IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 23, 2019, accepted May 1, 2019, date of publication May 17, 2019, date of current version May 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917451

Near Real-Time Three Axis Head Pose Estimation Without Training

ANDREA F. ABATE[1], (Member, IEEE), PAOLA BARRA[1], CARMEN BISOGNI[1], MICHELE NAPPI[1], (Senior Member, IEEE), AND STEFANO RICCIARDI[2], (Member, IEEE)

[1]Department of Informatics, University of Salerno, 84084 Fisciano, Italy
[2]Department of Biosciences and Territory, University of Molise, 80523 Campobasso, Italy

Corresponding author: Stefano Ricciardi (stefano.ricciardi@unimol.it)

This work was partly funded by the COSMOS (COntactlesS Multibiometric mObile System) PRIN (Research Project of National Interest) grant.

ABSTRACT Head pose estimation methods evaluate the amount of head rotation according to two or three axes, aiming at optimizing the face acquisition process, or extracting neutral-pose frames from a video sequence. Most approaches to pose estimation exploits machine-learning techniques requiring a training phase on a large number of positive and negative examples. In this paper, a novel pose estimation method that exploits a quad-tree-based representation of facial features is described. The locations of a set of landmarks detected over the face image guide its subdivision into smaller and smaller quadrants based on the presence or lack of landmarks within each quadrant. The proposed pose descriptor is both effective and efficient, providing accurate yaw, pitch and roll axis estimates almost in real-time, without need for any training or previous knowledge about the subject. The experiments conducted on both the BIWI Kinect Head Pose Database and the challenging automated facial landmarks in the wild dataset, highlight a pose estimate precision exceeding the state-of-the-art with regard to methods not involving training and machine learning approaches.

INDEX TERMS Biometrics, face recognition, image analysis.

I. INTRODUCTION

Face is currently considered one of the most diffused biometrics as well as one of the most accepted for both person authentication and identification, mainly because it is possible to capture it without direct contact by simply using a digital camera. It is a well-known fact that face's potentially high discriminant power can be significantly affected by subject's pose, and this is particularly true for uncontrolled/unsupervised acquisition often occurring for face capture at a distance and/or in-the-wild. The impact that a not-neutral pose may possibly have on the recognition accuracy depends on the intrinsic robustness of the feature-extraction algorithm considered, but it is always proportional to the overall head rotation extent with regard to each rotation axis and to the combination of rotations.

More precisely, for the same algorithm it is not unusual to have different values of recognition error for a head rotation involving, for instance, mainly the yaw axis, instead of mainly the roll axis or a combination of all three axis. It is worth noting that along with actual head rotations, also ''apparent'' head rotations due to the head-camera angles may have a negative impact on the subsequent processing stages.

This kind of situation is very common when face is captured at a distance by unattended imaging devices, such as the surveillance cameras typically present inside many buildings as well as in most urban contexts. In these scenarios, from the one side there is a high chance that in a randomly selected image the acquired face will not be in a neutral pose, but, on the other side, there is also a high chance that in at least one frame of the captured sequence the face will be close to the neutral pose.

The capability of selecting that optimal frame, possibly in real time, could actually improve the recognition performance. Moreover, in a multi-biometric system, by knowing the degree of head rotation affecting the acquired image, it would be possible to adjust the weight of this descriptor accordingly, thus optimizing the result in a score-level data-fusion strategy.

This paper describes a Head Pose Estimation (HPE) method capable of quickly measuring rotations of the head in a single intensity image according to yaw, pitch and

The associate editor coordinating the review of this manuscript and approving it for publication was Kien Nguyen.

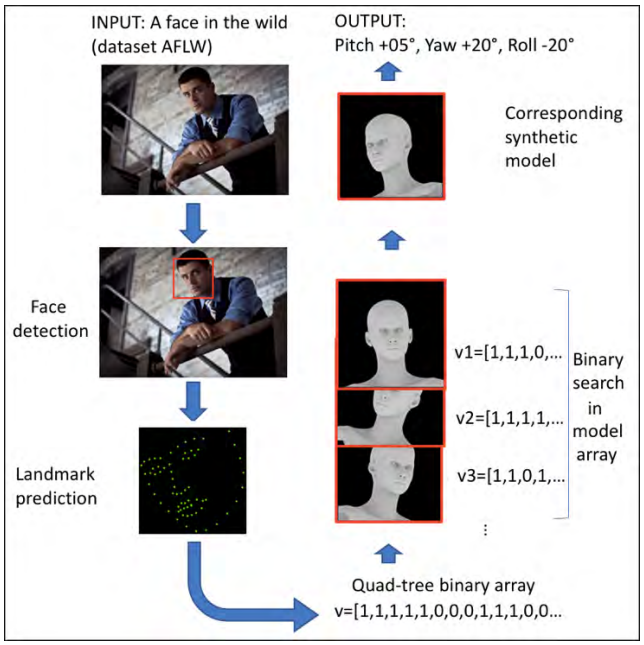

FIGURE 1. Overall workflow of proposed head pose estimation method.

roll axes. The proposed approach, schematically outlined in Figure 1, exploits a novel (for the HPE context) version of a well-known data structure, the quad-tree, expressly adapted to store a landmark-based representation of face orientation.

By measuring the distance between this representation and a reference model is possible to obtain a discrete estimate of actual head orientation in the three-dimensional space. The inherent efficiency of the quad-tree descriptor as well as the good measurement accuracy achieved, have been both confirmed by the experiments conducted on two databases, the BIWI Kinect Head Pose Database and the Automated Facial Landmarks in the Wild (AFLW) dataset. The proposed methodology, not involving any learning/training stage and the related optimization effort, not requiring to know anything in advance about the subject in input, is able to deliver MAE values lower than same-category state-of-the-art methods. We explicitly aim at improving performance without exploiting machine-learning techniques and related NN architectures, as well as without requiring depth-data and related hardware.

