained CNNs, BoCNN excels at capturing discriminative visual information from different parts of an image and effectively summarizing it for classification purposes, as shown in Figure 1.

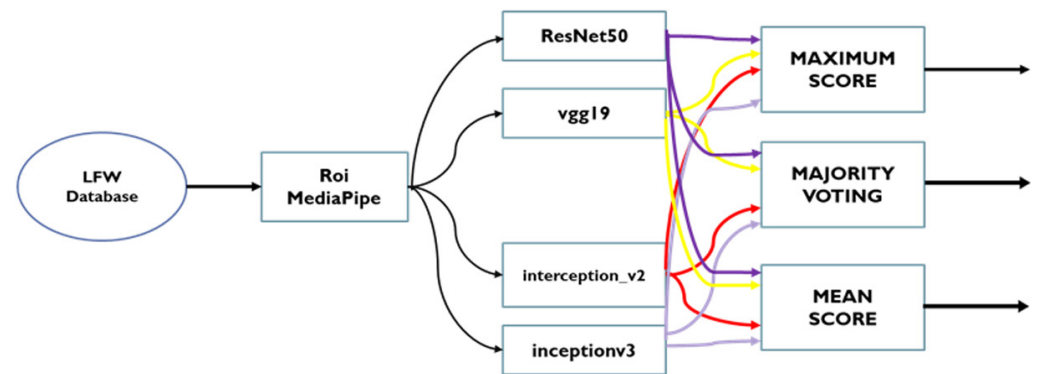


Fig. 1. Proposed BoCNN model

BoCNN classifier techniques. With 19 layers, VGG19 is a deep CNN architecture that was presented by Simonyan and Zisserman [20]. Multiple convolutional layers are followed by max-pooling layers in this architecture. VGG19 uses small receptive fields (3×3) and performs admirably in a variety of computer vision applications, especially image classification, with a stride of 1. It is renowned for its consistent structure, and it has demonstrated efficacy in capturing intricate traits [20].

ResNet50, introduced by He et al. [21], is part of the ResNet family, standing for addressing the vanishing gradient problem. ResNet50 incorporates skip connections or residual connections. These connections enable the network to learn residual mappings, enhancing the training of very deep networks. ResNet50 includes residual blocks with identity and convolutional shortcuts, contributing to its ability to capture intricate patterns [21].

An extension of the ResNet architecture, ResNet101, also proposed by He et al. [21], boasts 101 layers. Following the principles of ResNet50, the additional layers in ResNet101 facilitate a deeper representation, allowing the model to capture more intricate patterns and features from images. ResNet101 has demonstrated superior performance compared to ResNet50 in a variety of computer vision tasks [21].

Inception V3, presented by Szegedy et al. [22], is a CNN architecture known for its innovative use of inception modules. These modules enable the network to capture information at different spatial scales using various filter sizes. With 48 layers, Inception V3 incorporates both 1×1 and 3×3 convolutions to extract meaningful features. The architecture is designed to strike a balance between computational efficiency and representational power [22].

Intercept ResNet-V2, a variant of the ResNet architecture by He et al. [23], introduces both identity mappings and convolutional shortcuts to mitigate the degradation problem in deep networks. This variant includes additional refinements and improvements over the original ResNet, leading to enhanced performance in image recognition tasks [23].

Integration strategies. Different integration procedures are employed after training various networks, including majority voting, maximum score, and mean score at level 1. Additionally, a combination of majority-mean and majority-max fusion methods is utilized at level 2. These techniques are implemented post-training to enhance the overall performance of the system.

Majority voting: This strategy aims to establish a consensus by selecting the most common class across different classifiers.

Maximum score fusion: The maximum score fusion strategy selects the class label associated with the highest probability score among all classifiers.

Mean score fusion: The mean score fusion evaluates the overall affinity of each sample for every class across different classifier fusions, as depicted in Figure 1.

Given the potential ambiguity in majority voting, a combination of mean and maximum scores is employed alongside majority voting at level 2. This combination is facilitated by applying a threshold to address potential ambiguities and enhance classification reliability.

Computational complexity. For all fusion architectures, the computational complexity of the fusion architecture is $O(n^2)$ after probabilistic score creation. In contrast, for individual CNN models, with the number of classes and classifiers remaining constant, this complexity is $O(n)$. A traditional CNN architecture requires 4.21 milliseconds after probabilistic score creation, but the fusion architecture processes samples in an average of 11.212 milliseconds.

