Trustworthy Method for Person Identification in IIoT Environments by Means of Facial Dynamics

Aniello Castiglione, *Member, IEEE*, Michele Nappi, *Senior Member, IEEE*, and Stefano Ricciardi, *Member, IEEE*

***Abstract*—In industrial Internet of Things (IIoT) environments, dependability of a complex manufacturing process in which human operators play a key role can be improved by identity recognition/authentication of whoever is involved in various stages of a production process, according to where and when he/she is supposed to be. To this aim, we propose an approach that exploits the dynamic appearance and the time-dependent local features characterizing the face of an individual during speech utterance with regard to their spatial and temporal components. The proposed method models these dynamic facial patterns captured from edge Internet of Things devices by means of the Local Binary Pattern on Three Orthogonal Planes descriptor, which effectively extract both face's local features and movement at the fog level of the architecture. A deep feedforward network available in the cloud is trained and optimized to match the extracted features to a reference database. The achieved results highlight state-of-the-art performances of the proposed method with regard to robustness and trustworthiness of identification, especially for challenging IIoT scenarios.**

***Index Terms*—Biometrics, face recognition, image analysis, Industrial Internet of Things (IIoT).**

I. Introduction

IN THE past few years, the diffusion of the industrial Internet of Things (IIoT), as a network of a multitude of industrial devices connected by communications technologies [1], has become increasingly more pervasive in many industrial contexts. This diffusion is supported by the promise of bringing together smart machines, advanced analytics, and people at work to the aim of enabling unprecedented levels of efficiency, productivity, and performance [2].

Even though the "mainstream" application paradigm of the IIoT typically translates in a reduced need of involving humans throughout the industrial process as a side effect of machines' advanced data capture, analysis, and communication capabilities, there are cases in which human operators are necessary to the process itself. This is particularly true for complex manufacturing processes of mission-critical components and systems, such as those required in aerospace and defense industry, or even for applying the IIoT framework to the healthcare scenario. In several cases, indeed, the advantages of an IIoT environment can be fully exploited if the identity of the involved human operators can be reliably assessed by the same devices they are working with, according to where and when they are supposed to be present in the process.

To the aim of addressing the aforementioned challenges, we propose a highly trustworthy face recognition approach based on dynamic features, which can be acquired, extracted, and matched to a reference gallery, respectively, at the edge, fog, and cloud levels of an IIoT architecture.

The main idea behind this proposal lies in the increased trustworthiness provided by facial motion associated with speech for person recognition and authentication, compared to static face representations. According to this hypothesis, indeed, a dynamic descriptor of face local changes due to the utterance of a given passphrase results in a time-variable biometric template much harder to attack or counterfeit than conventional static descriptors.

More in detail, the proposed approach to face motion representation and matching is intrinsically more robust to presentation attacks than static descriptors, and it implies the liveness of the subject pronouncing the passphrase.

Consequently, most of the presentation attack strategies (i.e., presentation to the recognition system with the goal of interfering with its operation), such as subject impersonation, finding a look-alike, making appearance similar to the reference, and artefact presentation, became much more difficult, if not impossible, to apply with success. Malicious users should be able to mimic not only the overall appearance of the target subject, but also his specific facial dynamics related to the utterance of the passphrase.

Furthermore, system's trustworthiness is also possibly affected by false rejections, which may undermine the regular authentication of genuine users. This aspect, indeed, is particularly relevant in an industrial application context, where the incorrect recognition of authorized personnel due to a high sensitivity threshold would possibly lead to costly operation delays and reduced acceptability of the whole procedure.

Manuscript received February 7, 2020; accepted February 23, 2020. Date of publication April 27, 2020; date of current version November 18, 2020. This work was supported in part by the Italian National Research Project PRIN 2015 (201548C5NT) entitled "*COntactlesS Multibiometric mObile System in the wild: COSMOS.*" Paper no. TII-20-0629. *(Corresponding author: Aniello Castiglione.)*

Aniello Castiglione is with the Department of Science and Technology, Centro Direzionale di Napoli, University of Naples Parthenope, 80143 Naples, Italy (e-mail: castiglione@ieee.org).

Michele Nappi is with the Department of Computer Science, University of Salerno, 84084 Fisciano, Italy (e-mail: mnappi@unisa.it).

Stefano Ricciardi is with the Department of Biosciences, University of Molise, 86100 Campobasso, Italy (e-mail: stefano.ricciardi@unimol.it).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2020.2977774

To this regard, the proposed dynamic facial signature is designed to balance the requirement of the lowest possible false acceptance rate (FAR) with a good robustness to a certain degree of error in pronouncing the passphrase, thus achieving a reduction of false rejection rate (FRR) as well.

Briefly, the contributions in the proposed approach are as follows:

1) high recognition accuracy, coupling the distinctiveness of face's shape to the unique facial pattern series associated with speech;
2) inherent trustworthiness due to the spatial and temporal characteristics of facial dynamics;
3) reliable genuine subject authentication, even in the case of mispronounced or partially pronounced passphrase;
4) improved robustness to presentation attacks, compared to static approaches, due to the required dynamic facial signature;
5) effective and efficient processing at each stage of the biometric pipeline, thanks to the multilevel IIoT network architecture exploited.

According to the proposed operation flow, a sequence of frames captured during speech by devices at the edge of the IIoT network is processed at the fog level to extract dynamic local features related to the lower half of face using a variant of the Local Binary Pattern [3], namely the Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) descriptor. The resulting feature vector is then compared to a reference gallery through a (previously trained) deep feedforward neural network available at the cloud level of the architecture.

The experiments conducted show state-of-the-art recognition accuracy along with high robustness to the way the sentence is pronounced by the genuine subject, good independence from the choice of the sentence, and a fast matching time, enabling a near-real-time response to the input query compliant to industrial operation requirements.

The rest of this article is organized as follows. Section II resumes a selection of works related to this article. Section III describes the proposed approach to facial dynamics biometrics. Section IV describes the IIoT environment, in which the proposed approach has been tested. Section V presents the results of experiments carried out. Finally, Section VI concludes this article.

II. Related Works

Security issues and authentication strategies represent some of the most complex and crucial challenges related to the data interchange among the sensing, network, and application layers characterizing the Internet of Things (IoT) [4].