To resume its main contributions, the proposed HPE method:

- *does not require any learning/training stage;*
- *works on a single intensity image and does not require depth/3D data nor dedicated hardware;*
- *is capable of near real-time performance on ordinary single CPU hardware platform, thanks to the efficient quad-tree based pose representation;*
- *features an average pose estimation precision exceeding the state-of-the-art (with regard to methods only exploiting 2D images and not requiring a learning/training stage).*

The rest of the work is organized as follows. Section 2 takes up the main contributions related to the issue of estimating and normalizing facial pose. Section 3 contains a detailed description of proposed approach. Section 4 resumes the results of the experiments conducted to assessing the performance of proposed method. Finally, section 5 concludes summarizing the work done and providing guidance for future research.

II. RELATED WORKS

Head/face pose estimation is an active research topic for the computer vision community since almost three decades, as reported by Murphy-Chutorian and Trivedi in their survey [1] and more recently in [2]. There is no surprise, indeed, that a wide number of methods and algorithms has been proposed through the years. However, with regard to the aforementioned aims of the present work, it is possible to classify available approaches according to four main categories, depending on whether they work on 2D (intensity) or 3D (depth) images and whether they exploit some kind of training step (typically involving machine learning techniques) or not (i.e. they can work on completely unknown subjects).

It has to be remarked that the requirement of 3D (depth) data, implies the usage of a specialized sensor for capturing the subject, thus limiting the practical applications of such methods due to the limited subject-camera operative distance typical of this kind of equipment which are generally not suited to outdoor acquisition. Moreover, for 3D methods, head pose estimation from conventional video stream/footage is not possible. For these reasons, methods using both 2D and 3D image (intensity+depth) can be assimilated into the general ''3D'' category since they share the same operational limitations of the ''pure 3D'' methods. Though the proposed method works on intensity images without requiring any learning stage, we decided to include both 3D methods and/or training-based methods to allow a thorough comparison with the best available solutions to the head pose-estimation problem.

A. 2D METHODS WITH LEARNING/TRAINING

The largest category of pose estimation methods among those considered include approaches working on 2D images and involving machine learning techniques in general, with particular regard to deep-networks/convolutional-networks architectures. A convolutional neural network (CNN) is used indeed in [3] to project face images onto a low-dimensional pose-space, whereas a combination of CNN and adaptive gradient provide the best pose-estimation accuracy according to [4]. The method proposed in [5] fuse the hidden layers of a deep CNN via an additional CNN and a multi-task learning algorithm working on the fused features. DNN based multi-task learning is also used in [6] to learn shared features from low-res intensity images, while dictionary-learning and a classifier based on sparse representation are exploited in [7] for improving pose classification robustness.

Multi-task learning is also used in [8] since the authors aim at real-time performance in combined face detection and pose estimation by means of a multi-CNN cascade architecture. Continuous regression by means of a probabilistic framework is proposed by Aghajanian and Prince [9] to address pose estimation in uncontrolled conditions. Similarly, in [10] a CNN is trained over a synthetic dataset to learn head features by the procedurally annotated head pose and solve the regression problem. A mixture of tree-structures part models is exploited in [11] providing high accuracy even when trained on a limited number of examples, while a multi loss network using image intensities to estimate head pose Euler angles is described in [12]. Heatmap-CNN regressors are learnt in [13] by training on face's visibility, fiducials and 3D-pose to achieve key-points estimation and pose prediction. Multiple region-based classifiers are learned by means of FlExible GrAph-guided Multi-Task Learning in [14] to address head pose estimation in multiple-cameras monitored environments.

In [15], DNN for head pose estimation produce initialization shape according to two different initialization schemes, by either projecting a mean 3D face shape to the test image or searching nearest neighbor shapes from the training set based on head pose distance. Support Vector Regression (SVR) applied to histogram of oriented gradients (HoG) feature is used for head pose estimation on low-resolution images in [16].

Peng *et al.* [17] propose a coarse-to-fine framework exploiting a unit circle to model the coarse layer and a 3-sphere to model the fine layer within a generative approach to handle multiple head variations. A coarse-to-fine approach is also behind the work described in [18], where joint hierarchical head pose estimation and landmark detection is achieved by the learning system exploring both global and local CNN features.

B. 3D METHODS WITH TRAINING

Various training-based methods exploiting the additional information provided by 3D data have been proposed.

Face range data are used in [19] to address pose estimation through regression by means of a random forest framework. The proposed method addresses the need of training the regressor on labeled data, by training it only on synthetic examples not requiring laborious and error-prone annotations. Microsoft Kinect built-in depth sensor is exploited for real-time head pose estimation in [20] by means of a novel viewpoint invariant triangular surface patch (TSP) descriptor, mapping the shape of face's 3D surface into a triangular region and matching it to gallery TSPs.

Another approach involving the Kinect camera is represented by [21], where Viola Jones face detector (in frontal and profile versions) is used to locate the face in the RGB image, with size and position of the search window determined by means of the depth image. The head pose is then inferred from appearance-based features, extracted from both the face's depth and RGB images, using SVM regressors.

C. 2D METHODS WITHOUT TRAINING

Another category of methods not requiring any previous learning or training stage, relies solely on specific descriptors and metrics to estimate 3D face orientation from 2D images.