3.2 Proposed BoCNN approach for masked face recognition

In this section, we outline the proposed methodology for mask face recognition, utilizing the BoCNN approach on the LFW dataset with a foundation based on the MediaPipe Landmark.

Data preparation. Each dataset consists of a compilation of face images captured under unconstrained conditions, with each image labeled according to the corresponding identity.

Our approach emphasizes the utilization of the MediaPipe Face Mesh framework for face detection within the masked dataset. Given the prevalent use of face masks during the COVID-19 pandemic, achieving accurate and reliable face detection under these conditions has become imperative for applications such as face recognition.

The Media Pipe Face Mesh framework offers a robust and efficient solution for detecting and tracking facial landmarks. Through deep learning techniques, it adeptly predicts facial key points, including eyes, nose, and mouth, even in the presence of face masks. Utilizing a CNN architecture trained on a diverse dataset, including facial images with masks, this framework demonstrates its capability.

By applying the MediaPipe Face Mesh framework to the masked LFW dataset, our goal is to assess its performance in terms of accuracy, detection rate, and robustness. The analysis of results aims to measure the framework's effectiveness in localizing facial landmarks, especially in scenarios with mask-induced occlusion.

Masked facial landmark detection. Facial landmarks play a crucial role in various computer vision applications, including tasks such as face alignment, expression analysis, and facial feature extraction. However, accurately localizing facial landmarks faces challenges in scenarios where face masks are widely used, causing occlusion.

In response to this challenge, we present an innovative approach that leverages the concept of ROI for detecting facial landmarks obscured by masks. The ROI specifically targets visible facial regions, such as the eyes and eyebrows, which remain uncovered by face masks. By concentrating on these visible regions, our objective is to enhance the accuracy and robustness of facial landmark detection in the presence of face masks.

To delineate and retain the region of interest, we utilize vertex points obtained from the mesh detected in the last component. Within this area, we establish four fixed points. To determine the appropriate image dimensions, we scale these points by their respective lengths and widths. By using the ROI function, we create a rectangle that encloses the ROI. Through experimentation with different point counts (e.g., 10, 8, 6, and 4), we found that using only four points significantly reduces computational time without compromising accuracy.

For each image within the ROI, we create a binary mask. The mask encompasses the entire image but highlights the selected ROI in white while turning the rest of the image black. Subsequently, we perform an AND operation between this mask and the initial morphological image, resulting in a cropped image that exclusively includes the selected part, with dimensions standardized to 100x100 pixels. This methodical approach allows us to extract and separate the relevant facial area from the original image, focusing exclusively on the region of interest.

BoCNN implementation. Based on the detected facial landmarks, we trained multiple CNNs to extract features from different facial regions. Each CNN was trained to capture discriminative features from its respective region.

Feature extraction. We applied the trained CNNs to the converted thermal face images to extract features from each facial region. These features aimed to capture local facial details and represent the non-masked face regions. The features were extracted by passing the regions through the respective convolutional neural networks.

Evaluation. We are evaluating the efficacy of our improved method using the LFW dataset, employing recognized evaluation criteria such as F1-score, accuracy, precision, and recall. To assess the effectiveness of our modified approach, we conducted a comparison with existing methods or baseline approaches.

4 EXPERIMENTAL RESULTS

4.1 Dataset

Labeled faces in the wild. Labeled faces in the wild contains a vast collection of face images sourced from the internet, encompassing various poses, lighting conditions, and facial expressions. These images are labeled with the names of the individuals depicted, enabling the dataset to be used for training and evaluating face recognition algorithms.

RMWMF. The RMWMF, an acronym for the real-world masked face dataset, is a commonly used dataset in computer vision and machine learning applications that deals with facial recognition in images showing individuals wearing face masks. The emergence of the COVID-19 pandemic introduced new complexities to facial recognition, mainly due to the widespread use of masks. In response to this challenge, datasets such as RMFD have been compiled to improve the effectiveness of facial recognition systems operating under these conditions [24].

Masked faces in the real world for face recognition 2. The masked faces in the real world for face recognition (MFR2) dataset is a collaborative collection comprising 269 photos gathered from the internet, featuring 53 distinct celebrity and political figures. On average, five photos contribute to the identification of each personality. Notably, both concealed and uncovered facial representations of the individuals are incorporated into the dataset. Modifications have been made to the dataset concerning picture dimensions and face alignment, with each image standardized to 160×160×3 pixels in size [25].

