IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2024, accepted 22 February 2024, date of publication 26 February 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370437

RESEARCH ARTICLE

How Good Is ChatGPT at Face Biometrics? A First Look Into Recognition, Soft Biometrics, and Explainability

IVAN DEANDRES-TAME, RUBEN TOLOSANA, RUBEN VERA-RODRIGUEZ, AYTHAMI MORALES, JULIAN FIERREZ, (Member, IEEE), AND JAVIER ORTEGA-GARCIA, (Fellow, IEEE)

Biometrics and Data Pattern Analytics Laboratory—BiDA Laboratory, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Corresponding author: Ivan DeAndres-Tame (ivan.deandres@uam.es)

This study has been supported by INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER), Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2) and R&D Agreement DGGC/ UAM/FUAM for Biometrics and Cybersecurity.

ABSTRACT Large Language Models (LLMs) such as GPT developed by OpenAI, have already shown astonishing results, introducing quick changes in our society. This has been intensified by the release of ChatGPT which allows anyone to interact in a simple conversational way with LLMs, without any experience in the field needed. As a result, ChatGPT has been rapidly applied to many different tasks such as code- and song-writer, education, virtual assistants, etc., showing impressive results for tasks for which it was not trained (zero-shot learning). The present study aims to explore the ability of ChatGPT, based on the recent GPT-4 multimodal LLM, for the task of face biometrics. In particular, we analyze the ability of ChatGPT to perform tasks such as face verification, soft-biometrics estimation, and explainability of the results. ChatGPT could be very valuable to further increase the explainability and transparency of automatic decisions in human scenarios. Experiments are carried out in order to evaluate the performance and robustness of ChatGPT, using popular public benchmarks and comparing the results with state-of-the-art methods in the field. The results achieved in this study show the potential of LLMs such as ChatGPT for face biometrics, especially to enhance explainability. For reproducibility reasons, we release all the code in GitHub.

INDEX TERMS Large language models, ChatGPT, face recognition, soft biometrics, explainability.

I. INTRODUCTION

ChatGPT could be without doubt one of the most popular words in our society nowadays. ChatGPT[1] refers to an Artificial Intelligence (AI) chatbot created by OpenAI company that is capable of interacting with humans in a conversational way, making it possible to answer questions, summarize content, correct mistakes, provide suggestions, and write and debug code, among many other tasks. Since its launch in November 2022, ChatGPT has been the fastest-growing consumer application in history, reaching over 100 million monthly users just two months after launch [1]. In fact, ChatGPT has been already deployed with success in several real-world applications [2], [3]. But, what is the main reason for the success of ChatGPT and why now? In general, this has been possible thanks to the rapid advance produced in Large Language Models (LLMs) in the last years [4], [5] which offer impressive capabilities in different tasks like medicine [6], [7], [8], education [3], or coding [9], and also, the fine-tuning of the models through Reinforcement Learning from Human Feedback (RLHF)[2], improving the experience from a human perspective while interacting with them.

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

[1]https://openai.com/blog/chatgpt *(January, 2024)*

[2]https://openai.com/research/learning-from-human-preferences

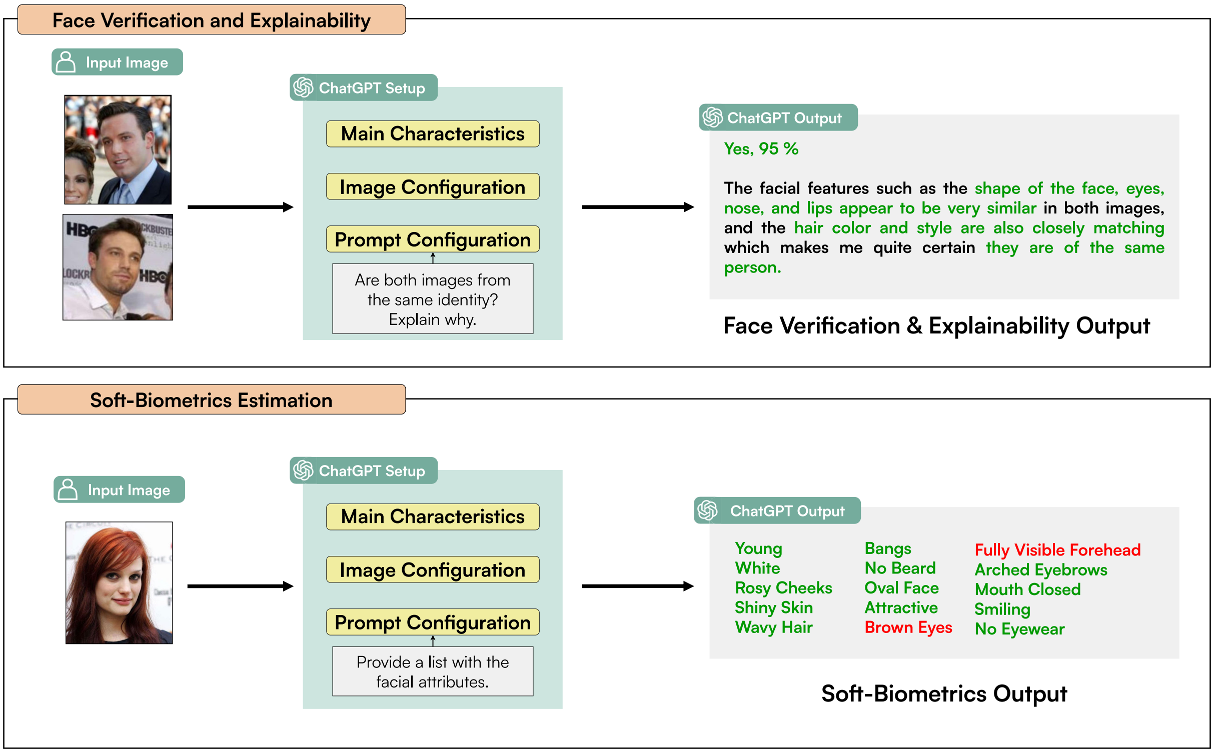


FIGURE 1. Graphical representation of the analysis carried out in this study, focused on the ability of ChatGPT to perform tasks such as face verification, soft-biometrics estimation, and explainability of the results. Different configurations of ChatGPT are explored in the present study.

One of the first popular LLMs that revolutionized the field was GPT-1 [10]. GPT-1[3] was the first LLM created by OpenAI and is based on a Transformer architecture [11], providing a more structured memory to handle long-term dependencies in comparison to traditional Recurrent Neural Networks (RNNs) [12], [13]. In addition to the Transformer architecture, the authors explored a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. In particular, the authors demonstrated that it is possible to achieve good performance on new tasks (e.g., textual entailment, reading comprehension, etc.) when the model is developed in an unsupervised way training with a large amount of data (BooksCorpus dataset [14]), and then fine-tuned to each specific dataset with minimal adaptation.