These challenges can be even more significant in the context of IIoT, where radio-based, acoustics-based, light-based, image-based, gesture-based, and biometrics-based authentication mechanisms have been proposed, as reported in [5].

Biometrics, in particular, exploits the something-the-user-have authentication paradigm to the aim of end-user-to-device authentication. Most diffused biometrics used for this purpose are fingerprints, iris, and face.

The latter is arguably the most suited to the IIoT context, since it is contactless and does not require specialized hardware but a digital camera, which is often (or can be easily) embedded in a variety of devices and equipment.

Usually, health signals, i.e., electrocardiogram (ECG) signal, is not strictly considered a biometric signature. However, since ECG reveals heart condition and cardiac risks, it has also been explored in healthcare IIoT approaches, to the aim of combining "things" and humans.

Works such as [6] and [7] propose architectures for tracking, monitoring, and storing people's health status and related data in elder people.

In this line of research, a more generalized approach to combine biometrics-based authentication and the IoT for smart health applications is proposed in [8].

However, even in more mainstream IIoT applications and services, security and privacy are paramount, as discussed in [9]. To this regard, Guo *et al.* [10] explore the potential benefits and technical difficulties related to the incorporation of biometric technologies into the IoT, focusing on reducing the risk of unauthorized access, tampering, and even reverse engineering of IoT devices.

Mainly aiming at cloud-based IIoT architectures, the authors of [11] propose a two-factor user authentication methodology for privacy preserving by means of user's smartcard and biometric signature, adopting a fuzzy biometric verification approach for the user and bitwise XOR plus cryptographic hash coding at the smart-devices' end.

The advantages of biometrics in enabling more secure end-to-end communication solutions among interconnected devices and services within IIoT environments are highlighted in [12], where user's face recognition by means of mobile devices increases the security of the IoT infrastructure.

In [13], the authors focus on the wireless sensor network component of the IIoT to address the limitations of typical protocols through a biometrics-empowered authentication protocol with elliptic curve cryptography, which results to be more reliable and effective than other state-of-the-art protocols.

The potential of biometrics for enhancing security and privacy can be even greater for fog and edge computing, which are particularly suited to IoT and IIoT scenarios when mobility, scalability, and reliability are required.

This is the observation behind [14], where an approach to enhance security of face biometrics by means of visual cryptography and zero-watermarking is proposed. In [15], the protection of fog computing environment in IIoT applications is addressed through a hybrid biometric smartcard authentication method, exploiting a combination of a short (maximum eight characters long) person identification number and fingerprint biometrics.

To the aim of achieving identity consistency between physical and cyber-space in the IoT, the authors of [16] propose a face-biometrics-based identification and resolution method enabled through fog computing to save bandwidth and reduce computing load of both identification and resolution tasks.

Also on this topic, Hu *et al.* [17] work on the confidentiality, integrity, and availability properties associated with the

aforementioned face identification and resolution framework in fog-computing-enabled IoT.

As a further evolution of the last two approaches, a unified cloud-enabled parallel matching of face identifier for IoT authentication and resolution of physical objects is described in [18], providing an effective improvement over the previous methods.

Finally, the work by Karimian *et al.* [19] investigates the implications of the next generation of IoT devices and technologies in the light of the incorporation of biometrics into the IoT design for a ubiquitous distributed authentication strategy delivering the Internet of Biometric Thing (IoBT) paradigm.

Though the aforecited works show the advantages of biometrics in general and particularly face for improving security through object and person authentication in IIoT environments, even biometrics can be forged, and face is no exception to this rule.

To address this challenge, a dynamic face descriptor is proposed along with an edge–fog–cloud architecture to efficiently capture, process, match, and resolve it, providing greater accuracy and trustworthiness against counterfeits.

The dynamic facial features exploited in this article are related to the way the uttering of a short sentence, captured through a frames sequence, locally affects the shape and texture of the lower portion of the face.

It is worth to remark that the proposed approach differs from lip-feature- and lip-motion-based methods, since our method is not limited to the lip component of the motion but is designed to analyze the entire surrounding area instead.

This distinction applies to any of the works based on stacked sparse autoencoders [20], particle-filter-based motion tracking [21], color and geometric components [22], orientation maps [23], [24], hidden Markov models [25], statistical analysis [26], deep neural networks [27], [28], recurrent neural networks [29], multiboosted learning [30], support vector machines [31], [32], time-series matching [33], or Gaussian mixture models [34].

Moreover, the adopted LBP-TOP descriptor is capable of embedding both spatial and temporal characteristics of dynamic facial patterns, which are crucial for dependable subject identification.

III. Method Description

As briefly anticipated in the previous sections, the proposed approach takes advantage from a three-level architecture designed to distribute the main steps of a general biometric pipeline among different kinds of nodes for maximum processing efficiency and result effectiveness.

More in detail, the acquisition step, in charge of capturing the dynamic biometrics and preprocessing it in real time to send a normalized video stream the next level, is located at the edge level.

The following step is aimed at extracting a (dynamic) feature vector from the previously acquired video and is performed on fog nodes, while the matching of the input to the reference templates is performed in the cloud, thus realizing an Industrial IoBT framework.

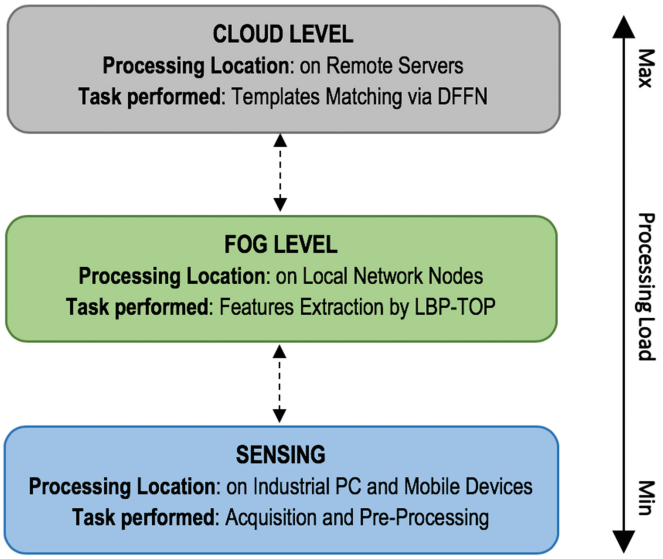

Fig. 1. Schematic view of the overall processing architecture for the proposed method.