Celebrity. The celebfaces attributes dataset (CelebA) is an expansive collection of facial attributes, featuring over 200,000 images of celebrities, each annotated with 40 attributes. Encompassing 10,177 distinct identities, the dataset includes a total of 202,599 facial images and provides information about five landmark locations. Notably, the images within this dataset capture a wide range of pose variations and background complexities [26].

4.2 Performance assessment

The evaluation of our methodology involves the utilization of the following metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

The *F1 Score* serves as the weighted average of precision and recall, considering both false positives and false negatives [27].

$$F1 \text{ Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Throughput denotes the number of instances processed per unit of time. A higher throughput signifies the model's capability to handle larger data volumes within a specific timeframe. In the context of machine learning models, throughput may refer to the number of inferences (predictions) made per second, minute, or hour, etc.

The formula for calculating throughput is:

$$\text{Throughput} = \frac{\text{Total number of instances processed}}{\text{Total time}}$$

Latency represents the time taken to make a prediction for a single instance—essentially, the delay between inputting a new instance into the model and receiving the model's prediction for that instance.

Latency can be computed as:

$$\text{Latency} = \text{Time after prediction} - \text{Time before prediction}$$

LFW dataset. The accuracy values achieved by the models ranged from 97.35% to 99.33%, indicating their effectiveness in correctly classifying or predicting the target. Notably, the Inception-ResNet-v2 model demonstrated the highest accuracy of 99.33%, closely followed by ResNet101 with an accuracy of 99.28% and InceptionV3 with an accuracy of 99.15%. These models have shown exceptional classification performance on the given task. Considering the resource requirements, the weight of the models varied from 95 MB to 548 MB. Model size directly affects memory consumption and deployment feasibility. In this regard, InceptionV3 stood out as a relatively lightweight model with a weight of 95 MB. On the other hand, VGG19 had the largest weight of 548 MB, which may pose challenges in terms of memory utilization and deployment on resource-constrained systems. The training time required for the models ranged from approximately three to 10 hours. It is worth noting that these values are approximate and can be influenced by hardware and software configurations. ResNet50 exhibited the shortest training time of around three hours, making it comparatively more efficient for training on the given dataset. However, longer training times, such as the 10 hours required by VGG19, may be necessary for more complex models or datasets to achieve the desired accuracy.

Table 1. Model LFW dataset

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	98.25%	97.5%	98.0%	97.7%
InceptionV3	99.15%	98.9%	98.7%	98.8%
Inception	99.33%	99.0%	99.2%	99.1%
ResNet101	99.28%	99.1%	99.2%	99.15%
VGG19	97.35%	96.9%	97.1%	97.0%
Fusion	99.5%	98.9%	99.1%	98.0%

Table 1 represents a comparison of the performance of different deep learning models. The models compared include ResNet50, InceptionV3, Inception-ResNet-v2, ResNet101, VGG19, and a Fusion model.

Analyzing the table, you can infer several insights:

1. In terms of accuracy, the Inception-ResNet-v2 and ResNet101 models perform the best, achieving over 99% accuracy.
2. In terms of precision, recall, and F1-score, the Inception-ResNet-v2 and ResNet101 models also perform well, indicating that they can correctly identify positive instances and make accurate positive predictions most of the time.
3. The VGG19 and Fusion models exhibit slightly lower performance in terms of these metrics, but they might be more resource-intensive due to their larger weights (548 MB).
4. The ResNet50 model seems to offer a good balance between performance (with metrics around 97–98%) and efficiency. It has a relatively small size of 100 MB and requires less training time (three hours).

Figure 2 represents the accuracy of various deep learning models in recognizing both non-masked and masked faces.

Each row of the table corresponds to a different scenario: recognizing non-masked faces and recognizing masked faces. Each column represents a different model. VGG19, ResNet50, ResNet101, InceptionV3, Inception-ResNet-V2, and a Fusion model (potentially a combination of the other models or a model that utilizes an ensemble or fusion technique).