To this aim, in [22], accurate pose estimation is addressed through a multi-level structured hybrid forest (MSHF). Head boundary is obtained from patches classified as either belonging to head region or to the background, then, selected patches sub-regions are used to develop the MSHF for head pose estimation. Gaussian mixture of locally-linear mapping (GLLiM) is the regression technique on which the approach described in [23] is based. More recently in [24], the same authors propose to learn with both head-pose parameters and bounding-box-to-face alignments, such that, at runtime both the head-pose angles and bounding-box shifts are predicted. This approach ensures that the predicted bounding-box-to-face alignments are similar with those used for training. Consequently, background variations have minimal influence on the observed feature vector from which the pose is being predicted. An approach aimed to provide fast head yaw/pitch estimation is proposed in [25]. It is based on an expressly adapted version of quad-tree to represent facial landmarks. By comparing this descriptor with previously stored templates, this method is able to provide a rough estimate of face rotations in a small amount of time. In [26], Diaz-Chito propose to combine HoG features and generalized discriminative common vectors within a continuous local regression approach to achieve low errors in head pose estimation.

D. 3D METHODS WITHOUT TRAINING

In the last category, fall methods exploiting 3D information but not based on machine-learning techniques.

3D pose of an unknown subject is estimated in [27] by finding nose shape in input range image and then using a GPU optimized generative algorithm to evaluates many pose hypotheses in parallel. In [28], Hough transform is applied to central profile, a unique characteristic curve defined over 3D face surface, to find the symmetry plane by means of a voting strategy. In [29] a framework based on random forests trained by the SIFT-HOG features is used to approach pose estimation as a regression problem. To cope with extreme poses and partial occlusions, a weighted-vertices morphable face model is registered to the 3D data captured by commodity depth cameras by combining particle swarm optimization and the iterative closest point algorithm in [30]. Finally, Darby *et al.* [31] explore real-time head pose estimation capability of Microsoft Kinect v2 High Definition Face Tracking (HDFT) component, evaluating the sensor's rotational and translational precision.

The head pose estimation approach proposed in this paper belongs to the aforementioned category of ''2D methods without learning'' and represents an evolution of [25] extending the pose estimation to three axis and significantly improving both estimate accuracy and speed.

III. PROPOSED METHOD

From an operational point of view, the proposed method aims at estimating a subject's pose on three axis, with an approximation of 5° , starting from a single intensity image or frame (in a video sequence), in near real-time. To achieve these results an effective and efficient processing pipeline is devised, made up of two main phases:

1. *the building of a reference pose-gallery, performed only once to create a gallery of quad-tree based descriptors of each of the head poses present in the angular range and step considered on each rotation axis*;
2. *the evaluation of the input image and its associated quad-tree based pose descriptor to find a matching pose in the reference pose-gallery and therefore the related pitch, yaw and roll values*;

A. REFERENCE POSE-GALLERY

The present method is optimized for estimating head pose in a range of ±45° for Yaw, ±30° for Pitch and ±20° for Roll discretized at 5° of angular step, accounting for a total of 2223 head poses which represent the method's discrete search space.

These ranges have been selected to reduce the search-space in the light of practical considerations such as the statistical prevalence of the yaw rotation values compared to pitch and roll values as well as the working limits of most face detection and facial landmark localization algorithms. Similar reasoning applies to the angular granularity of 5° adopted, which is reasonably small, however still visually significant. A smaller angular step would be barely noticeable, though it would have a significant impact on the efficiency of the method. It is worth to note that the proposed approach has no inherent limitations in terms of angular range and could work on large poses as well. Actually, apart from the considerations made above, the limit is more in the landmark localization algorithm we used, that provides optimal results within limited angular ranges and suggested the current number of poses distributed on the three axes. On the other side, there is a practical trade-off between pose-range and quad-tree size, that in turn has an impact on computing time. The required 2223 head poses can be obtained by synthetic generative methods, involving the procedural rendering of a 3D face (s) either modelled with CGI tools or captured via 3D scanning of a real subject. More in detail, a synthetic 3D head model is procedurally rotated and rendered in real-time across the chosen angular ranges and according to a 5° step for a total of 2223 poses. Within this procedure, there are three main time-consuming steps: rendering of head pose, landmarks association to the rendered image and quad-tree representation as a 1D vector; the bottleneck being mainly the landmarks predictor which on an average laptop may reach 30 Fps. In the end the overall procedure may require a few minutes on ordinary hardware and is performed only once. The procedural pose generation has the additional advantage of providing an implicit annotation of head pose angles, which will be valuable, at a later time, for the objective assessment of method's estimation accuracy. As shown in section V, both the pose generation methodologies cited above have been investigated in the experiments conducted. As an example, the synthetic head model depicted in Figure 2, was used to build one of the reference pose-galleries by procedurally rotating it with respect to pitch, yaw and roll axis.

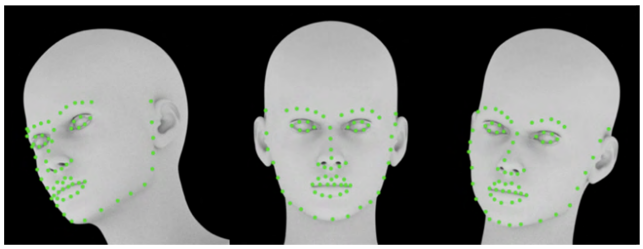

FIGURE 2. Simulated head pose variations involving three axis of rotation and the associated landmarks displayed in green.

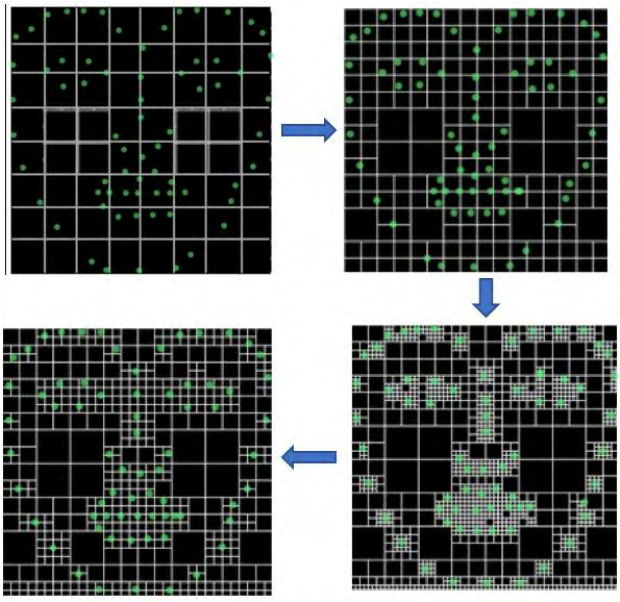

FIGURE 3. Example of four subsequent quad-tree subdivision steps from the coarser to the finer level.

