M2FRED: Mobile Masked Face REcognition Through Periocular Dynamics Analysis

LUCIA CIMMINO, (Graduate Student Member, IEEE), MICHELE NAPPI, (Senior Member, IEEE), FABIO NARDUCCI, (Member, IEEE), AND CHIARA PERO, (Member, IEEE)
Department of Computer Science, University of Salerno, 84084 Salerno, Italy
Corresponding author: Lucia Cimmino (lcimmino@unisa.it)

This work was supported by the Project “PREdiction of Activities and Events by Vision in an Urban Environment (PREVUE)” funded through the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2017 under Grant 2017N2RK7K.

ABSTRACT Recent regulations to block the widespread transmission of COVID-19 disease among people impose the use of facial masks indoor and outdoor. Such restriction becomes critical in all those scenarios where access controls take benefit from biometric recognition systems. The occlusions due to the presence of a facial mask make a significant portion of human faces unavailable for feature extraction and analysis. This work explores the contribution of the solely periocular region of the face to achieve a robust recognition approach suitable for mobile devices. Rather than working on a static analysis of the facial features, like largely done by work on periocular recognition in the literature, the proposed study focuses the attention on the analysis of face dynamics so that the spatio-temporal features make the recogniser frame-independent and tolerant to user movements during the acquisition. To obtain a lightweight processing, which is compliant with limited computing power of mobile devices, the spatio-temporal representation of the periocular region has analysed and classified through Machine Learning approaches. The experimental discussion has been performed on a new dataset, Mobile Masked Face REcognition Database, specifically designed to analyse the periocular region dynamics in presence of facial masks. For a wider comparative analysis, a publicly available dataset called XM2VTS has been considered as well as Deep Learning solutions have been experimented to discuss the challenging aspects of the recognition problem. Moreover, a summary of the state-of-the-art on periocular recognition driven by COVID pandemic has been presented, showing how the research efforts in this field focused on recognition of still images. Experimental results show promising levels of performance as well as limitations of the proposed approach, creating the premises for future directions.

INDEX TERMS Periocular analysis, feature extraction, temporal feature representation, machine learning, biometric recognition.

I. INTRODUCTION
Over the past decade, the field of periocular-based biometrics has registered meaningful improvement. It has been proved that the periocular region represents one of the most discriminating and significant areas of a subject and, consequently, can be effectively used for biometric recognition purposes. Usually, the facial area in proximity of the eyes including eyebrows, eyelids, eye folds and eye corners represents the region of interest adopted in periocular biometrics, as shown in Figure 1. Several techniques have been proposed over the years to extract potential features in the upper region of the human face. In this direction, iris characteristics have been deeply investigated since 1987 [1], followed by the sclera [2] and retina [3]. Even if the iris has been confirmed to be robust and a reliable biometric trait, it suffers from recognition at-a-distance. Iris scanners can be very accurate but their usefulness in practical scenarios is limited due to the constraints...
of acquisitions and the high user cooperation required. These limitations lead to consider more feasible solutions and the analysis of periocular region is one of them [4]. The introduction of the periocular area as a new biometric trait took place in 2009 by the work of Park et al. [5] who shown the reliability of using just the periocular region of the face for biometric recognition purposes. Later, they also shown that the eyebrow was the most distinctive characteristic of an individual and some scenarios, such as variability in expression, pose and occlusion subjects’ faces, represented performance degradation factors [6], [7]. Since then, the state-of-the-art of periocular biometric has registered successes and confirmations in the reliability of such a portion of the face, developing identification and verification systems that adopt one or more biometric characteristics belonging to the periocular area [8], also integrating them into traditional face and iris biometric systems. Periocular region has also acquired important role in soft biometric classification [9] (i.e., age, gender, race, and ethnicity of an individual) and in the analysis of medically altered facial acquisitions (e.g., images acquired before and after surgery). This region requires low user cooperation, making it particularly suitable for forensics analysis, security applications, mobile authentication, and surveillance systems [10].

The presence of partial occlusions, on the other hand, plays a significant role that may negatively impact on recognition rate. The anti-COVID regulations adopted to contain the widespread diffusion of the pandemic, have significantly promoted the research on the analysis of the periocular region. However, in presence of occlusions due to facial masks, most of the biometric approaches based on face recognition have been reconsidered. While wearing a mask, a significant portion of the lower face is totally unavailable. So, inspired from the COVID-19 emergence, researchers were motivated to understand how biometric systems should analyse occluded faces and how periocular recognition can be beneficial for a reliable recognition. Furthermore, Computer Vision field and especially Artificial Intelligence based on Deep Learning approaches have shown recent advances in various scenarios to help dealing with the COVID-19 pandemic [11]. However, COVID-19 regulations adopted worldwide are posing the basis for new challenges in biometric recognition. In this work, we introduce a novel periocular biometric recognition system in presence of facial masks and based on Machine Learning approaches. The literature presents a wide collection of approaches and results for periocular analysis to biometric purposes. However, the systems proposed in the literature consider the periocular region like a static biometric trait. Geometric features and landmarks, as well as pixel appearance and their distribution are often considered in a static way. Even though the experimental results show high performance of several models on datasets built on this premise, the operative condition of face recognition in real scenarios suggests that those high level of performance is hard to be achieved. In fact, the facial expressions and the presence of several disturbing factors (e.g., eyeglasses, makeup or, recently, facial masks) may introduce insightful new dynamic patterns that can uniquely trace back to the identity of a single individual. In general terms, facial expressions are undoubtedly affecting the approaches of periocular recognition [12]. The impact can be potentially negative when the analysis is performed on a static representation of the facial features. The key point of our work is indeed exploring the dynamics of facial features and taking into consideration the involuntary micro/macro expressions that people perform while speaking. In this direction, working with video acquisitions may provide more reliable information but, more importantly, makes it possible to provide an expression-invariant solution that can deal with those (in)voluntary expressions affecting the recognition performance. To the best of our knowledge, this work is the first attempt of exploiting the dynamics of facial features rather than working on still images or single video frames of a video sequence. We tried to put together these challenges, by providing a new dataset of video recording of people wearing facial masks and assessing the level of performance that standard classifier can achieve exploiting/analysing the dynamics of periocular features versus the performance achieved by using still images of the same subject. The main contributions of this work can be summarised in the following:

- Realize a robust biometric recognition system, which is tolerant to noise or expressions of a single acquisition, so reducing the misclassification error.
- Combine the periocular region features over time to achieve a compact and temporal representation that can be used as input for Machine Learning classifiers.
- Introduce a new multimodal face dataset, namely Mobile Masked Face REcognition Database (M2FRED),\(^1\) to analyse the periocular region dynamics in presence of facial occlusions and, consequently, improve facial recognition techniques in case just the upper area of the face is available.