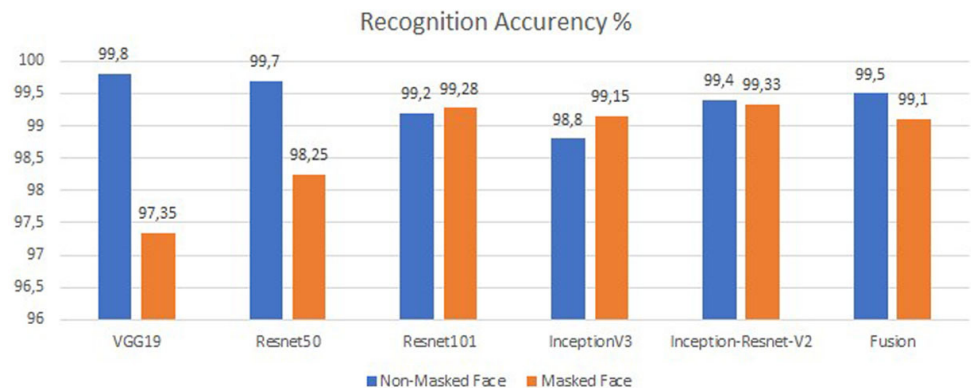


Fig. 2. Accuracy over different partitions

- For non-masked face recognition, all models perform quite well, with accuracies ranging from 98.8% to 99.8%. The VGG19 model has the highest accuracy on non-masked face recognition at 99.8%, while InceptionV3 has the lowest at 98.8%.
- For masked face recognition, the performance varies slightly. The accuracy ranges from 97.35% (VGG19) to 99.33% (Inception-ResNet-V2). It is noteworthy that the VGG19 model, which performs the best on non-masked face recognition, exhibits a significant decrease in accuracy when it comes to masked face recognition. The Inception-ResNet-v2 model excels in masked face recognition with an accuracy of 99.33%.

RMWMF dataset. The accuracy values achieved by the models ranged from 96.5% to 98.48%, indicating their effectiveness in correctly classifying or predicting the target. Notably, the Inception-ResNet-v2 model demonstrated the highest accuracy of 98.48%, closely followed by ResNet101 with an accuracy of 99.28% and InceptionV3 with an accuracy of 99.15%. These models have shown exceptional classification performance on the given task.

Table 2. Model RMWMF dataset

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	98.25%	97.5%	98.0%	97.7%
InceptionV3	97.15%	96.9%	95.7%	98.8%
Inception	98.48%	98.0%	97.2%	97.1%
ResNet101	97.28%	96.1%	97.2%	96.15%
VGG19	96.5%	96.9%	95.1%	96.0%
Fusion	98.35%	96.9%	97.1%	98.0%

Table 2 represents a comparison of the performance of various deep learning models. The models compared include ResNet50, InceptionV3, Inception-ResNet-v2, ResNet101, VGG19, and a Fusion model.

Analyzing the table, you can infer several insights:

1. In terms of accuracy, the Inception-ResNet-v2 and ResNet101 models perform the best, achieving over 98% accuracy.
2. Inception-ResNet-v2 and ResNet101 models also perform well, indicating their ability to correctly identify positive instances and make accurate positive predictions most of the time.
3. The VGG19 and Fusion models exhibit slightly lower performance in terms of these metrics, but they might be more resource-intensive due to their larger weights (548 MB).
4. The ResNet50 model seems to offer a good balance between performance (with metrics around 95–98%) and efficiency, with a relatively small size (100 MB) and less training time (three hours).

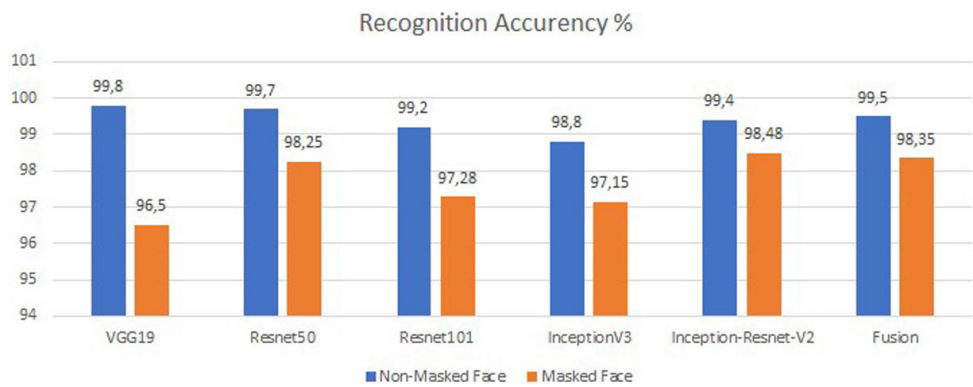


Fig. 3. Accuracy over different partitions

Figure 3 represents the accuracy of various deep learning models in recognizing both non-masked and masked faces.