In fact, as shown in Fig. 1, the proposed approach can be logically divided into three different layers in order to provide the best results in effectiveness and process efficiency.

The functionalities at the different layers are explained in detail in the following subsections, but the key idea is that different tasks are performed at different networking levels.

The level closest to the humans is the edge where the sensing activities are performed: here, at this level, the devices (mostly industrial tablet PCs and smartphones) performed the biometric acquisition, video-preprocessing, and sample normalization tasks. Those activities are the most privacy invading, since such devices are the only equipment that interact with human and that capture biometric data. Therefore, having such separation, it is fundamental to achieve a privacy-preserving sensing activity on humans.

In addition, at this level, there is the biggest amount of data to be processed, since the acquired video samples are not moved on the network and remain confined at this level (i.e., the edge level). This results in a very limited processing load (from a CPU perspective) and, from a networking prospective, in a quite limited amount of data to be moved across the network and to be sent to the upper level of the architecture (i.e., the fog level).

The main challenge at the edge level is to keep the biometric data as much private as possible by not exposing the video samples over any kind of communication network, since everything is processed locally.

Another important challenge to face up is to try to reduce the amount of data to propagate at the fog level. It is worth to note that the edge nodes communicate mostly using 4G/5G or Wi-Fi connections, thus using not so large bandwidth. In fact, it is well known that the IIoT is populated by different kinds of distributed devices that can communicate with a very low power consumption. Usually, such devices are referred to as LR-WPAN, that stands for low-rate wireless personal area network. The standard behind the LR-WPAN is the IEEE 802.15.4 [35], [36].

The second level is the fog level that receives the processed data from the lower level (i.e., the edge level). From a networking perspective, this level can be seen as a local area network aggregating several edge nodes. At this level, the privacy-preserving requirement can be easily achieved, since all the traffic coming from different edge nodes can be logically separated using the IEEE 802.1Q standard [37] to partition (and isolate) segment of (physical) LAN into different virtual LANs (the so-called VLANs).

This is also very important since the IEEE 802.1Q standard contains provisions for a quality-of-service prioritization scheme, such as the IEEE 802.1p [38]. Industrial environment is fundamental to guarantee enough communication bandwidth to all involved devices mainly to avoid any kind of denial-of-service attacks coming from some compromised devices operating at the edge level.

At this level, the available bandwidth is bigger than the one at the edge level, and it can be esteemed in a scale of 1–10 GB. Therefore, we can state that both the networking and the processing load are considerably bigger than the lower level. In particular, from a processing load point of view, the fog level is responsible for processing the video samples and performing the feature extraction. The latter can be performed on a mid-size server that can operate within a mid-size factory/plant.

The last level of the proposed architecture is the cloud. The cloud level receives all the data from the lower levels (edge and fog) and is responsible for the biometric signature matching. This is achieved by using a deep feedforward network for checking the reference database and perform the template matching.

At this stage, from the processing load point of view, we experience the maximum load. Also, from a networking point of view, we can say that there will be a fair and fast access to the Internet in order to interact with the cloud service provider on which all the processing are performed. Anyway, since most of the data processing is done at the lower levels, a bandwidth similar (around 1–10 GB) to the one at the fog level will be enough.

Regarding the aspects of privacy concerning the video clips, we can state that first of all no video clip is spread outside the edge/fog and, second, in this case, we can use the privacy-preserving functionalities provided by most of the cloud service providers [39]–[41].

The whole three-level architecture resumed above is designed to achieve near-real-time operation and response, which are likely to be considered a key requirement of an industrial-grade identification system, where minimal personnel distraction and interruption of the working routines should be guaranteed.

A. Edge Level: Acquisition

The first layer of the proposed architecture is made up by heterogeneous edge devices, possibly including smartphones, industrial tablets, embedded PCs, and any other kind of connected devices equipped with a camera and suited to perform a preliminary processing typically requiring limited computing resources.

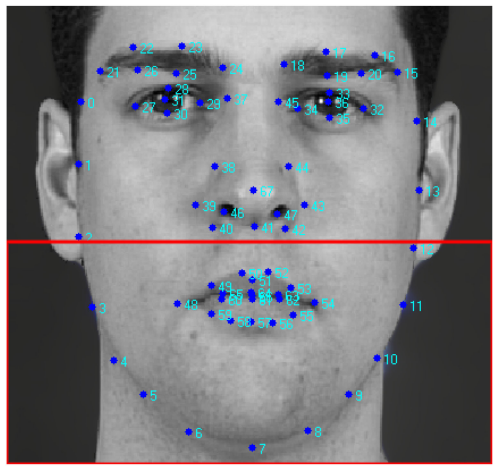

Fig. 2. Full set of 59 facial landmarks detected for cropping the region of interest, which is the region below the imaginary line passing between landmarks #2 and landmark #12, highlighted in red.

Subject acquisition is a double-step process: the recording of video footage via a digital camera and its preprocessing (prior to image analysis) to normalize it with regard to the number of frames. This latter step implies a resampling process to the aim of obtaining a clip whose length is consistent to the length of any gallery samples. Each frame of the resulting sequence is, therefore, analyzed by a face detector [42] that allows us to identify the image region in which the subject's face is present.

Subsequently, up to 59 facial features are found on the face crop previously detected by means of an efficient landmark predictor based on [43]. By exploiting these numbered landmarks, the frame is cropped again retaining only the lower face region comprised below the ideal line connecting landmark #2 to landmark #12 (see Fig. 2).

This choice is based on a specific analysis we conducted by testing the recognition accuracy achieved with three different cropping regions, respectively, including the whole face (except the hair), everything that is below the eyebrows (crop line passing between landmarks #22 and landmark #16), and everything that is below the nose (crop line passing between landmarks #2 and landmark #12).

The latter cropping region resulted the most discriminant and, at the same time, the smaller with regard to the number of pixels, so that the subsequent LBP-TOP feature descriptor would work more on less pixels with a relevant computing-time advantage.