The rest of the paper is organised as follows: Section II discusses similar approaches in the literature that deal with periocular analysis, Section III presents the proposed approach while Section IV focuses on a wide experimental discussion.

\(^1\)http://biplab.unisa.it/M2FRED
that involves two datasets; Section V draws the Conclusions of this work.

II. RELATED WORKS
As briefly mentioned before, Park et al. [6] analysed the potential impact on recognition accuracy of the periocular region of the face with an accuracy of 87.32%. Although periocular region is defined as an independent distinctive biometric trait, over the years it has also been adopted as a complimentary modality to increase the accuracy of recognition rate of other ocular features that achieve large intra-class variations and in several multi-biometric systems (e.g., with iris [13] or face [14]). The key challenges in analyzing the periocular region of the face are accurate detection of facial feature segmentation and representation in order to produce a reliable recognition system. About the extraction of the features, several methods have been proposed including scale invariant feature transform (SIFT), histograms of oriented gradients (HOG) [15], colour histograms [16], local binary patterns (LBP) [17], local phase quantization (LPQ) [18], binarized statistical image features (BSIF) and maximum response sparse filter [19]. They are, at different level of complexity, compliant with the limited mobile devices capability thus enabling to obtain a feasible feature extraction. Following an even simpler approach, Miller et al. [20] have observed the benefit of working on the solely green channel of an RGB image to recognise an individual through the periocular region. The segmentation is crucial step in periocular analysis. In fact, it has been extensively analysed over the past decade. According to [21], in [22] the authors performed different experiments to separate and analyse different portions of the periocular region to distinguish those performing best in terms of recognition accuracy. With the advent of Deep Learning, and in particular of Neural Networks-based approaches, the extraction of optimal characteristics has consequently improved the performance of the periocular biometric systems. Most of these correlated studies have adopted the Convolutional Neural Networks (CNNs), particularly suitable for image classification [23], [24], [25]. Using advanced Deep Learning models, different researchers have used Artificial Intelligence-based preventive measures such as face mask detection and image-based computed tomography scans [26], [27]. Currently, due to the recent COVID-19 pandemic and, therefore, the need to wear a facial mask, many applications require the use of periocular biometrics. In [28], the authors evaluated a deep learning approach for eyebrow-based user authentication using a fine-tuned lightCNN model. Recently, Kumari et al. [29] developed a periocular biometric authentication system using hand-crafted and non-handcrafted features, assisted with semantic information. The proposed approach was evaluated on several well-known image databases in the literature. A variant of a Siamese Deep Learning network that allows to recognise a person with a facial mask was proposed in [30]. Fernandez et al. [31] demonstrated that only the periocular area can effectively estimate soft-biometrics indicators. He also proposed in [32] a multi-algorithmic fusion approach at the score level, evaluating periocular recognition algorithms when the images are captured with different sensors. Wang et al. [33] analysed existing public face recognition benchmark and altered them with simulated self-built masked faces to evaluate a recognition model based on facial and periocular features. In particular, the authors built three masked face databases, namely Real-world Masked Face Recognition Dataset (RMFRD), Simulated Masked Face Recognition Dataset (SMFRD) and Masked Face Detection Dataset (MFDD). Cabani et al. [34] propose three types of masked face datasets: the Correctly Masked Face Dataset (CMFD), the Incorrectly Masked Face Dataset (IMFD), and their combination (MaskedFace-Net) for detecting whether or not people are wearing face masks and ensuring that the masks are correctly applied. A deformable model is also presented, which generates masked face images by applying specific masks to facial images. A non-synthetic masked face dataset is provided by Huang et al. [35], called Real-world Masked Face Recognition. In addition to the non-synthetic dataset (RMFRD), the authors make two other databases publicly available: Masked Face Detection Dataset (MFDD), and Synthetic Masked Face Recognition Dataset (SMFRD) to train and test Deep Learning based masked face recognition models. Facial-masked datasets have been mainly introduced for the purpose of detecting facial masks. Mbunge et al. [36] present a useful and comprehensive review of artificial intelligence models that have been used to detect face masks. The body of research in the literature is on the production of masked face image datasets and algorithms for detecting whether a subject is wearing or not wearing a mask. In [37], the authors propose an advanced deep learning model for face mask detection in real-time video streams. In this regard, Negi et al. [38] employ two well-known deep neural network architectures with transfer learning for face mask detection using the Simulated Masked Face Recognition Dataset. Recently, the focus has shifted to subject recognition in the presence of masks, which is nowadays known as Masked Face Recognition (MFR). As an example, Li et al. [39] propose a new method for masked face recognition by integrating a cropping-based approach with the Convolutional Block Attention Module (CBAM). Optimal cropping is explored for each case while the CBAM module is adopted to focus on the region around the eyes. In this direction, Damer et al. [40] present a study on the impact of facial masks on the behaviour of renewed face recognition systems. Their findings revealed the expected observation that wearing a mask has a considerable negative impact on the recognition performance, thus highlighting the need of alternative approaches to traditional facial recognition algorithms to deal with the difficulties introduced by the presence of masks.

III. PROPOSED APPROACH
The proposed approach is based on the hypothesis that facial dynamics can be an added value to biometric face recognition. In fact, it can be seen as a kind of signature that identifies
an individual uniquely and unmistakably. The changes that occur on the face while a subject is speaking are captured by an RGB camera, and represented through a time series that varies from subject to subject. The idea is that the identity can be recognised by the dynamics of the subject’s face when he/she pronounces phrases that imply facial movements. Although it is a behavioural characteristic, when combined with facial recognition, it can be a valuable tool in situations where partial facial occlusions occur.

A. SYSTEM ARCHITECTURE
The overall workflow is depicted in Figure 2. The architecture is the same of a generic facial recognition system and includes two main modules: the enrolment and recognition module. In the enrolment module, the videos are acquired through a sensor and a set of features are extracted. When the data relating to the subjects of interest have been obtained, they will be stored in a database and used for the training of the Machine Learning model. The model will work as a benchmark in the recognition process between the captured subject and the predicted identity. The purpose in this work is to test the effective discriminating power of the feature dynamics, using two different feature extraction approaches and various Machine Learning Models, to identify the most effective combination in terms of recognition performance.