B. FACE DETECTOR E LANDMARK PREDICTOR

For each head pose image, the first two steps of the processing pipeline perform whole face detection and, subsequently, facial landmarks localizations. The fast and robust Viola Jones algorithm [33] has been used to this purpose. Once the rectangular region containing a face has been detected, reliable facial key-points have to be found to generate a more compact face pose model, suited to be represented by the proposed quad-tree based descriptor. The localization of 68 2D facial landmarks is therefore rapidly performed through the algorithm described in [34] by means of an ensemble of regression trees and resulting in a feature vector of size 68x2.

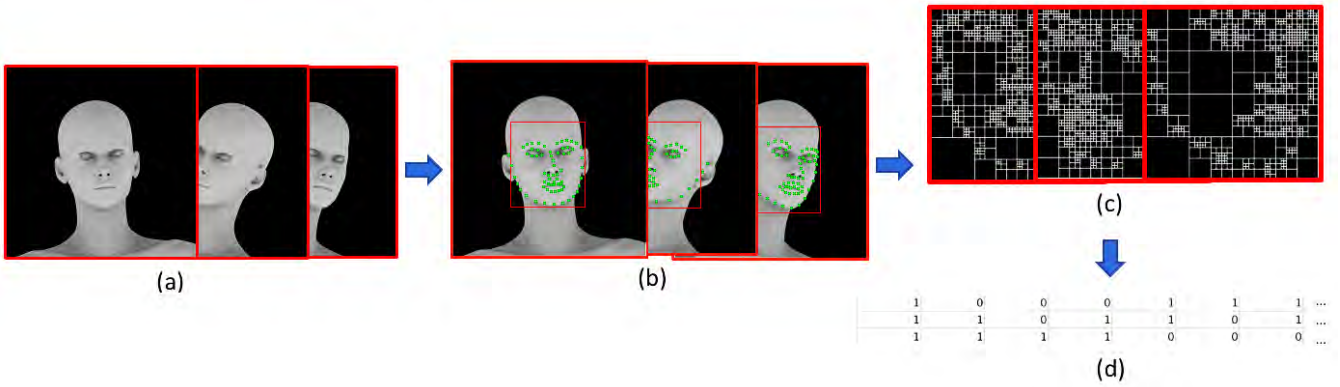

FIGURE 4. (a) Synthetic poses (b) Face detection and Landmark prediction (c) Quad-tree decomposition (d) Binary-tree array.

C. QUAD-TREE BASED HEAD POSE DESCRIPTOR

From this point on, the original image can be discarded and its associated landmarks-based description is used instead.

The following step, indeed, is to represent this description through a quad-tree [35].

This particular kind of unbalanced tree is often used for image representation by dividing the image into smaller and smaller quadrants based on the presence of information within each quadrant. In the proposed method, the root of the tree contains the entire face, that can always be divided into four quadrants due to the presence of the landmarks, which represent the relevant information. Each quadrant created is then subdivided into four quadrants or none, depending on whether or not at least one landmark is present in the quadrant. The subdivision continues in this way up to quadrants of 4x4 pixel size.

The process is illustrated in Figure 3 where both landmarks and local subdivisions are graphically represented and in Figures 4 and 5 in terms of the processing flow, respectively from a visual and a logical point of view.

Along with the subdivision process, a tree-vector is built to represent the structure of the specific tree associated to a given pose. This tree vector always contains 1365 nodes if the number of generations in the tree is set to 6, a value that has been found to be adequate to the purpose. The tree-vector is a binary vector, in which each element is either 1 (indicating the existence of the node and, therefore, that the parent quadrant of the node has been subdivided), or 0 (indicating the non-existence of the node and, therefore, that the parent quadrant of the node has not been subdivided). The resulting complete binary tree represents the distribution of the landmarks in the image.

The pose of the subject in the image is indeed strongly linked to this tree that is unbalanced on one side rather than on the other and that creates or does not create child nodes in relation to the position of the landmarks in the image and therefore to the rotation of the subject along the three axes.

This also explains the need for a complete tree, since associating to each pose a fixed-length vector, makes possible to organize all the poses in a sparse matrix (that will represent

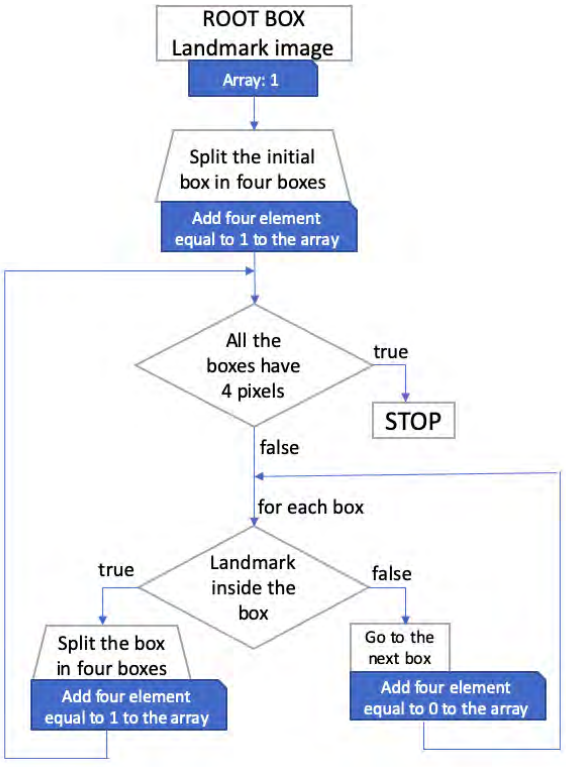

FIGURE 5. Quad Tree generation workflow.