Finally, the video segments thus obtained are converted into grayscale, spatially resampled to a resolution of 200×200 pixels and sent in this compact form to the fog nodes for further processing.

B. Fog Level: Feature Extraction

At this level, fog nodes take in input the preprocessed video of the lower face region of the subject pronouncing the passphrase to extract local spatial–temporal features. To this aim, a computationally less demanding version of the Volume Local Binary Pattern dynamic textures descriptor [44], namely LBP-TOP, is adopted. This more efficient descriptor, indeed, considers only

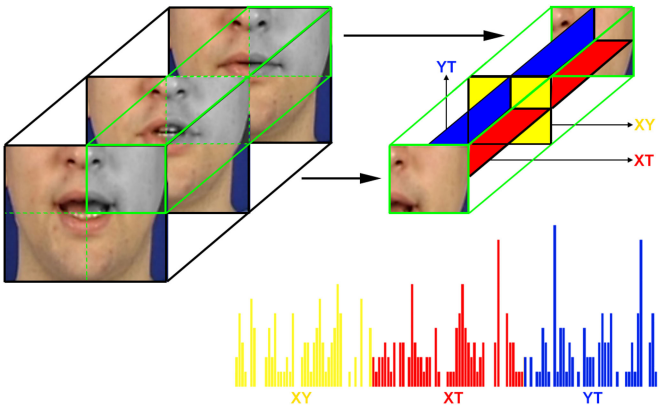

Fig. 3. LBP-TOP descriptor generation. From a sequence of frames capturing the motion in the region of interest, local spatial and temporal features are extracted according to three orthogonal planes YT, XT, and XY (where temporal dimension T corresponds to Z spatial axis) and represented through concatenated histograms.

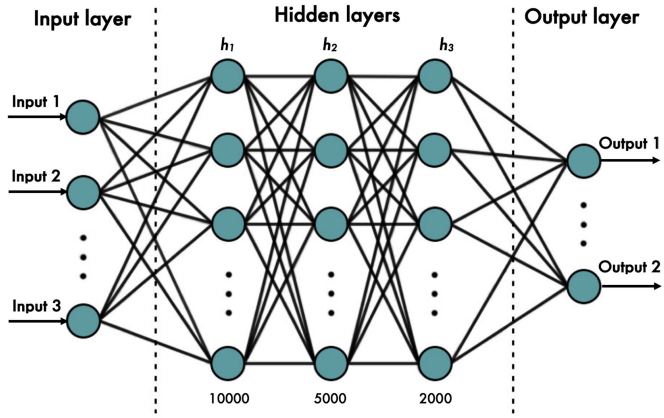

Fig. 4. Network layout of the fully connected deep feedforward network architecture used in the proposed method. The number of input nodes *n* is determined by the dimensions of the feature vector, whereas the number of output nodes *m* is equal to the number of subjects in the database.

three orthogonal planes for analyzing the local features and is extensively described in [45].

The LBP-TOP technique reduces the number of possible patterns by $2(3p + 2)$ (when considering only three planes in the Z dimension) to $3 \times 2p$, where p represents the number of neighboring points. In this article, 36 spaced points were used on a circumference of radius 6, centered on the pixel of interest. The patterns thus obtained are then scaled into integers that can be represented on 8 bits. The binary patterns obtained are extracted from the XY, XZ, and YZ planes. The histograms obtained from the three planes are linked together obtaining a single vector of features (see Fig. 3).

A useful extension of the original operator is the so-called uniform pattern, which can be used to further reduce the length of the feature vector. Indeed, some binary patterns occur more often than others in image textures. An LBP code is said to be uniform when it contains only binary patterns that have at most two transitions 0–1 or 1–0. The histogram relative to an LBP technique with uniform pattern will have a distinct bin for each uniform pattern, while it will have a single bin for all nonuniform patterns.

In the specific case, considering the value of the LBP code expressed on 8 pixels (with possible values between 0 and 255), there are 58 different uniform patterns, and therefore, the final histogram will consist of 59 bins, where the 59th represents the "other" class.

C. Cloud Level: Matching and Decision

At the cloud level, the feature vector resulting from previous levels of processing is, therefore, matched to a reference database by means of a (previously trained) fully connected deep feedforward neural network outlined in Fig. 4.

The network provides in output a percentage of probability of belonging to each class, for each sample shown in the testing phase. The class with the highest percentage is then selected without the use of particular thresholds.

The choice of parameters, activation functions, and architecture was determined on an experimental basis; a series of tests were, therefore, performed, modifying the combinations of these variables, and the best architecture and setup obtained a recognition accuracy of 98.83% on the test set. The rectified linear unit activation function was chosen for the input layer; for the three hidden layers, the sigmoid activation function was preferred; and the softmax was selected for the output layer. The number of input nodes was set equal to the size of the feature vector, while the number of output nodes is determined by the number of subjects in the database.

The network was implemented through the Keras framework with Tensorflow backend; the optimizer and the evaluation metric used are, respectively, stochastic descending gradient and accuracy. All the other parameters of the network, such as the number of epochs, batch size, learning rate, momentum, decay, and dropout, have been optimized experimentally.

The best performing configuration resulted to be the following: epochs $= 20$, batch size $= 32$, learning rate $= 0.1$, decay $= 0.000001$, and momentum $= 0$.

IV. Testing Environment

The IIoT environment, in which the proposed approach has been experimented, is an assembly and repair facility of mission-critical systems for aerospace and defense applications. In this environment, typically, three main kinds of activities take place: assembling (from on board mounting of electronic components up to board installing/wiring into larger units), testing (of boards and units throughout the assembling path and during diagnostic activities within repair procedures), and disassembling (of systems, subsystems and boards for local testing and repair).

This industrial context and the related manufacturing and servicing processes are strictly regulated by formally defined procedures, which, necessarily, involve the active presence of highly specialized workers and technicians, which have to be present in the right location at the right time according to a detailed operating schedule associated with each activity. In such an environment, there is a clear advantage in identifying/authenticating personnel at various locations by means of different kinds of devices they have at hand.

In this particular case, industrial tablet PCs have been used at the edge level for ubiquitously acquiring the subjects to be authenticated, while the local intranet and the related processing servers have been used, respectively, to transfer the captured video clips to the fog level and to process them for feature vector extraction before sending it to the deep learning network implemented via Google Cloud services.