B. GEOMETRIC FEATURES
1) LANDMARK MODEL
The first feature extraction approach is based on the geometric features. It analyses the facial features and the geometric relationships that exists between them over time. A face detection algorithm identifies the geometric structure of the human face through a set of 68 points. A subset of these points (located in the upper part of the face) is taken into consideration: some of them are connected in pairs by a segment representing their distance. A total of 14 key landmarks composed by points and distances that effectively represent the periocular area of a subject are stored. The time factor is expressed by the frames analysed for the features extraction in a specific video sequence. The variation of the distances captured frame by frame produces a time series that represents the variation of the subject’s expression while he/she is speaking. Formally, a segment that connects two points represents a feature $f_i$. Each features have a length $l_i$ and a weight $w_i$, where $1 \leq i \leq K$ and $K$ is the number of features. Let

$$T_{s_i} = (l_{i1} w_1, l_{i2} w_1, \ldots, l_{ij} w_1, \ldots, l_{iN} w_1)$$ (1)

be the i-th time series, relating to the variation of a specific feature in all the N frames of the video. The final vector, called “Dynamic Facial Feature” (DEF), is obtained as the sequence of all time series:

$$DEF = (T_{s_1}, T_{s_2}, \ldots, T_{s_i}, \ldots, T_{s_k})$$ (2)

The product of the number of features and the number of frames $K \times N$ represents the size of the DEF vector. Finally, a normalisation procedure is applied to the DEF vector to make the data structurally homogeneous and ready for the classification.
C. APPEARANCE-BASED FEATURES

1) LBP-TOP
Local Binary Pattern LBP [17] represents one of the most widely used local descriptors, applied in various fields of computer vision. The basic idea is that image texture can be described with two complementary measures: the local patterns and the contrast of gray scale image. LBP operator uses the neighbours of a pixel to generate a binary number, a function of the central pixel. The resulting histogram can be used as a descriptor for the image. Each pixel is compared with its \( n \times n \) neighbourhood, replacing it with 1 or 0 based on neighbour pixels value; the histogram generated by concatenating the LBP code for each pixel is then used as a descriptor of image texture. LBP is chosen mainly for static texture description. In presence of dynamic video frames, it is possible to distinguish a variant of LBP, namely LBP-TOP (Local Binary Pattern from Three Orthogonal Planes) [41]. LBP-TOP considers three orthogonal planes: \( XY \), \( XT \), \( YT \) (see Figure 4), that indicate, respectively, the spatial information and the horizontal or vertical motion information.

A video sequence can be considered as a series of successive \( XY \) planes over time, which ignores the other two orthogonal planes. Instead, \( XT \) and \( YT \) provide information on space-time transition. LBP-TOP analyse the three orthogonal planes that intersect in the central one. So, it considers the features on each plane separately, and after concatenating them, LBP-TOP creates a spatio-temporal feature vector. Once the patterns have been calculated for each block, they are concatenated into a single histogram. Through this representation is possible to provide a facial expression descriptor along the three axes that effectively describes appearance and temporal information at pixel level. The advantages of LBP-TOP approach are: (i) the robustness respect to the monotonic grayscale image transformations, (ii) the possibility to obtain the spatial texture and temporal motion information of a pixel and (iii) its computational speed. The above-mentioned are three fundamental aspects to analyse and to model the dynamic structure of a video-frame sequence.

2) DEEP LEARNING APPROACH
Convolutional Neural Networks (CNNs or ConvNets) are a type of Artificial Neural Network that is commonly used to handle computer vision and image processing tasks. 3D CNN applies a three dimensional convolutional filter to the dataset and the filter moves in three directions \((x, y, z)\) to calculate the low level feature representations. Their output shape is a 3D volume space such as cube or cuboid. They are suitable for learning spatio-temporal features: helpful in event detection in videos [42], 3D medical images [43], etc. They are not limited to 3D space but can also be applied to 2D space inputs such as images.

D. CLASSIFICATION ALGORITHMS AND EVALUATION METRICS

The classification algorithms considered in this study are linear Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and an Artificial Neural Network sequential model. In this regard we refer to a Multilayer Perceptron (MLP) network. An MLP is a class of feed forward networks, in which each neuron accepts the input of the previous layer, sends it to the next and there is no feedback loop. The training method is based on the gradient descedent approach: iterative change of network weights that gradually reduces the error on the training set. The algorithm that implements this idea is the backpropagation.

To evaluate the performance of the system and effectively compare the different classification models applied in this work, accuracy, precision, and recall score was calculate in a first reference. Accuracy is defined as the ratio between the correctly predicted positive values (True Positive, TP) and the correctly predicted negative values (True Negatives, TN) for the total classifications made, including False Positives FP and False Negatives FN.

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)
\]

Precision and Recall are mathematically defined as follows.

\[
\text{Precision} = \frac{TP}{TP+FP} \quad (4)
\]
\[
\text{Recall} = \frac{TP}{TP+FN} = \frac{TP}{TP+FP+FN} \quad (5)
\]

Recall is also defined as the True Positive Rate. Additionally, the False Positive Rate is defined as follow:

\[
\text{FPR} = \frac{FP}{FP+TN} \quad (6)
\]

Plotting the TPR against FPR it is possible to obtain the ROC (Receiver Operating Characteristic) curves, an immediate and simply overview of the models performance. The Area Under this Curve (AUC score) makes us better understand how efficient our model is and represents probability that
a random sample is correctly classified. Another graphic method to understand the behaviour of a model is the Confusion Matrix, a matrix where each row represents the actual class and each column the predicted class, comparing the parameters of true positives and negatives with false positives and negatives [44].

IV. EXPERIMENTAL RESULTS

A. DATASETS

The experiments have performed on two different datasets: M2FRED and XM2VTS. For the XM2VTS database, the experiments were divided according to subjects’ number: a first set including all 294 subjects and a second group with 50 individuals. The divisions described above are aimed to evaluate the performances increase based on the reduction the total number of 1:1 matching among test probe and the gallery, but also to make a fairer comparison with the M2FRED dataset which contains 43 subjects in total. Details of the two datasets are shown in the following subsections.