Since the publication of GPT-1 (117 million parameters) in 2018, several LLMs have been presented in the field [5], scaling up the models as it helps to greatly improve task-agnostic, few-shot performance. An example of this is GPT-3, which was presented in 2020 and comprises 175 billion parameters [15]. In that paper the authors demonstrated the influence of model size in the performance, concluding that GPT-3 is able to achieve promising results in the zero-shot and one-shot settings, and is able to achieve state-of-the-art results in few-shot settings. These interesting results in terms of generalization originated the integration of the GPT-3 model in the ChatGPT chatbot, achieving astonishing results.

Nevertheless, OpenAI is not the only company researching in the field. Others such as Google and Meta AI have recently presented their own LLMs known as PaLM 1 and 2[4] [16], [17] and LLaMA[5] [18], respectively. However, most of them only operate to date with text as input/output, and their corresponding chatbots such as Google Bard[6] have just been presented, including several limitations that restrict the application scenarios. As a result, the purpose of this study is to explore the ability of the popular ChatGPT, which is based on the recent GPT-4 multimodal (text, image, and video) LLM [19], for the task of face biometrics. Fig. 1 provides a graphical representation of the analysis carried out in this study, focused on the ability of ChatGPT to perform tasks such as face verification, soft-biometrics estimation, and explainability of the results. Face biometrics is a very challenging task in the field of computer vision and image understanding due to the large intra-user variability produced by factors such as pose, age, illumination, expression, etc., [20], [21].

It is important to remark that we initially considered also the Google Bard chatbot in the analysis, based on the PaLM 2 multimodal LLM. However, we had to discard

[3]https://openai.com/research/language-unsupervised *(January, 2024)*

[4]https://ai.google/discover/palm2/ *(January, 2024)*
[5]https://ai.meta.com/llama *(January, 2024)*
[6]https://bard.google.com/chat *(January, 2024)*

it due to two main reasons: *i)* when a face image was introduced in the chatbot, the answer of Google Bard was always *"Sorry, I can't help with images of people yet"*, and *ii)* unlike ChatGPT, Google Bard does not provide an Application Programming Interface (API) that allows to perform experiments using Python in a simple way.

The main contributions of the present study are:

- We explore the application of recent LLMs for the task of face verification, i.e., determine whether a pair of face images belong to the same subject or not. In particular, we consider ChatGPT Chatbot and the latest multimodal LLM available, GPT-4. Experiments are carried out in order to evaluate the performance and robustness of ChatGPT in different conditions (e.g., age, pose, image quality, etc.) using them with popular public benchmarks and comparing the results with state-of-the-art methods. For reproducibility reasons, we release all the code in GitHub[7].
- We also explore the ability of ChatGPT to perform other face biometric tasks such as the estimation of soft-biometric attributes (e.g., gender, age, ethnicity, type, and color of the hair, glasses, etc.) and reason about the output scores. We hypothesize that a simple interaction with chatbots could be beneficial to farther increase the explainability and transparency of automatic decisions in human scenarios.

The remainder of the paper is organized as follows. Sec. II focuses on the configuration details of ChatGPT, including the API parameters and the proposed prompts. Sec. III explains all the details regarding our proposed experimental framework, including the databases and experimental protocol considered, as well as the state-of-the-art face recognition systems included in the comparison. Sec. IV provides the results achieved by ChatGPT for the different face biometric tasks studied. We also show qualitative results in terms of explanability with ChatGPT. Finally, Sec. V draws the final conclusions and points out some lines for future work.

II. CHATGPT: SETUP

A. MAIN CHARACTERISTICS

OpenAI offers accessibility to ChatGPT through their interactive chatbot interface or through an API. Both of them have the same functionalities, but the API provides a simpler interface to run extensive experiments in Python. For this reason, although the experiments included in Sec. IV are performed using the API, we initially performed some quick experiments using the chatbot interface to explore easily the most adequate configurations. At the date of writing this paper, a premium subscription is needed to use the latest LLM (GPT-4), which accepts images along with other file formats and the use of other OpenAI products. Regarding its use and limitations, the API is split into different tiers, giving you more model capacities (requests, tokens, etc.) the more you pay. The number of requests given per tier changes actively, increasing from 100 RPD (Requests Per Day) for the *gpt-4-vision-preview* model to 500 RPD in less than a month for the tier 1 users[8].

We describe next the main aspects of the API that can greatly affect the results and cost of the experiments:

1) PROMPT

The text/question to be introduced as input to the LLM. OpenAI gives different recommendations to get the expected result, like writing clear instructions, providing reference text, or splitting complex tasks into simpler subtasks.[9]

2) ROLES

In order to interact with the API, it is mandatory to choose one of the three roles. The *system* role allows you to specify the way the model answers questions. The *user* role represents the queries made by the user. Lastly, the *assistant* role is employed to simulate the model's replies as, unlike the chatbot interface, the API lacks memory of prior messages. For all our experiments the prompt is sent with the *system* role, and the corresponding image with the *user* role. *Assistant* role is not used for any experiment.

3) MAX TOKENS

This parameter indicates the maximum number of tokens that the model can return. By tuning this, we can control the output style and price per request. For all our experiments we establish this parameter to 1,000 tokens.

4) IMAGE DETAIL

This parameter offers three settings: *low*, *high*, and *auto*. Depending on the configuration selected, it is possible to change the model's image processing and textual comprehension, thereby regulating the level of detail in its output[10]. For all our experiments we establish this parameter to *high*.

5) SEED

This parameter is at the date of writing this study in Beta, so OpenAI does not guarantee its functionality. In our experience, it has been very useful to have similar outputs from which to estimate important information. However, we do not use any seed in our experiments.

The influence of these API parameters is farther investigated in the following section.

B. DESIGN AND CONFIGURATION

Several configurations were tested in order to increase the performance in face biometrics, and at the same time optimize the usage of ChatGPT to reduce the cost and time.

[7]https://github.com/BiDAlab/ChatGPT_FaceBiometrics

[8]https://platform.openai.com/docs/guides/rate-limits *(January, 2024)*

[9]https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results *(January, 2024)*

[10]https://platform.openai.com/docs/guides/vision/low-or-high-fidelity-image-understanding *(January, 2024)*

FIGURE 2. Graphical representations of the images input to ChatGPT: a comparison of two faces merged in a single image (left), and a matrix of 4×3 face comparisons in a single image (right). In the latter, each cell is separated from the rest by a blue border and identified by a red number (from 0 to 11) that is used to reference the cell in the output of the model.