V. Experiments

The experiments described in the following were conducted on a custom-built database, containing short video clips of subject acquisition pronouncing brief sentences. More in detail, 48 subjects (33 of which were males and 15 females, selected to maximize interclass variability with regard to age, gender, and phenotype) were enrolled in the gallery by capturing their face through a camera embedded in an industrial equipment and featuring 1280 × 720 pixels of sensor resolution at 30 frames/s of acquisition rate. Edge node processing power was provided by an industrial-grade PC by Machine Vision, Inc., featuring a quad-core Intel i7@ 3.2 GHz with 16 GB of RAM, integrated GPU, and a 128-GB solid-state drive.

For each one of the enrolled subjects, the database contains three passphrases recorded throughout eight sessions over a span of five months under mildly controlled conditions. The temporal distance among the different acquisitions resulted in a rather ample intraclass variability of the same subject in the course of different sessions, for example, due to change of the hairstyle, growth of beard, and presence or absence of glasses.

The sentences pronounced were the following ones: *"My face is my key to this system"*; *"Please, authenticate me"*; *"This key is my system to my face,"* where the third sentence is deliberately a reassembling of all the words of the first sentence in a different order.

It is worth noting that a resampling operation has been necessarily performed on all video samples from all subjects in order to obtain a uniform feature vector, compensating the rather large variations in the number of frames associated with the utterance of each passphrase due to subject-characteristic speaking speed, passphrase-specific difficulties in memorizing and correctly repeating them (particularly for the third sentence which has no sense), and session-dependent utterance speed. Such factors, indeed, determined a variable length in the unprocessed videos ranging from a minimum of 90 frames to a maximum of 238 frames that have been normalized.

Three experiments were designed and carried out to evaluate the effectiveness of the proposed method within the IIoT testing environment. In all these three experiments, only the first passphrase *"My face is my key to this system"* was used for the training of the deep feed forward network (DFFN) network, partitioning the available samples into 75–25% sized subsets, respectively, for the train set and the test set.

In the first experiment, the same passphrase used for train the network was also used to test it. The overall performance of the biometric components of the system (from edge to fog and cloud levels) is graphically outlined by the receiver operating characteristic (ROC) (see Fig. 5) and FAR/FRR (see Fig. 6)

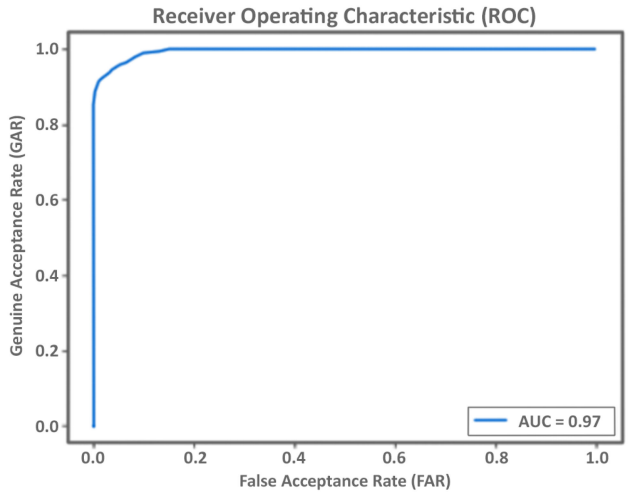

Fig. 5. ROC curve resulting from the first experiment. The area under the curve highlights a remarkable performance.

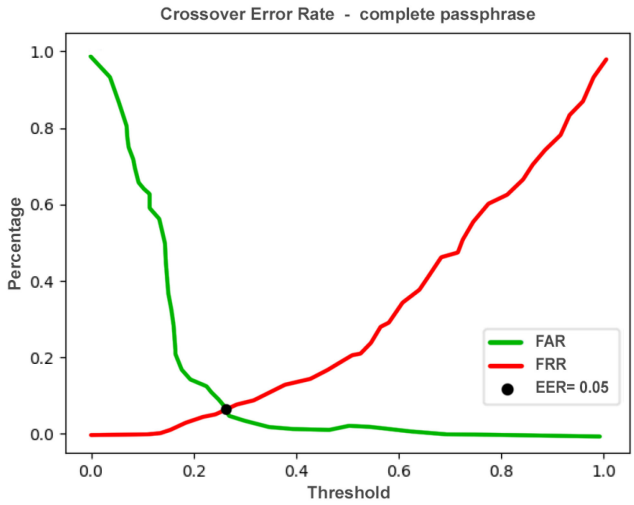

Fig. 6. FAR/FRR curves with equal error rate point located at 0.05%.

curves, which both show an almost ideal behavior, as further confirmed by the equal error rate (EER) value of 0.05 and a cumulative match curve (CMC) (see Fig. 7) of 98.8% already at rank 1, 99.7% at rank 2, and reaching 100% at rank 3.

In other terms, the system is able to perform a correct classification in almost all cases, even with a low decision threshold. Note that in the CMC curve, only the first five out of 48 rank have been reported in order to magnify the behavior between ranks 0 and 3.

In the second experiment, the network was tested on the third sentence *"This key is my system to my face"* (featuring misplaced words of the first sentence), unknown to the training process, and built to test the ability of the approach to correctly recognize a genuine subject even in case some words are unwillingly misplaced. Unsurprisingly, in this test, the percentage dropped slightly to 98.1%.

In the third experiment, the alternative passphrase *"Please, authenticate me"* was used for testing the robustness of the system to a wrong passphrase pronounced by a genuine subject.

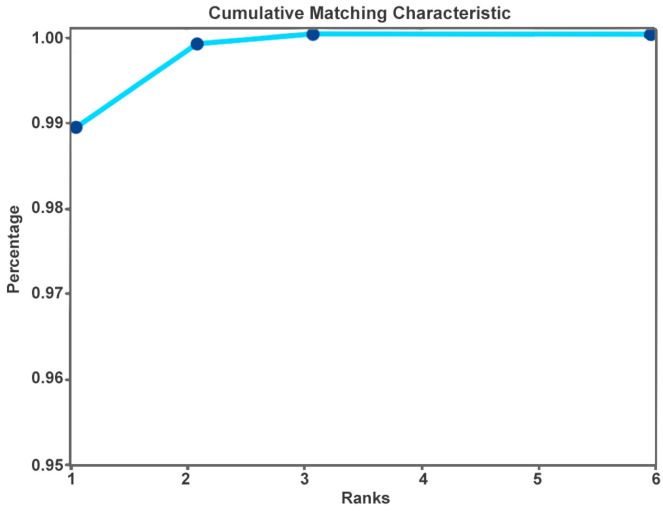

Fig. 7. CMC curve depicting the recognition performance for the first five ranks.