1) M2FRED

The M2FRED dataset (Mobile Masked Face REcognition Database) was developed by Biometric and Image Processing Lab (BIPLab) research laboratory at University of Salerno. M2FRED is a multimodal frontal face database that includes digital videos of 43 subjects. This dataset aims to analyse the dynamics of the periocular region when, in presence of facial occlusions, biometric recognition is only possible from the contribution of the upper area of the face. Each subject was required to register using their mobile device following precise acquisition rules for several distinct sessions. The protocol provides that each individual is recorded in several distinct and timed sessions over a period of time not limited to a single day. The construction of the dataset is in fact faithful to the normal conditions of use of a recognition system in real conditions, which foresee a variability of the same subjects over time. The subjects were asked to face the camera, making sure that their face is entirely framed in the video stream of the mobile phone used for recording. Participants videos were acquired with short recordings (a few minutes), pronouncing suggested phrases or words according to the defined protocol. The acquisitions were divided into 4 sessions, in physical and environmental conditions not necessarily the same. The position of the camera relative to the face was the same.

Each participant has pronounced the following sentences:

1) “Zero, one, two three, four, five, six, seven, eight, nine”; 
2) “Good for evil is charity, evil for good is cruelty”; 
3) “Hello, my name is “participant’s name”, my favorite color is “participant’s preferred color””; 
4) Tell an anecdote or a joke.

In each session, a single recording of all the phrases above mentioned was processed and separated by a 3 second pause. Within the same acquisition, the subject first recovered without the mask and then with the facial mask. Each session took place on a separate day of the week, interspersed with indoor and outdoor acquisitions. Figure 6 shows some images sample extracted from M2FRED videos.

In order to emphasise the contribution of M2FRED, a comparison of the proposed dataset against existing masked face datasets currently available in the literature is presented in Table 1. It is worth noting that M2FRED is the only multimodal dataset, with videos of 43 subjects wearing different facial masks. Furthermore, it offers the highest number of acquisitions for the same participants with and without mask, indicating a significant intra-class variability as well as different environmental conditions for each acquisitions. A comparison of the performance achieved by other datasets is also reported: it is possible to observe that M2FRED shows competitive results in terms of accuracy.

2) XM2VTS

The XM2VTS is a multimodal face dataset including digital videos of 295 participants. The database was created by the University of Surrey, England, as part of the M2VTS project (Multi Modal Verification for Tele-assistance and Security services). It comprises four videos for each subject, recorded taken over a period of four months. The entire dataset was acquired with the same digital camera. The records are in .avi format, with a 725 × 576 resolution. Additional technical details can be found in [48].

Figure 7 illustrates some images captured from XM2VTS dataset videos.
TABLE 1. Brief description of masked face datasets currently available in the literature. Images number and resolution along with the publication year are reported for each dataset. Mask type indicates whether the facial mask is Real (R) or Simulated/Fake (S). The number of mask describes the different types of facial masks present in the dataset. Number of acquisitions of the same subject with (#W/) and without (#W/o) mask are also reported.

| Dataset      | Year | Resolution | #Subjects | #W mask | #W/o mask | #Mask type | Mask type | #Images | #Videos | Accuracy |
|--------------|------|------------|-----------|---------|-----------|------------|-----------|---------|---------|----------|
| M2FRED       | 2022 | variable   | 43        | 16      | 16        | various    | R         | 412,800 | 1376    | 96%      |
| MFD [33]     | 2020 | 256x256    | 12,000    | 2       | 171       | half       | 1 R       | 95,000  | -       | 85%      |
| RMD [33]     | 2020 | 256x256    | 525       | 9       | 2         | 4 S        | 64,973   | -       | 49%      |
| MFR [45]     | 2020 | 160x160    | 55        | 171     | 5         | 2 S        | 411.70   | -       | 95%      |
| LFW-SM [45]  | 2020 | 160x160    | 5,749     | 8       | 2         | 4 S        | 64,973   | -       | 49%      |
| LFW-D [33],[46]| 2021 | 403x403    | 400       | 5       | 5         | 2 S        | 411.70   | -       | 95%      |
| LMD [47]     | 2021 | variable   | 700       | 1       | 1         | 2-3 S      | 1,510    | -       | -        |

TABLE 2. Accuracy, Precision and Recall score for XM2VTS 294 subjects static and dynamic landmarks with maximum (61) number of frames.

| XM2VTS #294 | MLP | SVM | DT | RF |
|-------------|-----|-----|----|----|
| Accuracy    | dynamic | 0.856 | 0.650 | 0.192 | 0.609 |
|             | static  | -    | 0.454 | 0.157 | 0.339 |
| Precision   | dynamic | 0.734 | 0.716 | 0.235 | 0.688 |
|             | static  | -    | 0.509 | 0.190 | 0.359 |
| Recall      | dynamic | 0.856 | 0.650 | 0.192 | 0.609 |
|             | static  | -    | 0.454 | 0.157 | 0.339 |

B. GEOMETRIC FEATURES

The following experiments use the geometric features extracted with the landmark detection model described in the section III-B1 and compare different classification models: Multilayer Perceptron network (MLP), linear Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF). Each experimental configuration uses both dynamics and the static landmarks to evaluate the real contribution of facial dynamics to geometric features representation. The results of the MLP network are sometimes ignored and not discussed (see Tables 2, 3, 4). This is due to the limited size of the training set that lead to unfair comparative analysis. Static features are extracted in the same way as dynamic features, but just considering one frame for each video (the first with a valid face detection). Extracting static features allows to compare the periorcular recognition results in static vs spatio-temporal setting and to observe the gain of performance of the use of dynamics into the periorcular analysis. A first set of experiments were conducted on the XM2VTS dataset described in section IV-A2 on a total of 294 subjects. The results reported in Table 2 resumes the level of performances in dynamic and static setting.

From these first results, it is possible to see how the facial dynamics have a positive impact on the performance of each model: i.e., the improvement in accuracy of the SVM in dynamic settings is about +20% respect to the performance in static settings. Also the ROC curves in Figure 9 show a better overall trend for almost all the classification models. The Table 3 summarises the results of the experiments considering a subset of 50 subjects for XM2VTS dataset. This setting has been chosen so to make fairer the comparative analysis between the samples in XM2VTS and the M2FRED datasets.

In addition, the number of frames was tuned, considering a maximum number of frames (sample-61) corresponding to a time window size equal to 61 consecutive video frames and a minimum number of frames (sample-10) corresponding to 10-frame time window size (such a double setting is not presented in Table 2 for XM2VTS with 294 subjects due to the insufficient discriminating information with sample-10 configuration).