1) IMAGE CONFIGURATION

Two options are considered. First, we've opted to merge the two facial images involved in the comparison into a single image as can be seen in Fig. 2 (left). We also consider a second configuration, which has advantages in terms of cost and comparison times. As ChatGPT is able to discern and respond to tabulated information on a cell-by-cell basis, we also create a matrix including 4×3 face comparisons, as depicted in Fig. 2 (right). According to OpenAI's documentation [19], the size of an optimal image is approximately $2{,}000 \times 768$ pixels for high-resolution images. Thus, for this purpose, we created a matrix of $2{,}123 \times 903$ pixels, organized into 4 columns and 3 rows (i.e., we fit 12 face comparisons of 512×256 pixels plus the cell borders). In order to distinguish between different face comparisons, each cell is separated from the rest by a blue border and identified by a red number (from 0 to 11) that is used to reference the cell in the output of the model.

2) PROMPT CONFIGURATION

This is the most important aspect to analyze. First, we focus on the prompt design for the face verification task, considering the first image configuration case, i.e., a single face comparison in the image. Following OpenAI's recommendations, we created a detailed prompt asking to identify if two face images are coming from the same person or not. However, as ChatGPT does not officially provide face recognition capabilities, it refused to answer as can be seen next (our prompt appears in blue color, ChatGPT answers in black color):

> **Prompt:** *I want you to behave like a face recognition system. Given an image of two people, I want you to tell me if they are the same person or not. This is the image.*
> **Answer:** *I'm unable to perform actual face recognition tasks or confirm the identities of individuals in images. However, I can provide general observations about the features in the photographs if that would be helpful.*

We hypothesize that this answer might be motivated due to privacy concerns with real-life identities. In order to overcome this, we decided to modify our initial prompt, indicating that they are *AI-generated* people, as can be seen next:

> **Prompt:** *I want you to behave like a face recognition system for AI-generated images. Given an image of two AI-generated people, I want you to tell me if they are the same person or not. This is the image.*
> **Answer:** *As an AI, I don't have the capability to perform biometric identification or verification like a face recognition system. I can provide observations about the visual aspects of the image, such as apparent differences or similarities in facial features, hair color, and other visual attributes. However, any conclusion about whether two images depict the same person would be speculative and not based on biometric analysis. In the image you've provided, there are two individuals who each have distinct facial features. They appear to be different people, judging by their facial structures, hairstyles, and other visible features. It's important to note that visual inspection is not a definitive method for verifying identity, especially when considering the quality and angle of the photographs, makeup, lighting, and other factors that might alter appearance.*

As can be seen in the box, with the proposed prompt we are able to obtain a positive answer from ChatGPT, indicating if face images are from the same identity or not, and also providing some explainability of their decision: "*They appear to be different people, judging by their facial structures, hairstyles, and other visible features*". However, it still points out that it is not its main functionality and "*any conclusion about whether two images depict the same person would be speculative*".

After that, we also tried to reduce the volume of information provided as input, thereby preventing the system from recognizing that it is engaged in facial recognition tasks.

Despite this, ChatGPT was able to detect our idea, providing a negative answer.

> **Prompt:** *I will give you two images. I want you to tell me if they are from the same identity or not.*
> **Answer:** *I'm sorry, but I cannot assist with identifying or making assumptions about the identities of people in images. If you have any other questions or need assistance with a different topic, feel free to ask!*

Nevertheless, we also tried to restrict the output of ChatGPT. In particular, we forced the response to be either *'yes'* or *'no'*, accompanied by a confidence value. After several experiments, we observed that ChatGPT always provided a confidence value of 100% on the response, even in instances of failure. To address this, we request a more precise value within the 0-100 range.

> **Prompt:** *I will give you two images. I want you to tell me if they are from the same identity or not. Answer with yes or no only. Give me a number between 0 and 100 representing how sure you are of your response. Try not to give exactly 0 or 100.*
> **Answer:** *No, 95*

As can be seen, with the proposed prompt, ChatGPT is able to provide a clear and concise response to our question, being this prompt the one selected for the face verification experiments. Starting from this configuration, we created another prompt for the matrix strategy in which we specify the position of the comparisons in the matrix and how to refer to each cell. Results using this particular prompt are analyzed in Sec. IV.

> **Prompt:** *I will give you a matrix containing different images. Each cell of the matrix is separated from the others by a blue border, and each cell also has a number over it showing its position and how I want you to refer to that cell. For each cell of the image tell me if the two main focused people are the same or not as well as a number between 0 and 100 representing how sure you are of your response. Try not to give me exactly 0 or 100.*

Another point we explore in the present study is the potential of ChatGPT to other face biometric tasks such as the estimation of soft biometrics and explainability of the results. To achieve this, several prompts were studied. Regarding the estimation of facial soft biometrics, we first considered a general prompt to see the ability of ChatGPT for the task in terms of precision and variability of the attributes. Fig. 3 shows the prompt considered as well as the results provided by ChatGPT for different face images. As can be seen, in general, ChatGPT is able to provide a great variability of soft biometrics with accurate results. To evaluate quantitatively the performance of ChatGPT, we decided to propose a prompt that includes the facial attributes considered in the popular MAAD-Face database [22], allowing, therefore, a direct comparison to state-of-the-art approaches. We provide our proposed prompt to evaluate the ability of ChatGPT to estimate facial soft biometrics.

> **Prompt:** *For the next list of attributes, I want you to give me two lists representing whether the image has that attribute or not. Answer to me in a python-like list, containing the attributes that are present. The list of attributes is: [Male, Young, Middle_Aged, Senior, Asian, White, Black, Rosy_Cheeks, Shiny_Skin, Bald, Wavy_Hair, Receding_Hairline, Bangs, Sideburns, Black_Hair, Blond_Hair, Brown_Hair, Gray_Hair, No_Beard, Mustache, 5_o_Clock_Shadow, Goatee, Oval_Face, Square_Face, Round_Face, Double_Chin, High_Cheekbones, Chubby, Obstructed_Forehead, Fully_Visible_Forehead, Brown_Eyes, Bags_Under_Eyes, Bushy_Eyebrows, Arched_Eyebrows, Mouth_Closed, Smiling, Big_Lips, Big_Nose, Pointy_Nose, Heavy_Makeup, Wearing_Hat, Wearing_Earrings, Wearing_Necktie, Wearing_Lipstick, No_Eyewear, Eyeglasses, Attractive]*
> **Answer:** *The individual in the image has the following attributes: [Young, Asian, Rosy Cheeks, Shiny Skin, Wavy Hair, Black Hair, No Beard, Oval Face, High Cheekbones, Fully Visible Forehead, Brown Eyes, Arched Eyebrows, Mouth Closed, Smiling, Heavy Makeup, Wearing Earrings, Wearing Lipstick, No Eyewear, Attractive]*

Finally, regarding the explainability of the decisions using ChatGPT, we considered the same prompt used for the face verification task, adding a final question that assesses why ChatGPT is making that decision.