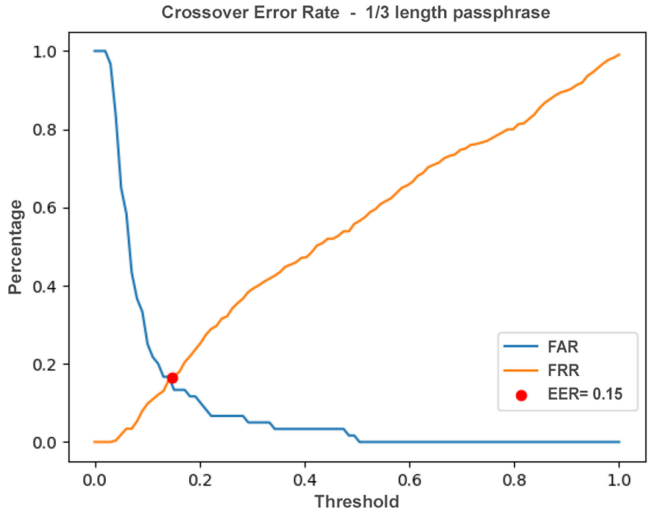

Fig. 8. FAR/FRR curves for 1/3 length passphrase. Equal error rate value is 0.15%.

TABLE I
RESULTS OF THE THREE MAIN EXPERIMENTS

	First Experiment	Second Experiment	Third Experiment
Train Passphrase	First sentence	First sentence	First sentence
Test Passphrase	First sentence	Second sentence	Third sentence
Features x block	177		
Sample size	19824		
Train /Test size	922/230	922/1152	922/1152
DNN Configuration	19824 - 10000 - 5000 - 2000 - 48		
Recognition Rate %	98.8	98.1	97.7

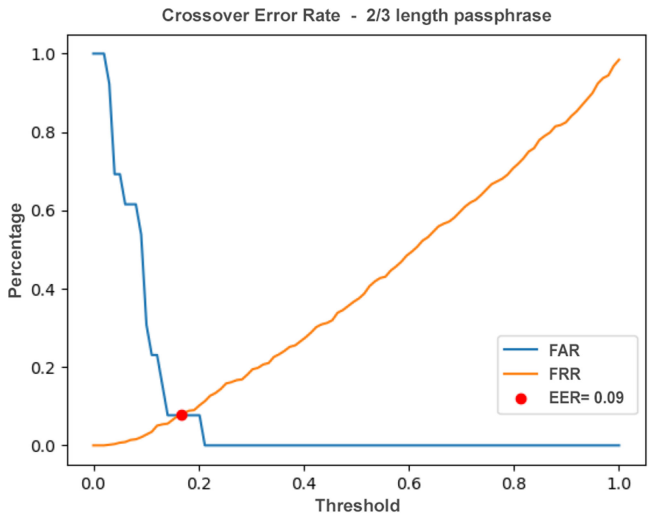

Fig. 9. FAR/FRR curves for 2/3 length passphrase. Equal error rate value is 0.09%.

In this case, the resulting accuracy was 97.7%, still a very high value. Finally, we wanted to verify the weight of the dynamic component of the feature vector (peculiar to the proposed approach) by providing in input to the system a fake video made up of a sequence of the same image of a genuine subject. The resulting accuracy was a mere 57%, proving the high sensibility of the method to the way the face locally moves and not just to its shape.

A quantitative summary of the aforementioned results, achieved in each of the three experiments, is found in Table I. The experiments proved that facial dynamics can not only represent a highly salient descriptor of an individual, but also their results are intrinsically much more difficult to forge by malicious users (compared to any static face descriptor) since they are time dependent in a way that is peculiar to whom is speaking, due to his/her anatomical and behavioral characteristics [46].

The temporal component of the facial descriptor, indeed, provides both uniqueness and reliability to the feature vector, as proved by the very low FAR and high GAR shown in the plots of Fig. 5 and 6. Furthermore, an implicit (yet indirect) liveness test is performed during acquisition, since only a time-changing face appearance could be a valid input to the processing pipeline.

As highlighted in the introduction, another aspect deserving investigation, since it is tightly related to the trustworthiness of the identification procedure, is the length of the sentence necessary for the recognition to take place successfully.

The experiment carried out to this aim consists of measuring the performance of a model trained on a complete passphrase and tested on a partially pronounced passphrase. The complete passphrase used for training is again "My face is my key to this system." Therefore, to obtain partial-passphrase video clips, the original complete-passphrase videos have been edited in two versions ("My face is my key" and "My face is") whose lengths are approximately 2/3 and 1/3 of the full length.

The aforementioned editing of video samples has been performed by means of a script. With this configuration, the percentage obtained for recognition by pronouncing one-third of the passphrase is 89.81%, while that obtained by pronouncing two-thirds of the passphrase is 97.79%. Remembering that the result obtained by testing the same model on the complete ordered sequence is 99.49%, it is useful to analyze the graphs relating to the crossover error rate shown in Fig. 8 and 9 and compare them with that of Fig. 6.

TABLE II
SYSTEM TRUSTWORTHINESS FOR A THRESHOLD SET AT 0.88

Test Passphrase	Recognition Rate (RR)	Equal Error Rate (EER)	Genuines not Accepted	Impostors Accepted
1/3 length	89.81 %	0.15 %	10 out of 100	0
2/3 lenght	97.79 %	0.09 %	2 out of 100	0
full lenght	99.49 %	0.05 %	1-0 out of 100	0
No motion	57.3 %	0.45 %	43 out of 100	0

Unsurprisingly, the minimum EER value is obtained with the full passphrase (0.05), though 2/3-length passphrase achieve EER = 0.09 and 1/3-length passphrase still reach a reasonably good EER = 0.15. It is also easy to note that the FAR curve becomes more and more segmented according to the length of the pronounced sentence. This is due to the increase in system performance, which has a lower number of false positives and, therefore, has a "step" trend in the FAR.