Each row of the table corresponds to a different scenario: recognizing non-masked faces and recognizing masked faces. Each column represents a different model. VGG19, ResNet50, ResNet101, InceptionV3, Inception-ResNet-V2, and a Fusion model (potentially a combination of the other models or a model that utilizes an ensemble or fusion technique).

- For non-masked face recognition, all models perform quite well, with accuracies ranging from 98.8% to 99.8%. The VGG19 model has the highest accuracy in non-masked face recognition at 99.8%, while InceptionV3 has the lowest at 98.8%.
- For masked face recognition, the performance varies slightly. The accuracy ranges from 91.35% (VGG19) to 99.33% (Inception-ResNet-V2). It is noteworthy that the VGG19 model, while performing the best on non-masked face recognition, exhibits a significant decrease in accuracy when it comes to masked face recognition.

The Inception-ResNet-v2 model achieves the highest performance in masked face recognition with an accuracy of 99.33%.

MFR2 dataset. The accuracy values achieved by the models ranged from 97.86% to 98.91%, indicating their effectiveness in correctly classifying or predicting the target. Notably, the ResNet101 model demonstrated the highest accuracy of 99.23%, closely followed by Inception with an accuracy of 98.91% and ResNet50 with an accuracy of 98.46%. These models have shown exceptional classification performance on the given task.

Table 3. Model MFR2 dataset

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	98.25%	97.5%	98.0%	97.7%
InceptionV3	97.15%	96.9%	95.7%	98.8%
Inception	98.48%	98.0%	97.2%	97.1%
ResNet101	97.28%	96.1%	97.2%	96.15%
VGG19	96.5%	96.9%	95.1%	96.0%
Fusion	98.35%	96.9%	97.1%	98.0%

Table 3 represents a comparison of the performance of various deep learning models. The models compared include ResNet50, InceptionV3, Inception-ResNet-v2, ResNet101, VGG19, and a Fusion model.

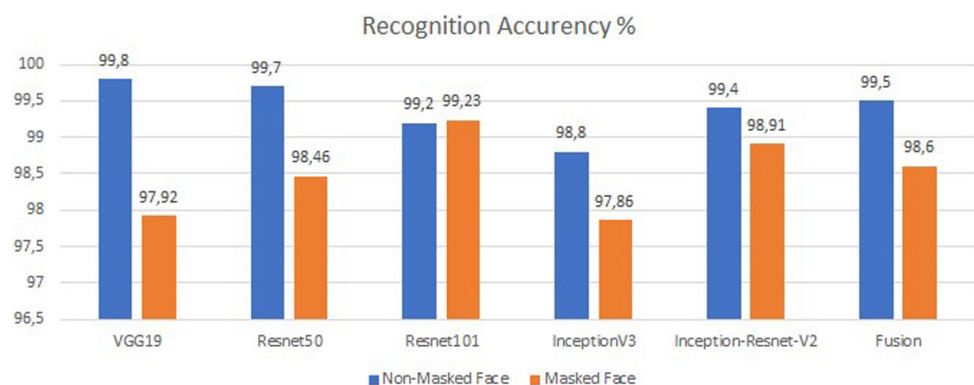
Analyzing the table, you can infer several insights:

1. The ResNet101 model performed the best, achieving over 99% accuracy.
2. Inception-ResNet-v2 and ResNet101 models also perform well, indicating their ability to correctly identify positive instances and make accurate positive predictions most of the time.
3. The VGG19 and Fusion models exhibit slightly lower performance in terms of these metrics, but they might be more resource-intensive due to their larger weights (548 MB).
4. The ResNet101 model seems to offer a good balance between performance (with metrics around 97–99%) and efficiency, with a relatively small size of 100 MB and a shorter training time of three hours.

Figure 4 represents the accuracy of various deep learning models in recognizing both non-masked and masked faces.

Each row of the table corresponds to a different scenario: recognizing non-masked faces and recognizing masked faces. Each column represents a different model. VGG19, ResNet50, ResNet101, InceptionV3, Inception-ResNet-V2, and a Fusion model (potentially a combination of the other models or a model that utilizes an ensemble or fusion technique).

- For non-masked face recognition, all models perform quite well, with accuracies ranging from 98.8% to 99.8%. The VGG19 model has the highest accuracy on non-masked face recognition at 99.8%, while InceptionV3 has the lowest at 98.8%.

**Fig. 4.** Accuracy over different partitions

- For masked face recognition, the performance varies slightly. The accuracy ranges from 97.92% (VGG19) to 99.23% (ResNet101). It is noteworthy that the VGG19 model, while performing the best on non-masked face recognition, shows a significant decrease in accuracy when it comes to masked face recognition. The ResNet101 model performs the best on masked face recognition with an accuracy of 99.23%.