> **Prompt:** *I will give you two images. I want you to tell me if they are from the same identity or not. Answer with yes or no only. Give me a number between 0 and 100 representing how sure you are of your response. Try not to give exactly 0 or 100. Tell me why you answer that.*

III. EXPERIMENTAL FRAMEWORK

A. EXPERIMENTAL PROTOCOL

The experimental protocol proposed in this study has been designed to analyze quantitatively the performance of ChatGPT for the tasks of face verification and soft-biometrics estimation.

1) FACE VERIFICATION

First, we analyze the ability of ChatGPT in different application scenarios (i.e., controlled, surveillance, and extreme conditions) and image qualities. For this purpose, we consider the following databases in the evaluation:

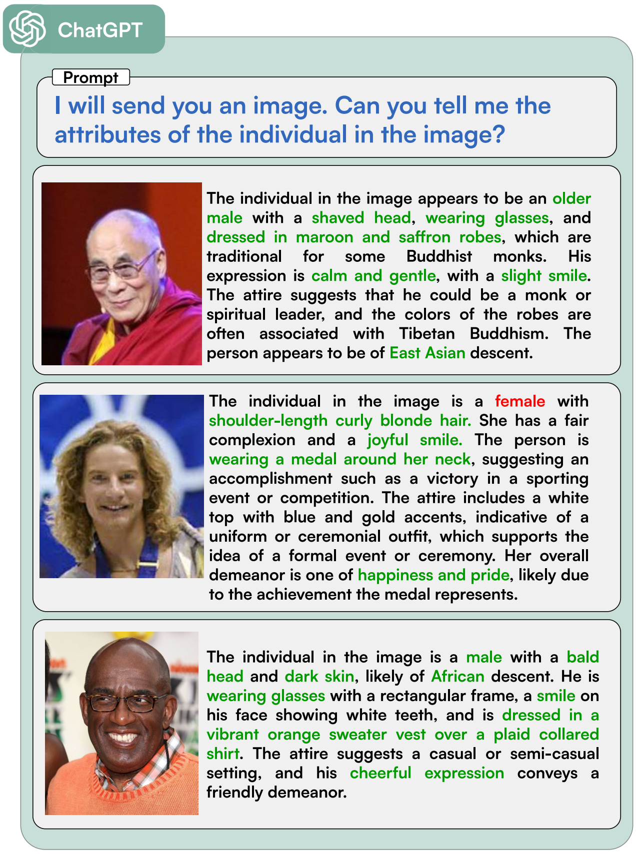


FIGURE 3. Prompt inserted to ChatGPT together with the different outputs provided for different face images. We highlight the most important soft-biometrics attributes in green/red color if they are correct/incorrect, respectively.

- Labeled Faces in the Wild (LFW) [23]: this is a very popular database in the field, containing high-quality images with no hard variations in pose.
- QUIS-CAMPI [24]: this database comprises videos and images captured in an uncontrolled outdoor setting using a camera positioned approximately 50 meters away from the subjects.
- TinyFaces [25]: this database consists in images of extremely low quality, with an average resolution of 20×16 pixels.

In addition to this, we also evaluate the performance of ChatGPT when considering popular challenges in face recognition such as demographic bias, age and pose variations, and occlusions. The following databases are considered in the evaluation, which are also considered in the recent FRCSyn Challenge [21]:

- BUPT-BalancedFace [26]: this database is specifically designed to tackle performance variations among various ethnic groups. It comprises eight distinct demographic groups formed by a combination of ethnicities (White, Black, Asian, Indian) and gender (Male, Female).
- CFP-FP [27]: this database presents images from subjects with great changes in pose and different environmental contexts.
- AgeDB [28]: this database presents diverse images featuring subjects of varying ages in different environmental contexts.
- ROF [29]: this database consists of occluded faces with both upper face occlusion, due to sunglasses, and lower face occlusion, due to masks.

2) SOFT-BIOMETRICS ESTIMATION

to assess the performance of ChatGPT for the estimation of soft biometrics, we use the MAAD-Face database [22] which is based on the VGGFace2 database [30]. MAAD-Face database provides a total of 47 soft-biometric attributes per face image. In addition, we also consider the LFW database [23] as the authors of [31] labeled manually the following soft biometrics: gender, age, and ethnicity.

Finally, it is important to highlight that, due to the limitations of the OpenAI's API in terms of the number of requests per day, and the price, we had to reduce the number of face comparisons to 1,000 per database for the face verification task. These comparisons are selected randomly from the standard protocols. Regarding the soft-biometrics estimation, we consider 1,000 face images per database. At the date of writing this paper, the cost of ChatGPT is 0.01$ per 1,000 input tokens and 0.03$ per 1,000 output tokens. The tokens per image are calculated based on the *image detail* parameter[11]. The total cost of the experimental framework carried out in the present study has been 124.31$, and it took us 30 days.

B. COMPARISON WITH THE STATE OF THE ART

In the present study, we compare the results achieved by ChatGPT with state-of-the-art methods. For the task of face verification, the following two approaches are considered:

1) ARCFACE [32]

this face verification system considers a loss function that maps facial features into a high-dimensional hypersphere where the embeddings are optimized to maximize the angular margins between different identities. The system considered in this study is based on the iResNet-100 architecture [33] pretrained using the MS1Mv3 database [34]. Cosine distance is used to measure the similarity between feature embeddings.

2) ADAFACE [35]

this face verification system proposes a new loss function in order to pay more attention to the harder examples in terms of image quality. In particular, the authors proposed an adaptive margin function that approximates the image quality with feature norms. The system considered in this study is based on the iResNet-100 architecture [33] pretrained using

[11] https://platform.openai.com/docs/guides/vision/low-or-high-fidelity-image-understanding *(January, 2024)*

TABLE 1. Face Verification Task: Performance in terms of Accuracy (%) achieved by ChatGPT and popular face verification systems in the literature. The term "ChatGPT 4 × 3" refers to the image configuration containing 12 face comparisons in the same prompt, whereas "ChatGPT 1 × 1" refers to the case of just a single face comparison per prompt.