our reference pose-gallery) and perform pose matching. The accuracy of the method depends from the capability of the quad-tree based representation of the input-subject to retain 3D orientation information according to the 2D landmarks coordinates. So, it is reasonable to expect that local tree-subdivision process could be somewhat affected by subject appearance and/or expression, yet the experiments have shown that by using 68 landmarks the impact of these variations is limited.

IV. REFERENCE DATASETS

Two reference datasets have been used for testing the proposed method and assessing its performance compared to the

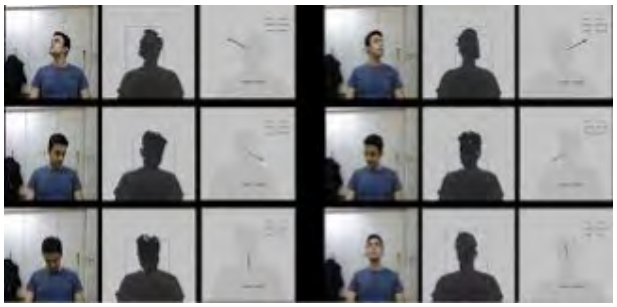

FIGURE 6. Samples from the BIWI Kinect Head Pose Database.

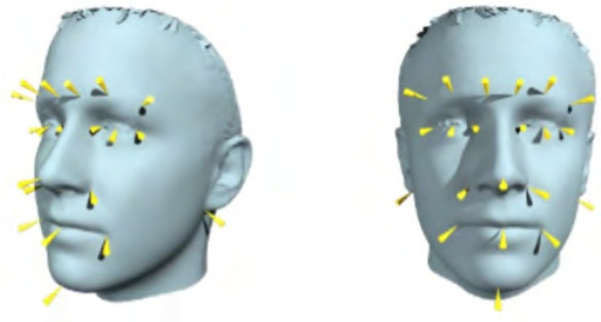

FIGURE 7. Landmarks positions provided in the AFLW Database.

state-of-the-art. The first database is the BIWI Kinect Head Pose Database originally introduced in [36]. Over 15000 rgb images captured from 20 subjects (14 males and 6 females, four of which have been acquired twice) are included, along with range data and ground truth annotations for both head position and rotation (see Figure 6). The second database is the Annotated Facial Landmark in the Wild (AFLW) database [37].

AFLW contains about 25000 sample faces gathered from Flickr, mostly rgb, featuring a wide range of poses, ages, expressions, ethnical traits, imaging and environmental conditions. Roughly sixty percent of these images depict female subjects, while the rest represents male subjects and, in limited cases, multiple faces. The images, which have not been resized or cropped, contain twenty-one facial landmarks each, manually annotated upon visibility as depicted in Figure 7.

V. DESCRIPTION OF EXPERIMENTS

Three groups of experiments, for a total of six trials, have been conducted on either the BIWI or the AFLW (Figure 8) datasets. The testbed was an i5 quad core single CPU Macbook Pro, with an integrated Intel Iris 540 GPU. No multithreading optimization has been used for the computing tasks. The first group is aimed at assessing the pose estimation accuracy of the proposed method (QT_PYR) whenever the probe image belongs to AFLW or BIWI and the reference pose-gallery is generated either from the synthetic model or from the 3D capture of a real subject from the BIWI database. This group includes three experiments. The first one is based on both BIWI and AFLW 2D images as probes and the synthetic model for generating the 2223 elements

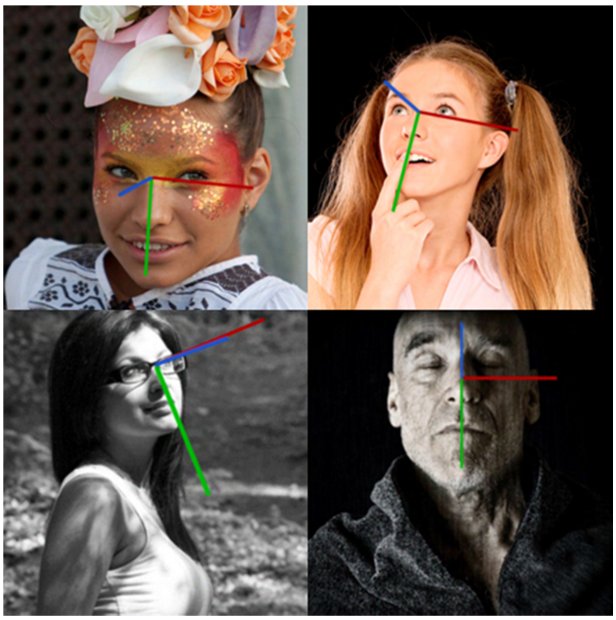

FIGURE 8. QT_PYR estimates on samples from AFLW Database.

TABLE 1. Pose estimation accuracy of proposed QT_PYR method. Probe: BIWI, AFLW - Gallery: generated via synthetic model.