Although with the sample-10 setting there is a loss of information in terms of frames considered, Table 3 shows how the performance on 50 subjects, for 2 out of 4 classifiers, is slightly better than the maximum number of frames considered. Instead, static landmarks experiments produce weaker effects than dynamic landmarks, with a general lowering of performance (about 10% in all three classifiers). This decrease in performance can also be seen in the ROC curves at Figure 9, showing the smallest AUC score for all the classifiers used.

It is useful to compare the result reported in Table 3 (50 subjects from XM2VTS) with the Table 4 (43 subjects from M2FRED) to understand how the uncontrolled environment negatively affect the performance. In fact, a general decrease in accuracy can be noted for all the models used: i.e., the accuracy for the MLP in Table 3 reaches the 89%, in Table 4 is equal to 60%. Finally, for the experiments at Table 3, the best configuration is sample-61 setting while in Table 4 the sample-10 setting shows better results. Table 4

| XM2VTS #50 | MLP | SVM | DT | RF |
|------------|-----|-----|----|----|
| Accuracy   | sample-61 | 0.890 | 0.780 | 0.378 | 0.753 |
|            | sample-10 | 0.800 | 0.792 | 0.382 | 0.766 |
| Precision  | sample-61 | 0.874 | 0.798 | 0.394 | 0.801 |
|            | sample-10 | 0.804 | 0.798 | 0.433 | 0.822 |
| Recall     | sample-61 | 0.890 | 0.781 | 0.378 | 0.753 |
|            | sample-10 | 0.800 | 0.792 | 0.382 | 0.766 |
|            | static    | 0.629 | 0.321 | 0.582 | 0.582 |
|            | sample-61 | 0.890 | 0.781 | 0.378 | 0.753 |
|            | sample-10 | 0.800 | 0.792 | 0.382 | 0.766 |
|            | static    | 0.633 | 0.308 | 0.511 | 0.511 |
TABLE 4. Accuracy, Precision and Recall score on 43 subjects from M2FRED static and dynamic landmarks with different time windows sizes, 61 and 10 frames respectively.

| M2FRED #43 | MLP | SVM | DT | RF |
|------------|-----|-----|----|----|
| Accuracy   |     |     |    |    |
| sample-61  | 0.604 | 0.693 | 0.224 | 0.564 |
| sample-10  | –    | 0.703 | 0.210 | 0.534 |
| static     | 0.371 | 0.554 | 0.213 | 0.347 |
| Precision  |     |     |    |    |
| sample-61  | 0.569 | 0.773 | 0.284 | 0.713 |
| sample-10  | –    | 0.818 | 0.345 | 0.613 |
| static     | 0.328 | 0.602 | 0.226 | 0.412 |
| Recall     |     |     |    |    |
| sample-61  | 0.604 | 0.693 | 0.224 | 0.564 |
| sample-10  | –    | 0.703 | 0.210 | 0.534 |
| static     | 0.371 | 0.554 | 0.213 | 0.347 |

TABLE 5. The results in terms of accuracy, precision and recall achieved in appearance-based modality on the three different configurations of the input data among the ML classifiers.

|         | MLP | SVM | DT | RF | 3DCNN Depth 16 | 3DCNN Depth 10 |
|---------|-----|-----|----|----|----------------|----------------|
| XM2VTS (n=254) |     |     |    |    | 0.951          | 0.951          |
| Accuracy | 0.951 | 0.951 | 0.471 | 0.951 | 0.951          | 0.951          |
| Precision| 0.529 | 0.529 | 0.020 | 0.529 | 0.529          | 0.529          |
| Recall   | 0.949 | 0.949 | 0.477 | 0.949 | 0.949          | 0.949          |
| XM2VTS (n=50)  |     |     |    |    | 0.951 | 0.951 |
| Accuracy | 0.875 | 0.875 | 0.440 | 0.875 | 0.951 | 0.951 |
| Precision| 0.512 | 0.512 | 0.084 | 0.512 | 0.951 | 0.951 |
| Recall   | 0.940 | 0.940 | 0.477 | 0.940 | 0.940 | 0.940 |
| M2FRED (n=43)  |     |     |    |    | 0.951 | 0.951 |
| Accuracy | 0.954 | 0.954 | 0.440 | 0.954 | 0.954 | 0.954 |
| Precision| 0.522 | 0.522 | 0.289 | 0.522 | 0.954 | 0.954 |
| Recall   | 0.955 | 0.955 | 0.290 | 0.955 | 0.955 | 0.955 |

Closes this first collection of experimentation confirming again the significance of using the dynamics of facial landmarks over a static acquisition. The model that reaches the best result terms of accuracy in static configuration is about 55% (refer to SVM column at Table 4).

The contribution of the dynamics is also visible in the Confusion Matrix in Figures 8 and 11: in both cases the matrix relating to the dynamics features are more stable and less sparse then the static ones.

Comparing the achieved results among the considered classifiers, the MLP reaches the highest accuracy on XM2VTS, but this is not true for M2FRED, where the SVM overpasses the competing classifiers. The different nature of the data collected in the two datasets is the reason behind this result. The MLP takes advantage of the laboratory acquisition conditions. They contribute to stabilising the input for the network model which is then able to train its parameters to separate the classes. Conversely, on the challenging cross-device conditions of M2FRED, the MLP is not able to achieve the same robustness, which is indeed higher for the SVM classifier. In this last case, the SVM classifier can better deal with the feature spaces and the noise which affects the acquisitions at different level of difficulties. Projecting the features representation in a higher dimensional space the SVM is less influenced by noise factors that affect the data in a sparse way thus resulting to separate the classes more effectively.

On the opposite side is the analysis of the decision tree (DT), which shows to suffer significantly on this problem. A quasi-exhaustive search of the parameters does not allow to look for a configuration that properly classifies the subjects. DT model is the lighter in terms of computing demand and it is also the easiest to feasibly implement on mobile devices, which would make it the one that best suits the requirements for a lightweight classifier. However, even if more demanding compared to DT, also SVM and MLP models involve a computation load that is compliant with the limitation of mobile devices.