	Controlled	Surveillance	Extreme	Demographic Bias	Pose	Age	Occlusions	Average
	LFW [23]	**QUIS-CAMPI** [24]	**TinyFaces** [25]	**BUPT** [26]	**CFP-FP** [27]	**AgeDB** [28]	**ROF** [29]	
ArcFace [32]	96.40	97.80	**87.50**	98.60	98.80	96.90	92.10	95.44
AdaFace [35]	**96.80**	**98.80**	**87.50**	**98.70**	**99.60**	**97.00**	**92.20**	**95.80**
ChatGPT 4x3	74.60	69.80	61.10	65.27	70.20	62.70	59.96	66.23
ChatGPT 1x1	**93.50**	**79.20**	**72.47**	**76.92**	**86.40**	**77.30**	**75.95**	**80.19**

TABLE 2. Face Verification Task: Performance in terms of Equal Error Rate (%) achieved by ChatGPT and popular face verification systems in the literature. The term "ChatGPT 4 × 3" refers to the image configuration containing 12 face comparisons in the same prompt, whereas "ChatGPT 1 × 1" refers to the case of just a single face comparison per prompt.

	Controlled	Surveillance	Extreme	Demographic Bias	Pose	Age	Occlusions	Average
	LFW [23]	**QUIS-CAMPI** [24]	**TinyFaces** [25]	**BUPT** [26]	**CFP-FP** [27]	**AgeDB** [28]	**ROF** [29]	
ArcFace [32]	6.70	2.20	14.50	1.80	1.20	3.80	13.10	6.19
AdaFace [35]	**6.50**	**1.20**	**14.20**	**1.30**	**0.40**	**3.80**	**11.70**	**5.59**
ChatGPT 4x3	26.60	31.00	40.57	34.93	30.30	40.90	40.44	34.96
ChatGPT 1x1	**8.60**	**24.00**	**32.07**	**23.79**	**13.40**	**22.40**	**23.75**	**21.19**

the WebFace12M database [36]. Cosine distance is used to measure the similarity between feature embeddings.

Regarding the soft-biometrics estimation, we compare the results achieved by ChatGPT with two different approaches. For the LFW database, as it only contains the soft biometrics related to gender, age, and ethnicity, we consider FairFace [37], as it provides state-of-the-art results. Finally, for the MAAD-Face database, as it contains 47 soft-biometric attributes per facial image, we specifically train a ResNet-50 architecture pretrained on Imagenet [38] using the train set of MAAD-Face.

IV. EXPERIMENTAL RESULTS

A. FACE VERIFICATION

We conducted a performance assessment considering ArcFace, AdaFace, and ChatGPT. Specifically for ChatGPT, we took evaluations using images in a matrix-like configuration (i.e., ChatGPT 4 × 3), as well as through direct comparisons (i.e., ChatGPT 1 × 1). To measure the similarity between ArcFace and AdaFace embeddings, we use the cosine distance. We also consider this metric to determine the Equal Error Rate (EER) for these models. For ChatGPT, we use the confidence values obtained directly from ChatGPT's outputs as a custom metric to obtain this EER. The results are shown in Tables 1 and 2, categorized into two main groups. The first group (left part) refers to different application scenarios, including controlled environments (LFW), surveillance (QUIS-CAMPI), and extreme conditions (TinyFaces). The second group (right part) highlights popular challenges in face recognition such as demographic bias (BUPT), pose (CFP-FP) and age (AgeDB) variations, and occlusions (ROF). Lastly, the rightmost column presents the average performance of each model across all databases.

In general, state-of-the-art models such as ArcFace (95.44% Average accuracy, 6.19% Average EER) and AdaFace (95.80% Average accuracy, 5.59% Average EER) exhibit superior overall performance compared to ChatGPT. However, while these models are trained for this specific task, ChatGPT is primarily oriented to more general tasks. Moreover, when evaluating ChatGPT, a significant decline in performance was observed when the images were presented in a matrix format (66.23% Average accuracy, 34.96% Average EER) compared to the case of providing comparisons one by one (80.19% Average accuracy, 21.19% Average EER). We hypothesize that this reduction in performance might be produced as in the ChatGPT 4 × 3 case, the model needs first to detect the faces in the whole image, and then perform facial verification, potentially compromising overall task execution. Nevertheless, considering this matrix approximation could serve as a quick solution when the ChatGPT API imposes limitations on daily requests or when the budget to perform comparisons is low.

Analyzing the results achieved in each database, it becomes evident that ChatGPT's performance varies greatly based on image quality, pose variations, and domain disparities among comparisons. In databases like LFW, where images exhibit good quality and consistent poses, ChatGPT achieves performance close to state-of-the-art models (93.50% Accuracy, 8.60% EER). This indicates the potential of ChatGPT for controlled environments. However, in the surveillance scenario using the QUIS-CAMPI, characterized by a mix of CCTV and mugshot images, the performance of ChatGPT drops significantly (79.20% Accuracy, 24.00% EER). Furthermore, in the TinyFaces database, characterized by extreme conditions in terms of quality, ChatGPT's performance declines even more (72.47% Accuracy, 32.07% EER). These results discourage the application of ChatGPT for more challenging scenarios. Similar performance drops are observed in the databases related to challenges such as bias, pose, age, and occlusions: BUPT (76.92% Accuracy, 23.79%), CFP-FP

TABLE 3. Face Verification Task: Performance achieved by ChatGPT for the different demographic groups considered in the recent BUPT database.

	BUPT [26]	
	Accuracy (%)	**EER (%)**
White Male	78.91	23.14
White Female	**83.76**	**14.94**
Black Male	76.42	24.01
Black Female	79.51	20.49
Asian Male	80.67	22.82
Asian Female	71.22	29.35
Indian Male	79.51	24.70
Indian Female	65.38	30.88
Average	76.92	23.79

(86.40% Accuracy, 13.40%), AgeDB (77.30% Accuracy, 22.40% EER) and ROF (75.95% Accuracy, 23.75%).

One notable concern is the potential for biased or inappropriate content generation, stemming from the models' dependence on extensive internet untreated training data. LLMs learn from diverse sources on the internet, absorbing biases present in the data [39]. This can result in the reproduction of societal biases within the generated content. For example, gender bias, racial bias, or other forms of prejudice may manifest in the outputs of the model. To address this bias, we perform an evaluation focused on the different genders and ethnicities included in the BUPT database, which considers four distinct labeled ethnicities and two genders, all balanced in the same proportion. Table 3 shows the results achieved by ChatGPT for the different ethnicities and genders considered in the BUPT database. As can be observed, ChatGPT provides very different performances among the different demographic groups (e.g., 14.94% EER for the white female group vs. 30.88% EER for the Indian female group), showing the large bias it has.