QT_PYR		Pitch (deg)	Yaw (deg)	Roll (deg)	Mean (deg
BIWI	MAE	12.80	5.41	6.33	8.18
	RMSE	15.88	7.38	8.91	10.73
	STD	9.36	5.01	6.22	6.86
AFLW	MAE	7.60	7.60	7.17	7.45
	RMSE	9.58	12.42	10.39	10.80
	STD	7.51	11.66	9.14	9.44

making up the pose-gallery. The results achieved are resumed in Table 1 including MAE, RMSE and STD values for each axis and the three axes mean value as well. The two graphs in Figure 9-10 show the percentage of correctly estimated poses for a given value of estimation error, with regard to to Pitch, Yaw and Roll respectively in the BIWI and AFLW testing.

For what concerns the time-cost of the algorithm, time required for the whole process is 0.11 seconds which enables operations at 9-10 Fps. In the second experiment, the reference pose-gallery is the same of the previous one.

However, instead of building the quad-tree with regard to Pitch, Yaw and Roll, the image is previously normalized with respect to the Roll and only Pitch and Yaw are considered (QT_PY+R). More in detail, starting from the rotation of the eyes (by measuring the angular coefficient of the straight line passing through the two points represented by the external corners of the eyes) face's Roll is roughly estimated. The image is therefore normalized by rotating it so as to eliminate the Roll component. The resulting image will possibly have posing variations only referred to Pitch and Yaw axes, accounting for a total of only 207 images

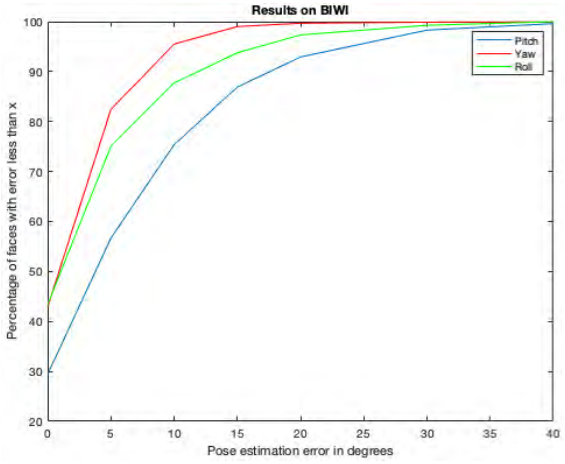

FIGURE 9. Correctly estimated poses for a given value of estimation error (BIWI).

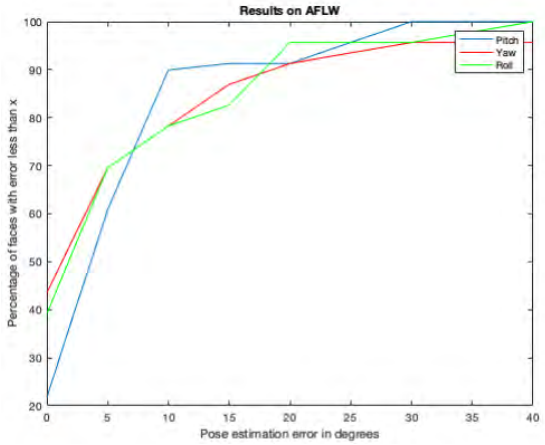

FIGURE 10. Correctly estimated poses for a given value of estimation error (AFLW).

TABLE 2. Pose estimation accuracy of proposed QT_PY+R method. Probe: BIWI, AFLW - Gallery: generated via synthetic model.

QT_PY+R		Pitch (deg)	Yaw (deg)	Roll (deg)	Mean (deg)
BIWI	MAE	14.95	6.28	4.12	8.45
	RMSE	18.11	8.58	5.42	10.70
	STD	10.19	5.82	3.48	6.50
AFLW	MAE	17.84	9.33	3.44	10.20
	RMSE	23.73	14.28	4.84	14.28
	STD	15.12	11.24	3.36	9.90

comprising the usual ranges of ±30° for Pitch, ±45° for Yaw with 5° step (see Table 2). This much smaller search-space suggests the possibility of achieving an even faster performance. The results confirm this prevision, with a time cost of 0.044 seconds or 22 Fps of operating speed, at the cost of a higher Pitch error.

Though these results do not match the typical 30 Fps requirement to allow the method to be fully considered

TABLE 3. Pose estimation accuracy of proposed QT_PYR method. Probe: BIWI - Gallery: generated from each of ten different subjects in BIWI.

Subject	Pitch (deg)	Yaw (deg)	Roll (deg)	Mean (deg)
01	9.41	4.64	5.31	6.45
02	7.84	4.19	5.78	5.94
03	9.61	7.93	14.41	10.65
04	**7.51**	**4.07**	**5.50**	5.69
05	7.17	4.90	6.61	6.23
06	7.32	4.77	5.51	5.87
07	9.35	4.25	5.52	6.37
08	7,46	7,46	5,91	6.94
09	7.58	4.13	5.59	5.76
10	7.96	5.48	6.61	6.68

TABLE 4. Comparison of proposed methods to state-of-the-art on the AFLW database. † Implemented on Nvidia GTX Titan-X GPU. †† Implemented on Nvidia GTX 1080-Ti GPU.

Methods	Training + NN required	Pitch (deg)	Yaw (deg)	Roll (deg)	Mean (deg)	Time (sec)
[5]	Yes	6,13	7,61	3,92	5.87	0,1†
[13]	Yes	**5,85**	**6,45**	8,75	7.02	0.3††
[12]	Yes	6,56	6,47	5,43	**6.15**	-
QT_PYR	No	7,60	7,60	7,17	7.45	0,11
QT_PY+R	No	17.84	9.33	**3.44**	10.20	**0,044**

TABLE 5. Comparison of proposed methods to state-of-the-art on the BIWI database. †††Implemented on Nvidia GTX Titan Black GPU.