C. APPEARANCE-BASED FEATURES

The following experiments use the appearance-based features extracted with LBP-TOP technique, described in the section III-C1. It is possible to notice a general improvement of the classification performance in comparison to the geometric features using the same datasets and the same number of subjects. Another improvement can be found among the performance achieved on M2FRED compared to XM2VTS. The first one shows better results even if the video acquisitions are made in-the-wild compared to the most controlled acquisitions of XM2VTS. Moreover, the experiments on M2FRED were conducted both on subjects with and without masks. The increase in performance is also reflected in the ROC curves which, in the best case, for the M2FRED dataset, show an AUC equal to 0.99 (Figure 12).

D. 3D CNN

The following experiments use 3D CNN: a Deep Learning approach described in the section III-C2. As for the previous methods, the comparative analysis is carried out considering the XM2VTS and M2FRED datasets. Two experimental settings were taken into consideration, tuning the depth level of the 3D CNN input space: depth=16 and depth=10. As shown in Table 5, 3D CNN reaches the best results in comparison with all other methods, in terms of accuracy (96% in its best configuration). Unlike the geometric features and LBP-TOP, the experiments that involve the XM2VTS dataset show better results in comparison with the M2FRED dataset. According to the ROC curves in figure 12 (b), the best results are obtained with the XM2VTS#50, which show an AUC score of 0.96. Even considering all optimizations possible for the 3D CNN, such a temporal deep model cannot run smoothly on mobiles. In this point of view, the MLP is a more adequate choice for mobile devices even though the hardware limitations for a real-time processing. All other standard classifiers suffer when the number of different subjects to recognise increases. By an accurate inspection of the ROC curves, separately from the aggregated results in Table 5, it can be observed that the performance achieved is not suitable for a reliable biometric system in the majority of the settings. On the other hand, the results achieved on M2FRED show again that the challenging acquisition conditions and the diverse contributions of the noise affecting the video recordings do not limit the performance of the SVM over the
competitors. It even overpass, in accuracy, the performance of the 3D CNN thus resulting, for that dataset, the choice to be preferred. Moreover, the SVM classifiers can be feasibly implemented on mobile devices, so meeting another goal of our work. Comparing the experiments on a fair number of subjects in XM2VTS and M2FRED, again it can be observed that MLP is able to effectively deal with stabilised acquisition conditions of XM2VTS, training the hyperparameters to correctly classify the subjects. SVM and RF indeed collapse in performances on XM2VTS#50 while on M2FRED#43 show promising classification accuracy. As it happens for geometric approach, the SVM does not take advantage of the laboratory acquisition conditions of XM2VTS dataset, rather the noise affecting the uncontrolled acquisition of M2FRED reveals useful to separates the classes. In appearance-based setting, the SMV is moreover the most performing as authentication approach. The ROC curve in Figure 12 clearly shows the superior trend of the curve compared to the others, even if compared with the behaviour of the 3D CNN.

**E. ANALYSIS OF THE FAILURE CASES**

An analysis of the failure cases was carried out to identify any significant drawbacks of the implemented approaches as well as possible gender or age-related biases. By an aimed inspection into the experimental results achieved on the two datasets.
considered in this study, we identify three classes of failures for which the proposed approach misclassifies the identities. They are the gender, the age, and the presence of the glasses, this last one can be considered as noise in periocular-based
TABLE 6. Distribution of the subjects in XM2VTS and M2FRED across the three classes considered in the analysis of failures, that are gender, age and eyeglasses.

|        | XM2VTS(#295) | M2FRED(#43) |
|--------|--------------|-------------|
| Gender |              |             |
| Male   | 53.56%       | 72.09%      |
| Female | 46.44%       | 27.91%      |
| Age    |              |             |
| Young adults (18-30) | 24.06% | 76.74% |
| Middle aged (31-50)  | 27.19% | 2.33%  |
| Older adults (≥ 51)  | 48.75% | 18.60% |
| Glasses |              |             |
| Yes    | 34.58%       | 39.53%      |
| No     | 65.42%       | 60.46%      |

Identification systems. We identify three classes for age in both datasets that are (i) young adults (18-30); (ii) middle aged (31-50); (iii) older adults (≥ 51) (for M2FRED dataset it must be noted that the middle-age class contains one subject only). On the whole dataset population, we found the subjects with the lowest classification scores (up to cases of incorrect detection). Considering the solely extraction stage of the processing pipeline, the geometric landmark extraction algorithm failed to extract the information needed to create a template in case of two different subjects. They are the subject 002 from XM2VTS and the subject 004 in M2FRED dataset, they are shown in Figure 13 and Figure 14 respectively.

By observing Figure 14, the geometric features extraction algorithm fails for the significant environmental noise in the acquisitions: in the indoor acquisition the scene in the background is such to produce reflections of the artificial light in the room that alter the segmentation of the foreground (the face of the subject in our case) thus impacting negatively on the feature extraction. Problems related to lighting can also be observed in outdoor acquisition conditions. The environmental light in this case is responsible for a non-uniform illumination of the face and, therefore, the precision of the detection of the landmark is poor. Moreover, the subject 004 is a critical sample on the dataset because his acquisition is also affected by the reflections onto the eyeglasses he worn, which contributes to those facial shadowing and occlusions determining the detection algorithm to fail. Concerning eyeglasses, it may be supposed that their presence is a sufficient condition for the extraction algorithms to fail. But this is not true. As it can be seen in Table 6 there is a good percentage of people wearing glasses (34.58% for M2FRED and 39.53% for XM2VTS respectively) and most of them are correctly acquired and classified. In some cases, the eyeglass frames occlude the periocular region and add features that hard to be separated from the true facial features of the acquired subject. In case of non-ideal conditions, we have just above discussed that the reflections of the light can again be considered as a case of occlusion. However, eyeglasses are a true source of issues when the lenses alter the shape and the proportions of the face (it can be better realised in samples shown in Figure 15). This, in turn, causes the loss of facial geometries and, hence, the crash of periocular feature extraction. Moreover, since the proposed method in this study is based on a continuous collection of frames, it is important to point out that the geometric feature extraction algorithm might fail for some frames of the total in the time window. A failure in these cases is explained by the fact that the number of faulty frames is high. The biometric template becomes unavailable and the training of the classifiers impossible to be performed properly. The number of incorrect classifications was calculated during the testing phase by using the static and dynamic features.
on the two datasets considered. There were 13 erroneous classifications for M2FRED and 61 for XM2VTS when static features were taken into account.