Furthermore, another sign indicating the reliable behavior of the system is the steepness, with which the FAR curve tends to zero. Also, as regards the FRR, the slope of the curve provides valuable indications. In the first case, the growth is immediate, and even for low threshold values, the percentage of FRR increases considerably. By increasing the length of the passphrase, in contrast, the FRR values remain lower, and consequently, the ability of a system to recognize genuine subjects increases.

These results demonstrate that the dynamic characteristic extracted on a shorter subsequence is sufficient to obtain discrete performance, but the use of a longer passphrase leads to a considerable increase in the robustness of the system. Overall, the experiments confirm the validity of the proposed approach and its trustworthiness with regard to person identification in challenging industrial contexts, as summarized in Table II.

VI. CONCLUSION

In this article, we presented a method for identifying/authenticating persons operating within a three-level IIoT environment by means of a dynamic biometric signature. Facial motion around the mouth region captured by edge-level devices was processed at fog nodes for extracting discriminant features through the LBP-TOP local spatial–temporal descriptor and further matched to a reference gallery via a deep network implemented on the cloud.

The proposed approach and the related edge–fog–cloud architecture proved to be highly effective in increasing the trustworthiness of the IIoT environment, thanks to the intrinsic difficulty in forging such a time-dependent descriptor. The experiments conducted on a custom-built database resulted in state-of-the-art recognition accuracy reaching 98.7% at rank 1 and showing high robustness to the way the passphrase is pronounced if the subject is genuine, yet reliable rejection of imposters even with a low decision threshold.

Future research will concern experiments aimed at assessing the efficiency of the three-level architecture compared to a more conventional solution based on local processing. An extension of this article could also include the audio component of the speech samples for implementing a bimodal biometric system to further improve both accuracy and reliability of the proposed method.

REFERENCES

[1] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[2] C. Perera, C. Liu, and S. Jayawardena, "The emerging Internet of Things marketplace from an industrial perspective: A survey," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 4, pp. 585–598, Dec. 2015.

[3] Z. Xia, L. Jiang, X. Ma, W. Yang, P. Ji, and X. Xiong, "A privacy-preserving outsourcing scheme for image local binary pattern in secure Industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 629–638, Jan. 2020.

[4] M. Ferrag, L. Maglaras, H. Janicke, J. Jiang, and L. Shu, "Authentication protocols for Internet of Things: A comprehensive survey," *Secur. Commun. Netw.*, vol. 3, no. 4, pp. 585–598, 2017.

[5] U. Qureshi, G. Hancke, T. Gebremichael, U. Jennehag, S. Forsström, and M. Gidlund, "Survey of proximity based authentication mechanisms for the Industrial Internet of Things," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc.*, May 2018, pp. 5246–5251.

[6] M. Hossain and G. Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT) enabled framework for health monitoring," *Comput. Netw.*, vol. 101, pp. 192–202, 2016.

[7] A. Castiglione, K. R. Choo, M. Nappi, and S. Ricciardi, "Context aware ubiquitous biometrics in edge of military things," *IEEE Cloud Comput.*, vol. 4, no. 6, pp. 16–20, Nov. 2017.

[8] H. Hamidi, "An approach to develop the smart health using Internet of Things and authentication based on biometric technology," *Future Gener. Comput. Syst.*, vol. 91, pp. 434–449, 2019.

[9] K. Choo, S. Gritzalis, and J. Park, "Cryptographic solutions for industrial Internet-of-Things: Research challenges and opportunities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3567–3569, Aug. 2018.

[10] Z. Guo, N. Karimian, M. Tehranipoor, and D. Forte, "Hardware security meets biometrics for the age of IoT," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2018, pp. 1318–1321.

[11] A. Das, M. Wazid, N. Kumar, A. Vasilakos, and J. Rodrigues, "Biometrics-based privacy-preserving user authentication scheme for cloud-based Industrial Internet of Things deployment," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4900–4913, Dec. 2018.

[12] M. Hossain, G. Muhammad, S. Rahman, W. Abdul, A. Alelaiwi, and A. Alamri, "Toward end-to-end biomet rics-based security for IoT infrastructure," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 44–51, Oct. 2016.

[13] X. Li, J. Niu, M. Bhuiyan, F. Wu, M. Karuppiah, and S. Kumari, "A robust ECC-based provable secure authentication protocol with privacy preserving for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3599–3609, Aug. 2018.

[14] W. Abdul, Z. Ali, S. Ghouzali, B. Alfawaz, G. Muhammad, and M. Hossain, "Biometric security through visual encryption for fog edge computing," *IEEE Access*, vol. 5, pp. 5531–5538, 2017.

[15] K. Munir and L. Mohammed, "Biometric smartcard authentication for fog computing," *Int. J. Netw. Secur. Appl.*, vol. 10, no. 6, pp. 35–45, 2018.

[16] P. Hu, H. Ning, T. Qiu, Y. Zhang, and X. Luo, "Fog computing based face identification and resolution scheme in Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1910–1920, Aug. 2017.

[17] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in Internet of Things," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1143–1155, Oct. 2017.

[18] P. Hu, H. Ning, T. Qiu, Y. Xu, X. Luo, and A. Sangaiah, "A unified face identification and resolution scheme using cloud computing in Internet of Things," *Future Gener. Comput. Syst.*, vol. 81, pp. 582–592, 2018.

[19] N. Karimian, P. Wortman, and F. Tehranipoor, "Evolving authentication design considerations for the Internet of Biometric Things (IoBT)," in *Proc. 11th IEEE/ACM/IFIP Int. Conf. Hardware/Softw. Codesign Syst. Synthesis*, Oct. 2016, pp. 1–10.

[20] Y. Lu, K. Gu, and S. He, "Research on visual speech recognition based on local binary pattern and stacked sparse autoencoder," in *Human Systems Engineering and Design*, T. Ahram, W. Karwowski, and R. Taiar, Eds. New York, NY, USA: Springer, 2019, pp. 1082–1087.