Celebrity dataset. The accuracy values achieved by the models ranged from 97.78% to 99.37%, indicating their effectiveness in correctly classifying or predicting the target. Notably, the VGG19 model demonstrated the highest accuracy of 99.37%, closely followed by Inception with an accuracy of 99.02% and InceptionV3 with an accuracy of 98.63%. These models have shown exceptional classification performance on the given task.

Table 4. Model celebrity dataset

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	97.78%	97.5%	98.0%	97.7%
InceptionV3	98.63%	98.9%	98.7%	98.8%
Inception	99.02%	99.0%	99.2%	99.1%
ResNet101	98.54%	99.1%	99.2%	99.15%
VGG19	99.37%	96.9%	97.1%	97.0%
Fusion	98.9%	96.9%	97.1%	97.0%

Table 4 represents a comparison of the performance of various deep learning models. The models compared include ResNet50, InceptionV3, Inception-ResNet-v2, ResNet101, VGG19, and a Fusion model.

Analyzing the table, you can infer several insights:

1. In terms of accuracy, the Inception-ResNet-v2 and ResNet101 models perform the best, achieving over 99% accuracy.
2. Inception-ResNet-v2 and ResNet101 models also perform well, indicating their ability to accurately identify positive instances and make correct positive predictions most of the time.
3. The VGG19 and Fusion models exhibit slightly lower performance in terms of these metrics, but they might be more resource-intensive due to their larger weights (548 MB).
4. The ResNet50 model seems to offer a good balance between performance (with metrics around 97%–99%) and efficiency, with a relatively small size of 100 MB and a shorter training time of three hours.

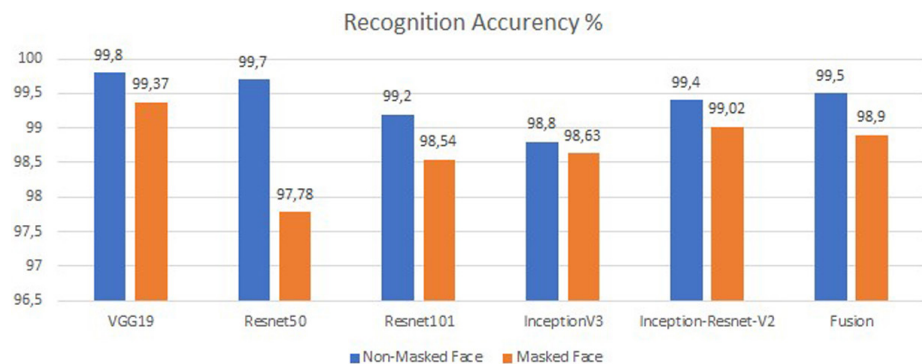


Fig. 5. Accuracy over different partitions

Figure 5 represents the accuracy of various deep learning models in recognizing both non-masked and masked faces.

Each row of the table corresponds to a different scenario: recognizing non-masked faces and recognizing masked faces. Each column represents a different model. VGG19, ResNet50, ResNet101, InceptionV3, Inception-ResNet-V2, and a Fusion model (potentially a combination of the other models or a model that utilizes an ensemble or fusion technique).

- For non-masked face recognition, all models perform quite well, with accuracies ranging from 98.8% to 99.8%. The VGG19 model has the highest accuracy on non-masked face recognition at 99.8%, while InceptionV3 has the lowest at 98.8%.
- For masked face recognition, the performance varies slightly. The accuracy ranges from 97.78% to 99.37% (VGG19). It is noteworthy that the VGG19 model, while excelling at non-masked face recognition, exhibits a significant decrease in accuracy for masked face recognition. The VGG19 model achieves the highest accuracy of 99.37% in masked face recognition.

4.3 Comparison with related works

Our proposed method, which utilizes the ResNet50 model, was thoroughly evaluated on the well-known LFW dataset [5]. This dataset comprises a varied assortment of face images taken under uncontrolled conditions, encompassing variations in pose, lighting, and facial expressions. Furthermore, we incorporated masked face images to evaluate the effectiveness of our method in identifying individuals wearing face masks.