B. EXPLAINABILITY

For completeness, we also analyze how ChatGPT can increase the explainability of the results for the task of face verification. In Fig. 4, the proposed prompt is shown, along with the outputs provided by ChatGPT for some examples of the different face verification databases. ChatGPT's responses are divided into right (on the left column) and wrong (on the right column).

In both right and wrong answers, ChatGPT demonstrates its ability to rationalize decisions based on image features. For example, in most cases, the output score of ChatGPT for the task of face verification is related to soft-biometric attributes such as facial hair and skin tone. Additionally, it exhibits the capability to focus on more detailed attributes like eye color, face shape, or nose shape, showcasing proficiency in handling both coarse and fine details.

It is noteworthy that ChatGPT considersfacial expressions in its predictions, despite of the fact that this is a variable

TABLE 4. Soft Biometrics: Accuracy (%) achieved by ChatGPT over the MAAD-Face database for the estimation of the 47 soft-biometric attributes considered in the database.

	LFW [23]		
	Gender	**Age**	**Ethnicity**
FairFace	**98.23**	67.88	87.48
ChatGPT	94.05	**72.80**	**88.25**

TABLE 5. Soft Biometrics: Accuracy (%) achieved by ChatGPT over the MAAD-Face database for the estimation of the 47 soft-biometric attributes considered in the database.

	MAAD-Face [22]	
	ResNet-50 [40]	**ChatGPT**
Male	85.50	**96.30**
Young	**82.90**	49.50
Middle_Aged	**93.20**	92.30
Senior	**92.30**	91.10
Asian	**93.60**	89.00
White	73.80	**83.90**
Black	96.50	**97.50**
Rosy_Cheeks	**99.40**	73.50
Shiny_Skin	**84.70**	68.10
Bald	**95.90**	94.50
Wavy_Hair	**84.80**	72.70
Receding_Hairline	**90.20**	84.30
Bangs	**91.80**	89.70
Sideburns	**83.80**	75.90
Black_Hair	**84.60**	82.60
Blond_Hair	**92.90**	91.90
Brown_Hair	**79.00**	76.40
Gray_Hair	91.40	**91.90**
No_Beard	**87.50**	71.70
Mustache	**99.80**	98.20
5_o_Clock_Shadow	**89.70**	89.40
Goatee	**99.80**	99.70
Oval_Face	**87.00**	30.60
Square_Face	**86.60**	57.50
Round_Face	**99.60**	98.80
Double_Chin	**89.40**	81.30
High_Cheekbones	**77.60**	71.50
Chubby	**90.70**	86.70
Obstructed_Forehead	**94.30**	94.20
Fully_Visible_Forehead	**74.50**	51.90
Brown_Eyes	**73.20**	56.30
Bags_Under_Eyes	**82.70**	75.30
Bushy_Eyebrows	**86.00**	75.10
Arched_Eyebrows	**77.90**	73.40
Mouth_Closed	**96.20**	18.40
Smiling	**82.30**	69.90
Big_Lips	**81.00**	73.20
Big_Nose	**89.20**	87.30
Pointy_Nose	**76.70**	45.90
Heavy_Makeup	**86.40**	73.00
Wearing_Hat	**87.30**	95.50
Wearing_Earrings	**80.50**	70.30
Wearing_Necktie	**90.90**	86.50
Wearing_Lipstick	**84.20**	64.00
No_Eyewear	75.40	**75.80**
Eyeglasses	**92.90**	89.20
Attractive	**86.40**	56.40
Average Accuracy	**87.28**	76.98