For the dynamic features, was observed 4 misclassifications for M2FRED and an average of 22.5 misclassifications for XM2VTS: 24 misclassifications were produced by applying geometric features and 21 with LBP-TOP features. The study of the misclassification cases focuses on the best configurations found in Section IV for the two different approaches. In case of the geometric landmark model, we identify MLP and SVM as best classifier for the XM2VTS dataset and SVM for M2FRED. For the appearance-based approach that implements LBP-TOP technique, the best model is the MLP in both experimental settings.

Figures 16 and 17 show the number of misclassified subjects falling into the three classes of failures, measured as a percentage of the subjects belonging to each class. Dynamic and static features are considered respectively. It can be observed that the percentage of males is higher than the females, in each configuration. The age-related attribute is the most variable of the three, while the presence of glasses, in general, does not appear to generate specific issues for a classification algorithm or another. Based on the overall percentage of total subjects (summarised in Table 6), the gender bias in XM2VTS is significant, since the balanced proportion of males and females in this dataset. The same cannot be inferred from M2FRED, where the ratio of males is considerably higher than the ratio females. Similar conditions happen for age in both datasets while in M2FRED with LBP-TOP no failures are reported, a result that is largely justified by the limited number of subjects in the dataset but that, on another side, confirms the feasibility of the proposed method on practical authentication scenarios.

V. DISCUSSION AND CONCLUSION

The periocular region of the face represents a valid biometric trait to be used as authentication mechanism. Over the years, the state of the art focused on techniques for the extraction and the analysis of static periocular features, meaning the features coming from a single acquisition of the face. Recently, the analysis of periocular region gained a privileged position. COVID-19 pandemic has led several public health institutes worldwide to impose the use of facial masks to counter the transmission of the virus among people. Such restrictions represent a crucial factor for face recognition systems, which are nowadays adopted in several public areas as well as used as unlocking systems for personal mobile devices (e.g., smartphones and tablet). In this work, we focus the attention on the analysis of the periocular recognition aiming at discussing the contribution of the dynamics of the facial periocular
features may introduce in biometric authentication systems. Macro- and micro-expressions can significantly transform the appearance and the distribution of facial features. Approaches based on a static acquisition could not well generalise the possible different appearances that an authorised identity might show over time. With this assumption, the proposed study explores the potentials and the robustness of analysing the dynamics of the periocular facial features through traditional Machine Learning classifiers. Such a choice sounds to be in contrast with the most famous and established Deep Learning approaches that have been proved very effective in several computer vision tasks. In our work, we compare the results achieved by Machine Learning classifiers with a Deep Learning model suitable for our problem and based on temporal Convolutional Neural Network (CNN), which is known as 3D CNN. The comparative analysis shows that Deep Learning solutions are not significantly superior in terms of achieved recognition accuracy thus confirming the challenging points of the problem. Occlusions and noise can relevantly impact on the quality of the acquired trait limiting the reliability of the recognition. On the other hand, using Machine Learning solutions has a potential positive impact on the adoption of the proposed approaches on mobile devices. To prove the effectiveness and the limitations of the approaches, a dataset recording videos of people wearing a mask through mobile devices were necessary but not publicly available. For this reason, a dataset called M2FRED (Mobile Masked Face Recognition through periorcular Dynamics analysis) has been presented in this study. According to an established protocol aimed at inducing micro and macro facial expressions while speaking, the M2FRED dataset collects the acquisitions of 43 different subjects in a cross-device setting thus resulting in a challenging dataset for biometric recognition. Another dataset has been considered, namely XM2VTS, that collects videos of the faces of a bigger number of people but without wearing a mask. To achieve a fairer comparison, the XM2VTS samples have been pre-processed to virtually apply a mask over all faces. This could be considered a redundant step since the proposed work explores the solely features of the periorcular region. However, working on the whole visible face and discarding the lower part would not create those conditions that make the XM2VTS totally comparable with the proposed M2FRED dataset (where masks can occlude parts of the faces that also involve the periorcular region). The analysis of the periorcular region for biometric recognition has been performed in two different modalities that differ in the representation of the facial features. The first is a geometric modality, consisting in the detection of facial landmarks and the computation of their dynamics over time. The second modality starts from the appearance of the periorcular region frame-by-frame and builds a temporal feature map. In both cases, the representations are used as input for Machine Learning classifiers to discuss the level of performance achieved and to compare those results with the 3D CNN (in such last case just for appearance-based modality). Separately discussing the results achieved, an improvement has been reported for the geometric modality when the representation of the feature was dynamic. In all cases of comparison with static features, the classifiers exhibited better performances in accuracy, precision and, recall thus confirming the positive contribution that the dynamics of the features introduced in the analysis. Apart from the significance of the level of performance achieved, another relevant contribution of using geometrical features is that they are relatively easy to extract from the video stream. Once the periorcular area has been correctly detected in a video frame, the extraction of the landmarks and their tracking over time and space is computationally light. This is a suitable condition to consider the geometrical approach as a viable solution for mobile devices computing. On the other hand, the expectation that appearance-based model could perform better has been confirmed. In the second modality, the analysis is carried out at pixel-level. The experimental results show that the overall level of performance achieved is comparable with the results of the geometric-based approach, but a slightly higher level of performance has been observed. This, however, at a computational demand of the feature extraction process and the analysis that is notably higher in many cases. Moreover, some extra considerations can be done, especially by comparing with the Deep Learning solution included in the experimentation. From one side, it can be observed unstable results with classical classifiers, except for MLP. On the other side, the 3D CNN better generalises the problem among different datasets, thus representing an interesting result that may reduce the significance of the results achieved by the ML classifiers. However, the objectives of our work were analysing the feasibility of solutions for mobile devices and 3D CNN are undoubtedly not a choice to prefer. Conversely, the SVM exhibited a reasonable level of performance on M2FRED dataset as an authentication system, confirming as a feasible option for authentication approach to implement on mobile devices. Biometric recognition systems can be relatively accurate and precise, especially when conditions are favourable. The challenging conditions on which this paper has been conceived do not enable a level of performance comparable with well-assessed authentication systems. On the other hand, it shows how traditional Machine Learning solutions can compete, under some assumptions, with well-established Deep Learning solutions. The experimental analysis is supported by a discussion of the failure cases which is useful to focus on those conditions that make the problem even more challenging.

The experimental results also emphasize the gap of performance to get a really reliable face biometric recognition suitable for mobile devices in in-the-wild conditions. As a future improvement of this work, an accurate analysis of the facial features and their single contribution on total recognition could allow us to understand which real facial features weigh most heavily on recognition and which, if any, can even be ignored if they have a negative impact. Evaluating the importance of facial features would thus enable the achievement of a model that is not only more robust but also computationally
lighter in case an effective strategy for mining facial features can be identified.