[21] S. Nainan and V. Kulkarni, "Lip tracking using deformable models and geometric approaches," in *Information and Communication Technology for Intelligent Systems*, S. C. Satapathy and A. Joshi, Eds. Singapore: Springer, 2019, pp. 655–663.

[22] G. Chetty and M. Wagner, "Automated lip feature extraction for liveness verification in audio-video authentication," in *Proc. Image Vis. Comput.*, 2004, pp. 17–22.
[23] M. I. Faraj and J. Bigun, "Motion features from lip movement for person authentication," in *Proc. 18th Int. Conf. Pattern Recognit.*, Aug. 2006, vol. 3, pp. 1059–1062.
[24] M.-I. Faraj and J. Bigun, "Audio-visual person authentication using lip-motion from orientation maps," *Pattern Recognit. Lett.*, vol. 28, no. 11, pp. 1368–1382, 2007.
[25] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2879–2891, Oct. 2006.
[26] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 133–137, Jan. 2009.
[27] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[28] A. Y. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567
[29] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, vol. 5, pp. 3771–3779.
[30] X. Liu and Y. Cheung, "Learning multi-boosted HMMs for lip-password based speaker verification," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 2, pp. 233–246, Feb. 2014.
[31] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.
[32] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of face spoofing using visual dynamics," *IEEE Trans. Informat. Forensics Secur.*, vol. 10, no. 4, pp. 762–777, Apr. 2015.
[33] K. Mendhurwar, S. Mudur, and T. Popa, "Time series matching for biometric visual passwords," in *Proc. ACM SIGGRAPH Posters*, 2017, Art. no. 87.
[34] Y. Yuan, J. Zhao, W. Xi, C. Qian, X. Zhang, and Z. Wang, "SALM: Smartphone-based identity authentication using lip motion characteristics," in *Proc. IEEE Int. Conf. Smart Comput.*, May 2017, pp. 1–8.
[35] S. Yoo *et al.*, "Guaranteeing real-time services for industrial wireless sensor networks with IEEE 802.15.4," *IEEE Trans. Ind. Electron.*, vol. 57, no. 11, pp. 3868–3876, Nov. 2010.
[36] D. Striccoli, G. Boggia, and L. A. Grieco, "A Markov model for characterizing IEEE 802.15.4 MAC layer in noisy environments," *IEEE Trans. Ind. Electron.*, vol. 62, no. 8, pp. 5133–5142, Aug. 2015.
[37] P. Lindgren, J. Eriksson, M. Lindner, A. Lindner, D. Pereira, and L. M. Pinho, "End-to-end response time of IEC 61499 distributed applications over switched ethernet," *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 287–297, Feb. 2017.
[38] A. Willig, "Recent and emerging topics in wireless industrial communications: A selection," *IEEE Trans. Ind. Informat.*, vol. 4, no. 2, pp. 102–124, May 2008.
[39] Y. Zhao, L. T. Yang, and J. Sun, "Privacy-preserving tensor-based multiple clusterings on cloud for Industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2372–2381, Apr. 2019.
[40] D. He, M. Ma, S. Zeadally, N. Kumar, and K. Liang, "Certificateless public key authenticated encryption with keyword search for Industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3618–3627, Aug. 2018.
[41] L. Lyu, J. C. Bezdek, X. He, and J. Jin, "Fog-embedded deep learning for the Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4206–4215, Jul. 2019.
[42] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. I511–I518.
[43] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.
[44] T. Ojala, M. Pietikainen, and T. Maenpaa, "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, Jul. 2002.
[45] G. Zhao, M. Pietikainen, and T. Maenpaa, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007.
[46] M. Nappi, S. Ricciardi, and M. Tistarelli, "Deceiving faces: When plastic surgery challenges face recognition," *Image Vis. Comput.*, vol. 54, pp. 71–82, 2016.

Aniello Castiglione (Member, IEEE) received the Ph.D. degree in computer science from the University of Salerno, Fisciano, Italy, in 2007.

He is currently with the Department of Science and Technology, University of Naples Parthenope, Naples, Italy. He authored more than 200 papers in international journals and conferences. He has served in the organization of more than 200 international conferences. He served as a Reviewer for approximately 100 international journals and the Managing Editor for two ISI-ranked international journals. He was a Guest Editor for around 20 special issues and served as an Editor on more than ten Editorial Boards of international journals. One of his papers (published in the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING) was selected as the "Featured Article" in the "IEEE Cybersecurity Initiative" in 2014. His current research interests include information forensics, digital forensics, security and privacy on distributed systems, steganography, communication networks, applied cryptography, and sustainable computing.

Michele Nappi (Senior Member, IEEE) received the Laurea degree (*cum laude*) in computer science from the University of Salerno, Fisciano, Italy, in 1991, the M.Sc. degree in information and communication technology from the Istituto Internazionale per gli Alti Studi Scientifici "E.R. Caianiello," Vietri sul Mare, Italy, in 1991, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Padua, Italy, in 1997.

He is currently a Full Professor of Computer Science with the University of Salerno. He is also a Team Leader of the Biometric and Image Processing Lab. He has coauthored more than 200 papers in international conferences, peer-reviewed journals, and book chapters in his research interests. His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human–computer interaction, and virtual reality/augmented reality.

Prof. Nappi was a member of the International Association for Pattern Recognition. 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.

Stefano Ricciardi (Member, IEEE) was born in Naples, Italy. He received the B.Sc. degree in computer science, the M.Sc. degree in informatics, and the Ph.D. degree in sciences and technologies of information, complex systems, and environment from the University of Salerno, Fisciano, Italy, in 2002, 2004 and 2015 respectively.

He has been Co-Founder/Owner of a videogame development company. He is currently an Assistant Professor with the Department of Biosciences, University of Molise, Campobasso, Italy. He served as an External Expert of the European Commission's Research Executive Agency for the Horizon-2020 research and innovation program. He served as a Reviewer for several international journals and coauthored more than 80 research papers, including international journals, book chapters, and conference proceedings. His current main research interests include biometry, virtual and augmented/mixed reality, haptics systems, and human–computer interaction.

Dr. Ricciardi is a member of the Italian Group of Researchers in Pattern Recognition, International Association for Pattern Recognition.