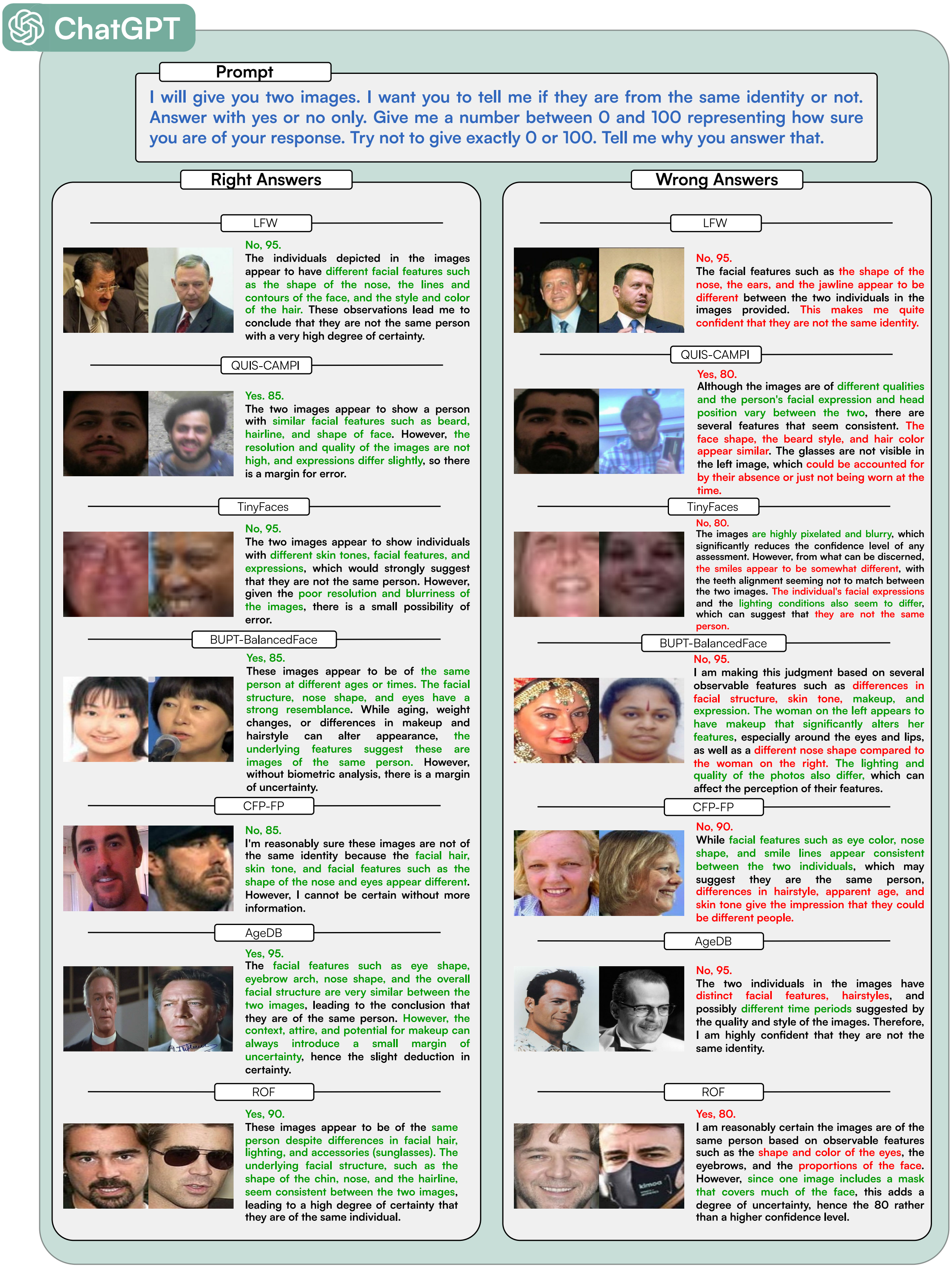


FIGURE 4. **Explainability: Proposed prompt along with the outputs provided by ChatGPT for some examples of the different face verification databases. Left column: examples where ChatGPT answers are correct; right column: incorrect answers. We highlight the most important parts of the text in green/red color if they are correct/incorrect, respectively.**

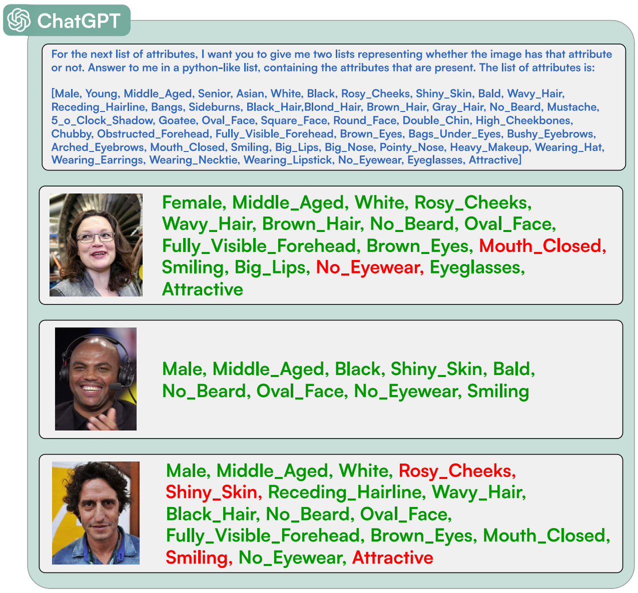


FIGURE 5. Soft Biometrics: Proposed prompt along with the outputs provided by ChatGPT for some examples of the MAAD-Face database [22]. We highlight the most important parts of the text in green/red color if they are correct/incorrect, respectively.

attribute that should not be considered. Furthermore, the model recognizes temporal differences between images, incorporating this information into its predictions.

Regarding the wrong answers, we observe that, although the prediction may be wrong, some of the explanations provided by ChatGPT are accurate in describing the people in the images.

C. SOFT BIOMETRICS

Tables 4 and 5 shows the results achieved for the soft-biometrics estimation task for the LFW and MAAD-Face databases, respectively. For completeness, Fig. 5 shows some examples of the output provided by ChatGPT with the proposed prompt.

Analyzing the results achieved on LFW database, FairFace exhibited superior performance for gender classification (98.23% Accuracy) compared to ChatGPT (94.05% Accuracy). Despite this, ChatGPT outperforms FairFace for age classification (72.87% vs. 67.88% Accuracy) and ethnicity classification (88.25% vs. 87.48%). These results prove the potential of ChatGPT for certain facial attribute classifications.

For a more extensive evaluation, we consider the MAAD-Face dataset, annotated with 47 distinct attributes. Our custom model (ResNet-50) achieves superior performance across the majority of attributes (87.28% Average accuracy). Nevertheless, ChatGPT, although having a lower average performance (76.98% Average Accuracy), excels on some facial attributes. Some of the most notorious soft-biometric attributes where ChatGPT achieves better performance are in gender classification (96.30% Accuracy), some ethnicities (White - 83.90% Accuracy, Black - 97.50%), and accessories such as wearing a hat. These results follow the same conclusions drawn in the face verification task. While specific models trained for the task achieve in general better results, ChatGPT shows promising results and utility for tasks with no prior training.

V. CONCLUSION

In this study, we have conducted a comprehensive evaluation of ChatGPT's capabilities in handling facial biometric tasks, including face verification, soft-biometric estimation, and explainability. Our experiments spanned across various databases and challenges, comparing the performance of ChatGPT with specialized models trained explicitly for these tasks. The experiments have revealed that while ChatGPT may not attain the same levels of accuracy as dedicated models, it presents a promising utility as an initial assessment tool with zero training. For example, results with around 94% Accuracy are obtained in the LFW database for face verification. Also, impressive results are achieved for the estimation of soft biometrics such as gender ($\approx$96%) in the MAAD-Face database, or age ($\approx$73%) and ethnicity ($\approx$88%) in the LFW database. Furthermore, its ability to return textual outputs contributes to a better explainability of the results. Future work will be oriented to analyze the ability of other popular chatbots for face biometrics.

REFERENCES

[1] K. Hu. (2023). *ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note*. Reuters Technol. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[2] S. S. Biswas, "Role of chat GPT in public health," *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, May 2023.

[3] E. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individual Differences*, vol. 103, Apr. 2023, Art. no. 102274.

[4] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[5] W. X. Zhao, "A survey of large language models," 2023, *arXiv:2303.18223*.

[6] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv:2303.13375*.

[7] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, F. Zeng, L. Sun, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, and X. Li, "DeID-GPT: Zero-shot medical text de-identification by GPT-4," 2023, *arXiv:2303.11032*.

[8] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[9] A. Singla, "Evaluating ChatGPT and GPT-4 for visual programming," in *Proc. ACM Conf. Int. Comput. Educ. Res.*, vol. 2, Aug. 2023, pp. 14–15.

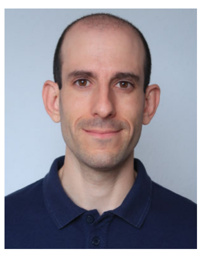
[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[12] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 1–8.