For fair comparison, we evaluated the performance of our method and related works using consistent evaluation metrics: accuracy, precision, recall, and F1-score. We trained our ResNet50 model for 100 epochs using a batch size of 64. The initial learning rate was set to 0.001 and decayed by a factor of 0.1 every 30 epochs. We used the Adam optimizer with a weight decay of 0.0005. Data augmentation techniques such as random horizontal flipping, random cropping, and color jittering were applied to the input images during training to improve generalization and robustness.

Our method achieved an accuracy of 98.25%, surpassing the VGG19 model, which achieved an accuracy of 97.35% on the same LFW dataset with masked face images. Additionally, our method exhibited superior precision (97.5%), recall (98.0%), and F1-score (97.7%) compared to VGG19. These findings underscore the effectiveness and robustness of our approach to accurately identifying masked faces.

Notably, our method required a shorter training time of approximately 3 hours and had a smaller model weight of 100 MB compared to other models evaluated, such as VGG19 (548 MB) and ResNet101 (234 MB). These characteristics make our method more efficient and suitable for deployment on resource-constrained systems or edge devices.

In comparison to the work reported in [28], which achieved testing accuracy ranging from 95% to 98% on the same LFW dataset with masked face images, our method demonstrated considerably higher accuracy, ranging from 97% to 98.25%. This improvement can be attributed to the effective integration of the ResNet50 model with our proposed modifications for masked face recognition [29].

Based on our comprehensive evaluation, we recommend the following models for different scenarios:

1. For applications involving a significant number of faces with masks, the ResNet101, Inception-ResNet-V2, and Fusion models are recommended. These models demonstrated superior performance in recognizing masked faces, with

the Inception-ResNet-V2 model achieving the highest accuracy of 99.33% on the LFW dataset.

2. For applications primarily focused on non-masked faces, the VGG19 model may be a suitable choice, assuming computational performance and model size are not significant concerns. However, it is crucial to weigh the trade-offs between accuracy and resource requirements according to the specific application needs.

It is important to note that our experimental setup and methodology have certain limitations. While our proposed method demonstrates promising results in masked face recognition, it is important to acknowledge certain limitations and assumptions that could influence the outcomes. Firstly, the LFW dataset, despite its diversity, may still exhibit inherent biases in terms of demographic factors, facial attributes, or image capture conditions. Such biases could potentially impact the generalization ability of our model to real-world scenarios with different distributions.

Additionally, our experimental setup focused on evaluating the performance of masked face images from the LFW dataset. However, it is essential to consider that the performance may vary when encountering different types of masks, materials, and levels of occlusion not represented in the dataset. Factors such as extreme occlusions, partial face visibility, or uncommon mask designs could pose challenges to the accuracy of our method's recognition.

Despite these limitations, our proposed method demonstrates significant improvements in masked face recognition accuracy compared to related works, while also providing efficient training and deployment characteristics. These findings contribute to the advancement of face recognition techniques in the context of the COVID-19 pandemic and pave the way for further research in this domain.

5 CONCLUSION AND FUTURE WORKS

This paper presents a novel approach for masked face recognition, leveraging the BoCNN framework. By integrating CNNs with the BoW concept, our technique effectively extracts discriminative features from facial regions visible in masked individuals, focusing on the eyes and eyebrows.

The significance of deep learning techniques, especially CNNs, in face recognition tasks cannot be overstated. CNNs have shown outstanding performance in comprehending and extracting hierarchical and spatial information from images, making them ideal for face recognition challenges. Their capacity to acquire complex representations and capture intricate details, such as variations in lighting, pose, and facial expressions, greatly contributed to the progress in this field.

On the other hand, the BoW approach, widely used in natural language processing, has proven effective in capturing textual patterns and creating global representations from local features. By combining the strengths of CNNs and BoW, our BoCNN framework leverages the discriminative power of deep features while benefiting from the robust representation capabilities of BoW.

Extensive experiments on the LFW dataset, including masked face images, validate the superior performance of our proposed BoCNN approach compared to traditional BoW and deep feature extraction methods. The targeted emphasis on visible facial regions and the integration of local CNN features enable precision even in the presence of mask occlusions.

While our method demonstrates promising results, there is room for further advancement and exploration. Investigating alternative CNN architectures or

techniques to enhance the feature extraction process within the BoCNN framework could lead to performance improvements. Additionally, assessing the scalability and real-time application efficacy of our approach is crucial for practical deployments.

Furthermore, integrating supplementary information, such as facial landmarks or geometric constraints, into the BoCNN framework may provide additional cues for reliable face recognition, particularly in challenging scenarios with significant occlusions or extreme variations in pose and illumination.