[29]	No	8.5	8.8	7.4	8.23	0.067
[10]	Yes	**6.1**	6.0	5.7	5.94	0.76†††
[24]	No	7.65	6.06	5.62	6.44	-
[12]	Yes	6,97	**5,16**	**3,38**	5.17	-
QT_PYR	No	7.51	**4.07**	5.50	**5.69**	0.11
QT_PY+R	No	14.95	6.28	4.12	8.45	**0.044**

real-time, that goal can be easily achieved if a slightly more performing hardware is used. To further clarify the relevance of the hardware adopted, it is worth noting that some of the recent real-time CNN-based HPE methods cited in Table 4 and 5, reach this performance on workstation-class machine equipped with a high-end GPU (typically Nvidia GTX 1080-Ti or GTX Titan-X) to efficiently implement the neural network. If the most diffused GPU benchmarks are used for comparing the integrated graphics board within the notebook used in our experiments (Intel Iris Graphics 540) to the aforementioned Nvidia GTX 1080-Ti, a speed increment of 1411% is found for the latter. According to the 14**x**boost provided, the actual performance of the aforementioned GPU-enhanced real-time methods would drop of more than an order of magnitude on ordinary hardware, on which, probably, the most diffused DL environments would not work

FIGURE 11. Search for the desired pose in a sequence of frames extracted from a video interview (a) Search for an image that matches the frontal pose of the synthetic model with angles: P +00, Y +00, R +00 (b) Front-most frame in the sequence (c) Search for an image that matches the synthetic model with angles: P +10, Y +30, R +05 (d) Frame more similar to the required pose.

at all. The third experiment is based on a different pose-gallery, made up from real 3D data captured from each of the real subjects included in the BIWI database as indicated in the leftmost column of Table 3.

According to the results shown, using one of the subjects of BIWI as a pose-gallery allows to reduce errors, especially in Pitch. In particular, subject 04 appears to be the best candidate for the generation of the reference pose-gallery.

The second group of experiments is designed to compare the proposed approach to the state-of-the-art of head pose estimation algorithms. To this aim, and consistently to the categories of methods described in section II, some of the best performing training based and not-training based methods have been considered. With regard to training-based the results are shown in Table 4 and Table V. As can be seen from Table 4, on a competitive dataset like AFLW our method competes with neural networks, approaching the best error levels in Pitch $(+1, 75°)$, Yaw $(+1, 15°)$ and Roll $(1, 74°)$. In Table 5 the proposed method is compared to state-of-the-art training-based approach [12] on the same BIWI database. Though this kind of comparison would usually be considered unfair toward a method not taking advantage of learning by example, it is noteworthy that the results are very close. Indeed, with regard to Pitch, the error in our method is less than one degree higher than in [12], the Yaw error is lower for our method than more than one degree and the Roll error is higher for our method for just over two degrees. In Table 6, results from a comparison to state-of-the-art methods belonging to the same category of proposed method (2D, no training) are reported. It is worth noting that the method [24] works on a manual annotation of the faces, whereas our methods exploits a synthetic pose generation process.

Nevertheless, our method outperforms the other approaches. The accuracy of the method depends from the capability of the quad-tree based representation of the input-subject to retain 3D orientation information according to the 2D landmarks coordinates. So, while it is reasonable to expect that local tree-subdivision process could be somewhat affected by subject appearance, the experiments proved that by using 68 landmarks this impact is limited. This is confirmed by comparing the results achieved on AFLW to those achieved on BIWI. While AFLW includes a much larger number of variations, the performance obtained are close to those scored on the less-challenging BIWI, and sometimes even better.

Finally, to demonstrate further use of the method presented, we performed tests on video sequences. The aim was to search in a sequence of frames the one with the desired pose. In this case then the method has been implemented in a slightly different way (see Figure 11). Instead of using the entire reference pose-gallery, only the tree-vector of the desired pose is used as a reference. As the frames are acquired, they are processed as described in section III-B and III-C, obtaining the related tree vectors that are compared with the reference tree-vector (here the comparison is one by one).

In this way, the distance of each frame is obtained from the reference pose used. After video capture, the shortest among these distances is chosen and the required frame is the one featuring the requested head pose.

VI. CONCLUSION

We presented a head pose estimation method from a single intensity image not requiring any previous learning/training stage. The proposed method, based on a quad-tree adaptation to represent facial landmarks is able to estimating head pose with regard to pitch, yaw and roll axis with a discrete angular resolution of five degrees at an operating speed close to 10 Fps on a single CPU computing hardware.

According to the results of experiments carried out on both BIWI and AFLW reference datasets, the reported pose estimation accuracy significantly exceeds that of state-of-the-art

methods not based on training and gets very close to the best performances achieved by state-of-the-art training-based methods. The proposed quad-tree based HPE method has no inherent limitations in terms of angular range and could work on large poses as well.

We are currently working to exploit more advanced landmark predictors to fully exploit the potential of proposed method.