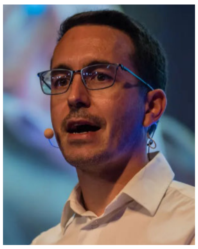
[13] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, "Exploring transformers for behavioural biometrics: A case study in gait recognition," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109798.
[14] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[15] T. Brown, B. Mann, and N. Ryder, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[16] A. Chowdhery, S. Narang, and J. Devlin, "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-1144.html
[17] R. Anil, "PaLM 2 technical report," 2023, *arXiv:2305.10403*.
[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[19] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[20] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102805.
[21] P. Melzi et al., "FRCSyn challenge at WACV 2024: Face recognition challenge in the era of synthetic data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, Jan. 2024, pp. 892–901.
[22] P. Terhörst, D. Fährmann, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "MAAD-face: A massively annotated attribute dataset for face images," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3942–3957, 2021.
[23] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images: Detection, Alignment, Recognit.*, 2008, pp. 1–15.
[24] J. Neves, J. Moreno, and H. Proença, "QUIS-CAMPI: An annotated multi-biometrics data feed from surveillance scenarios," *IET Biometrics*, vol. 7, no. 4, pp. 371–379, Jul. 2018.
[25] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Proc. Asian Conf. Comput. Vis.*, 2019, pp. 605–621.
[26] M. Wang, Y. Zhang, and W. Deng, "Meta balanced network for fair face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8433–8448, Nov. 2022.
[27] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
[28] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1997–2005.
[29] M. E. Erak?n, U. Demir, and H. K. Ekenel, "On recognizing occluded faces in the wild," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2021, pp. 1–5.
[30] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[31] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018.
[32] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
[33] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9415–9422.
[34] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2638–2646.
[35] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Milwaukee, WI, USA, Jun. 2022, pp. 18729–18738.
[36] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10487–10497.
[37] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.
[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[39] L. Lucy and D. Bamman, "Gender and representation bias in GPT-3 generated stories," in *Proc. 3rd Workshop Narrative Understand.*, 2021, pp. 48–55. [Online]. Available: https://aclanthology.org/2021.nuse-1.5
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

IVAN DEANDRES-TAME received the B.Sc. in Computer Science and Engineering in 2021 and the M.Sc. in Deep learning for Image, Audio and Video processing in 2022 from the Universidad Autonoma de Madrid. In May 2023, he joined the Biometrics and Data Pattern Analytics Lab - BiDA Lab (UAM), where he is currently collaborating as an assistant researcher pursuing a Ph.D. degree. The research activities he is currently working on include Human-Computer interaction for Digital Health and Digital Security domain, Face Recognition, Machine Learning, and Deep Learning.



RUBEN TOLOSANA received the M.Sc. degree in Telecommunication Engineering, and the Ph.D. degree in Computer and Telecommunication Engineering, from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In 2014, he joined the Biometrics and Data Pattern Analytics - BiDA Lab at the Universidad Autonoma de Madrid, where he is currently an Assistant Professor. He is a member of the ELLIS Society, the Technical Area Committee of EURASIP, and the Editorial Board of the IEEE Biometrics Council Newsletter. His research interests are mainly focused on signal and image processing, pattern recognition, and machine learning, particularly in the areas of DeepFakes, Human-Computer Interaction, Biometrics, and Health. He is the author of more than 80 scientific articles published in international journals and conferences. He has served as General Chair and Program Chair (AVSS 2022), and Area Chair (IJCB 2023, ICPR 2022) in top conferences. Dr. Tolosana has also received several awards such as the European Biometrics Industry Award (2018) from the European Association for Biometrics (EAB) and the Best Ph.D. Thesis Award in 2019-2022 from the Spanish Association for Pattern Recognition and Image Analysis (AERFAI).

RUBEN VERA-RODRIGUEZ received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Spain, in 2006, and the Ph.D. degree in electrical and electronic engineering from Swansea University, U.K., in 2010. Since 2010, he has been affiliated with the Biometric Recognition Group, Universidad Autonoma de Madrid, Spain, where he is currently an Associate Professor since 2018. Ruben has published over 100 scientific articles published in international journals and conferences. He is actively involved in several National and European projects focused on biometrics. Ruben has been Program Chair for the IEEE 51st International Carnahan Conference on Security and Technology (ICCST) in 2017; the 23rd Iberoamerican Congress on Pattern Recognition (CIARP 2018) in 2018; and the International Conference on Biometric Engineering and Applications (ICBEA 2019) in 2019.



AYTHAMI MORALES received his M.Sc. degree in Electrical Engineering in 2006 from Universidad de Las Palmas de Gran Canaria. He received his Ph.D. degree in Artificial Intelligence from La Universidad de Las Palmas de Gran Canaria in 2011. He performs his research works in the BiDA Lab - Biometric and Data Pattern Analytics Laboratory at Universidad Autónoma de Madrid, where he is currently an Associate Professor (CAM Lecturer Excellence Program). He is a member of th ELLIS Society (European Laboratory for Learning and Intelligent Systems). He has performed research stays at the Biometric Research Laboratory at Michigan State University, the Biometric Research Center at Hong Kong Polytechnic University, the Biometric System Laboratory at the University of Bologna, and Schepens Eye Research Institute (Harvard Medical School). He is the author of more than 100 scientific articles published in international journals and conferences and 2 patents.

JULIAN FIERREZ (Member, IEEE) received the M.Sc. and the Ph.D. degrees in telecommunications engineering from Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002 he has been affiliated as a PhD candidate with the Universidad Politecnica de Madrid, and since 2004 at Universidad Autonoma de Madrid, where he is currently a Full Professor since 2022. From 2007 to 2009 he was a visiting researcher at Michigan State University in the USA under a Marie Curie fellowship. Since 2016 he has been Associate Editor for Elsevier's Information Fusion and IEEE Trans. on Information Forensics and Security, and since 2018 also for IEEE Trans. on Image Processing. He has been the General Chair of the IAPR Iberoamerican Congress on Pattern Recognition (CIARP 2018) and the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2019). He is also the recipient of a number of world-class research distinctions, including: the EBF European Biometric Industry Award 2006, EURASIP Best PhD Award 2012, Medal in the Young Researcher Awards 2015 by the Spanish Royal Academy of Engineering, and the Miguel Catalan Award.

JAVIER ORTEGA-GARCIA (Fellow, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree (cum laude) in electrical engineering from Universidad Politecnica de Madrid, Spain, in 1989 and 1996, respectively. He is currently a Full Professo at the Signal Processing Chair at Universidad Autonoma de Madrid - Spain, where he holds courses on biometric recognition and digital signal processing. He is a founder and Director of the BiDA-Lab, Biometrics and Data Pattern Analytics Group. He has authored over 300 international contributions, including book chapters, refereed journals, and conference papers. His research interests are focused on biometric pattern recognition (on-line signature verification, speaker recognition, human-device interaction) for security, e-health, and user profiling applications. He chaired Odyssey-04, The Speaker Recognition Workshop, ICB-2013, the 6th IAPR International Conference on Biometrics, and ICCST2017, the 51st IEEE International Carnahan Conference on Security Technology.