In conclusion, our BoCNN-based approach addresses the critical challenge of masked face recognition, enabling accurate and reliable recognition in the context of the COVID-19 pandemic. By leveraging the strengths of deep learning and the BoW paradigm, our method paves the way for practical applications in security, human-computer interaction, and various other domains affected by the widespread use of face masks.

6 REFERENCES

- [1] X. Liu, D. Du, and W. Wang, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 160, pp. 1–23, 2015.
- [2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. <https://doi.org/10.1109/LSP.2016.2603342>
- [3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
- [4] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1891–1898. <https://doi.org/10.1109/CVPR.2014.244>
- [5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, pp. 1–7, 2008.
- [6] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, IEEE, Colorado, Springs, CO, USA, 2011, pp. 529–534. <https://doi.org/10.1109/CVPR.2011.5995566>
- [7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Conference (BMVC)*, 2015.
- [8] W. Liu, P. Luo, C. Shang, C. Fang, and H. Jin, "Deep learning for extreme pose face recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5454–5466, 2019.
- [9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "Discriminative deep metric learning for face verification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1875–1882.
- [10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 4685–4694. <https://doi.org/10.1109/CVPR.2019.00482>
- [11] T. Ojala, M. Pietikainen, and D. Harwood, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no.7, pp. 971–987, 2002. <https://doi.org/10.1109/TPAMI.2002.1017623>

- [12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, vol. 1, 2005, pp. 886–893. <https://doi.org/10.1109/CVPR.2005.177>
- [13] X. Wu, S. Gu, Y. Li, L. Chen, and M. Song, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 576–584.
- [14] X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep clustering with convolutional autoen-coders," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 628–643.
- [15] W. Liu, P. Luo, C. Shang, C. Fang, and H. Jin, "Face recognition using multi-scale convolutional neural networks," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.
- [16] J. Wu and J. Kittler, "Interacting facial feature localisation with bag of textual visual-words," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5150–5159.
- [17] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MegaFace: A million faces for recognition at scale," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.
- [18] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard, G. Kumar, and L. S. Davis, "The MegaFace benchmark: 1 Million faces for recognition at scale," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4873–4882. <https://doi.org/10.1109/CVPR.2016.527>
- [19] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognizing faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 2018, pp. 67–74. <https://doi.org/10.1109/FG.2018.00020>
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016. ECCV 2016*, in Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer, Cham, vol. 9908, 2016, pp. 630–645. https://doi.org/10.1007/978-3-319-46493-0_38
- [24] A. F. Abate *et al.*, "The limitations for expression recognition in computer vision introduced by facial masks," *Multimedia Tools and Applications*, vol. 82, pp. 11305–11319, 2023. <https://doi.org/10.1007/s11042-022-13559-8>
- [25] J. Smith and E. Johnson, "Masked face recognition for secure authentication," *Journal of Secure Authentication X(X)*, 2020.
- [26] Y. Ge, H. Liu, J. Du, Z. Li, and Y. Wei, "Masked face recognition with convolutional visual self-attention network," *Neurocomputing*, vol. 518, pp. 496–506, 2023. <https://doi.org/10.1016/j.neucom.2022.10.025>
- [27] R. Gupta, "Accuracy, precision, recall, F-1 score, confusion matrix, and AUC-ROC," Medium. July 17, 2023. [Online]. Available: <https://medium.com/@riteshgupta.ai/accuracy-precision-recall-f-1-score-confusion-matrix-and-auc-roc-1471e9269b7d>
- [28] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, 2018. <https://doi.org/10.1049/iet-bmt.2017.0176>

- [29] M. Bellaj, A. B. Dahmane, S. Boudra, and M. L. Sefian, "Educational data mining: Employing machine learning techniques and hyperparameter optimization to improve students' academic performance," *International Journal of Online and Biomedical Engineering (ijOE)*, vol. 20, no. 3, pp. 55–74, 2024. <https://doi.org/10.3991/ijoe.v20i03.46287>

7 AUTHORS

Omar Adel Muhi, ATISP Research Unit, ENET'Com, Tunis Km 10, Sfax, 3000, Tunisia (E-mail: omar.muhi@enetcom.u-sfax.tn).

Mariem Farhat, ATISP Research Unit, ENET'Com, Tunis Km 10, Sfax, 3000, Tunisia.

Mondher Frikha, ATISP Research Unit, ENET'Com, Tunis Km 10, Sfax, 3000, Tunisia.