From another point of view, Generative Adversarial Networks (GANs) have shown impressive result in generating likely images of different objects, including the human faces. Their ability to extract salient features from samples and use them to generate new samples has been discussed as a potential risk for face recognition. In our work, GANs could be used to generate new samples of the subject wearing masks and to understand is the model is able to replicate the solely periorcular features to provide a identity-safe reconstruction. However, in consideration of the goal of our work, the challenge of using GANs is that the neural model should generate a video (and not a simple image of the face) mimicking the gestures and the expressions of the subject, which presents an engaging research challenge.

REFERENCES

[1] L. Flom and A. Safir, “Iris recognition system,” U.S. Patent 4,641,349, Washington, D.C., Dec. 3, 1987.
[2] N. L. Thomas, Y. Du, and Z. Zhou, “A new approach for sclera vein recognition.” in Proc. SPIE, vol. 7708, pp. 770–805, Apr. 2010.
[3] R. B. Hill, “Apparatus and method for identifying individuals through their retinal vessel patterns,” U.S. Patent 4,109,237, Aug. 22, 1978.
[4] M. De Marsico, M. Nappi, F. Narducci, and H. Proença, “Insights into the results of MICHE I—Mobile iris challenge evaluation,” Pattern Recognit., vol. 74, pp. 286–304, Feb. 2018.
[5] U. Park, A. Ross, and A. K. Jain, “Periorcular biometrics in the visible spectrum: A feasibility study,” in Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst., Sep. 2009, pp. 1–6.
[6] U. Park, R. Jillela, A. Ross, and A. K. Jain, “Periorcular biometrics in the visible spectrum,” IEEE Trans. Inf. Forensics Security, vol. 6, no. 1, pp. 96–106, Mar. 2011.
[7] P. Kumari and K. R. Seexa, “Periorcular biometrics: A survey,” J. King Saud Univ.-Comput. Inf. Sci., vol. 34, no. 4, pp. 1086–1097, Apr. 2022.
[8] I. Nigam, M. Vatsa, and R. Singh, “Ocular biometrics: A survey of modalities and fusion approaches,” Inf. Fusion, vol. 26, pp. 1–35, Nov. 2015.
[9] V. Cantoni, L. Cascone, M. Nappi, and M. Porta, “Demographic classification through pupil analysis,” Image Vis. Comput., vol. 102, pp. 983–989, Oct. 2020.
[10] F. Alonso-Fernandez and J. Bigun, “A survey on periorcular biometrics research,” Pattern Recognit. Lett., vol. 82, pp. 92–105, Oct. 2016.
[11] A. Ulhaq, J. Born, A. Khan, D. P. S. Gomes, S. Chakraborty, and M. Paul, “COVID-19 control by computer vision approaches: A survey,” IEEE Access, vol. 8, pp. 179437–179456, 2020.
[12] E. Barroso, G. Santos, L. Cardoso, C. Padole, and H. Proença, “Periorcular recognition: How much facial expressions affect performance?” Pattern Anal. Appl., vol. 19, no. 2, pp. 517–530, May 2016.
[13] G. Santos and E. Hoyle, “A fusion approach to unconstrained iris recognition,” Pattern Recognit. Lett., vol. 33, pp. 984–990, Sep. 2012.
[14] R. Jillela and A. Ross, “Mitigating effects of plastic surgery: Fusing face and ocular biometrics,” in Proc. IEEE 5th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS), Sep. 2012, pp. 402–411.
[15] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2005, vol. 1, no. 1, pp. 886–893.
[16] D. L. Woodard, S. J. Pundlik, J. R. Lyle, and P. E. Miller, “Periorcular region appearance cues for biometric identification,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops, Jun. 2010, pp. 162–169.
[17] T. Ojala, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Aug. 2002.
[18] A. Gangwar and A. Joshi, “Robust periorcular biometrics based on local phase quantisation and Gabor transform,” in Proc. 7th Int. Congr. Image Signal Process., Oct. 2014, pp. 714–720.
[19] R. Raghavendra and C. Busch, “Learning deeply coupled autoencoders for smartphone based robust periorcular verification,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2016, pp. 325–329.
LUCIA CIMMINO (Graduate Student Member, IEEE) received the B.S. and M.S. degrees (cum laude) in computer science from the University of Salerno, Fisciano, Salerno, Italy, in 2018 and 2020, respectively, where she is currently pursuing the Ph.D. degree in computer science with the Biometric and Image Processing Laboratory (BIPLAB). Her research interests include machine learning, biometrics, facial recognition, and image processing.

Michele Nappi (Senior Member, IEEE) received the Laurea degree (cum laude) in computer science from the University of Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from I. I. A. S. S. “E. R. Caianiello,” Vietri sul Mare, Salerno, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Italy. He is currently a Full Professor in computer science with the University of Salerno. He is also a Team Leader of the Biometric and Image Processing Laboratory (BIPLAB). He has coauthored more than 190 papers in international conferences, peer review journals and book chapters in these fields. His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human–computer interaction, and VR/AR. He was a member of IAPR. He received several international awards for scientific and research activities. He was the President of the Italian Chapter of the IEEE Biometrics Council, from 2015 to 2017.

Chiara Pero (Member, IEEE) received the B.S. and M.S. degrees (cum laude) in computer science from the University of Salerno, Italy, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree in computer science with the Biometric and Image Processing Laboratory (BIPLAB). Her research interests include machine learning techniques in facial recognition, image processing, behavioral profiling, and activity recognition.

Fabio Narducci (Member, IEEE) received the Ph.D. degree in computer science from the Virtual Reality Laboratory, University of Salerno, in 2015. He is currently an Assistant Professor with the University of Salerno. He served as the TPC for international conferences and as a reviewer of several international journals. He was the Guest Editor of Special Issues on IEEE Cloud Computing, Pattern Recognition Letters (Elsevier) and Multimedia Tools and Applications. He is currently an Associate Editor of IET Electronic Letters and an Editorial Board Member of International Journal of Embedded Systems (Inderscience) and Electronics (MDPI). His research interests include machine learning and deep learning, image processing and pattern recognition, biometrics, mobile, and wearable computing.

Open Access funding provided by ‘Università degli Studi di Salerno’ within the CRUI CARE Agreement