REFERENCES

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[2] B. Czupryński and A. Strupczewski, "High accuracy head pose tracking survey," in *Proc. Int. Conf. Active Media Technol.* Cham, Switzerland: Springer, Aug. 2014, pp. 407–420.
[3] M. Osadchy, Y. Le Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *J. Mach. Learn. Res.*, vol. 8, pp. 1197–1215, May 2007.
[4] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.
[5] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[6] B. Ahn, D.-G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robot. Auton. Syst.*, vol. 103, pp. 1–12, May 2018.
[7] H. Liao, S. Lu, and D. Wang, "Tied factor analysis for unconstrained face pose classification," *Optik*, vol. 127, no. 23, pp. 11553–11566, Dec. 2016.
[8] H. Wu, K. Zhang, and G. Tian, "Simultaneous face detection and pose estimation using convolutional neural network cascade," *IEEE Access*, vol. 6, pp. 49563–49575, 2018.
[9] J. Aghajanian and S. Prince, "Face pose estimation in uncontrolled environments," *BMVC*, vol. 1, no. 2, p. 3, Sep. 2009.
[10] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1289–1293.
[11] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2879–2886.
[12] N. Ruiz, E. Chong, and J. M. Rehg. (2017). "Fine-grained head pose estimation without keypoints." [Online]. Available: https://arxiv.org/abs/1710.00925
[13] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 258–265.
[14] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1070–1083, Jun. 2016.
[15] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. (2015). "Face alignment assisted by head pose estimation." [Online]. Available: https://arxiv.org/abs/1507.03148
[16] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar, "Estimating head pose orientation using extremely low resolution images," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, Mar. 2016, pp. 65–68.
[17] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas, "From circle to 3-sphere: Head pose estimation by instance parameterization," *Comput. Vis. Image Understand.*, vol. 136, pp. 92–102, Jul. 2015.
[18] X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local CNN features," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 642–649.
[19] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 617–624.
[20] C. Papazov, T. K. Marks, and M. Jones, "Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4722–4730.
[21] A. Saeed and A. Al-Hamadi, "Boosted human head pose estimation using Kinect camera," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1752–1756.
[22] Y. Liu, Z. Xie, X. Yuan, J. Chen, and W. Song, "Multi-level structured hybrid forest for joint head detection and pose estimation," *Neurocomputing*, vol. 266, pp. 206–215, Nov. 2017.
[23] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud, "Head pose estimation via probabilistic high-dimensional regression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4624–4628.
[24] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1428–1440, Mar. 2017.
[25] P. Barra, C. Bisogni, M. Nappi, and S. RicciardiEmail author, "Fast quadtree-based pose estimation for security applications using face biometrics," in *Proc. Int. Conf. Netw. Syst. Secur.* Cham, Switzerland: Springer, Aug. 2018, pp. 160–173.
[26] K. Diaz-Chito, J. M. Del Rincón, A. Hernández-Sabaté, and D. Gil, "Continuous head pose estimation using manifold subspace embedding and multivariate regression," *IEEE Access*, vol. 6, pp. 18325–18334, 2018.
[27] M. D. Breitenstein, D. Kuettel, T. Weise, L. van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[28] D. Li and W. Pedrycz, "A central profile-based 3D face pose estimation," *Pattern Recognit.*, vol. 47, no. 2, pp. 525–534, Feb. 2014.
[29] B. Wang, W. Liang, Y. Wang, and Y. Liang, "Head pose estimation with combined 2D SIFT and 3D HOG features," in *Proc. 7th Int. Conf. Image Graph. (ICIG)*, Jul. 2013, pp. 650–655.
[30] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3D head pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3649–3657.
[31] J. Darby, M. B. Sánchez, P. B. Butler, and I. D. Loram, "An evaluation of 3D head pose estimation using the microsoft kinect v2," *Gait Posture*, vol. 48, pp. 83–88, Jul. 2016.
[32] [Online]. Available: http://www.makehumancommunity.org
[33] 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. (CVPR)*, vol. 1, Dec. 2001, p. 1.
[34] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1867–1874.
[35] H. Samet, "The quadtree and related hierarchical data structures," *ACM Comput. Surv.*, vol. 16, no. 2, pp. 187–260, Jun. 1984.
[36] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.* Berlin, Germany: Springer, Aug. 2011, pp. 101–110.
[37] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.

ANDREA F. ABATE received the Laurea degree *(cum laude)* in computer science from the University of Salerno, Salerno, Italy, in 1991, and the Ph.D. degree in applied mathematics and computer science from the University of Pisa, Pisa, Italy, in 1998. He has been serving as an Associate Professor with the University of Salerno, since 2006. He is the Scientific Coordinator of the Computer Graphics Lab (VR_Lab), Dipartimento di Matematica e Informatica, University of Salerno. He authored many scientific papers published in scientific journals and proceedings of refereed international conferences and he is a co-editor of a book. His research interests include computer graphics, virtual reality, augmented reality, haptics, biometrics, and multimedia databases. He is a member of the IEEE Haptics Technical Committee. He is a Steering Committee Member of several international conferences. He serves as a Reviewer for international scientific journals.

PAOLA BARRA received the B.S. degree in computer science from University of Salerno and the M.S. degree in business informatics from the University of Pisa. She is currently pursuing the Ph.D. degree in computer science with the Biometric and Image Processing Laboratory (BIPLAB), University of Salerno, Italy. Her research interests include machine learning technics in facial and gait recognition, image processing, and video games development. She is a member of GIRPR/IAPR.

CARMEN BISOGNI received the B.S. and M.S. degrees *(cum Laude)* in mathematics from the University of Salerno, in 2015 and 2017, respectively. She is currently pursuing the Ph.D. degree in computer science with the Biometric and Image Processing Laboratory (BIPLAB), University of Salerno, Italy. Her research interests include applied mathematics for machine learning, biometrics, image processing, and statistical analysis. She is a member of GIRPR/IAPR.

MICHELE NAPPI was born in Naples, Italy, in 1965. He received the Laurea degree *(cum laude)* in computer science from the University of Salerno, 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, Padova, Italy. 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 (BIPLAB). His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human–computer interaction, and VR/AR. He has coauthored more than 120 papers in international conference, peer review journals and book chapters in these fields. 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.

STEFANO RICCIARDI 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. He has been a Co-Founder/Owner of a Videogame Development Team, focused on 3D sports simulations. He is currently an Assistant Professor with the Department of Biosciences, University of Molise. He has coauthored more than 70 research papers including international journals, book chapters, and conference proceedings. His current research interests include biometry, pattern recognition, virtual and augmented/mixed reality, haptics systems, and human–computer interaction. He is a member of GIRPR/IAPR. He serves as a Reviewer for several scientific journals and has been a Technical Committee Member for international conferences. He also serves as an Expert for the Research Executive Agency of the European Commission.