IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 May 2024, accepted 25 May 2024, date of publication 28 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406478

RESEARCH ARTICLE

Uniss-FGD: A Novel Dataset of Human Gazes Over Images of Faces

PIETRO RUIU[1], MAURO FADDA[1], (Senior Member, IEEE), ANDREA LAGORIO[1], SETH NIXON[1], MATTEO ANEDDA[2], (Senior Member, IEEE), ENRICO GROSSO[1], AND MARINELLA IOLE CADONI[1]

[1]Department of Biomedical Science, University of Sassari, 07100 Sassari, Italy
[2]Department of Electrical and Electronic Engineering (UdR CNIT of Cagliari), University of Cagliari, 09123 Cagliari, Italy

Corresponding author: Matteo Anedda (matteo.anedda@unica.it)

This work was supported by Italian Ministry for Research and Education (MUR) through the Project e.INS-Ecosystem of Innovation for Next Generation Sardinia within the National Recovery and Resilience Plan (NRRP)-MISSION 4 COMPONENT 2, "From Research to Business" INVESTMENT 1.5, "Creation and Strengthening of Ecosystems of Innovation," and Construction of "Territorial Research and Development Leaders" (Spoke 06—Digital Transformation) under Grant cod. ECS 00000038.

ABSTRACT Face detection and recognition play pivotal roles across various domains, spanning from personal authentication to forensic investigations, surveillance, entertainment, and social media. In our interconnected world, pinpointing an individual's identity amidst millions remains a formidable challenge. While contemporary face recognition techniques now rival or even surpass human accuracy in critical scenarios like border identity control, they do so at the expense of poor explainability, leaving the underlying causes of errors largely unresolved. Moreover, they demand substantial computational resources and a plethora of labeled samples for training. Drawing inspiration from the remarkably efficient human visual system, particularly in localizing and recognizing faces, holds promise for developing more efficient and interpretable systems, with high gains in scenarios where misidentification can yield grave consequences. In this context, we introduce the Uniss-FGD dataset, which captures gaze data from observers presented with facial images depicting diverse expressions. In view of the potential uses of Uniss-FGD, we propose two baseline experiments on a subset of the dataset in which we perform a comparative analysis juxtaposing the attention mechanisms of ViTs, multi-scale handcrafted features, and human observers when viewing facial images. These preliminary comparisons pave the way to future investigation into the integration of human attention dynamics into advanced and diverse image analysis frameworks. Beyond the realms of Computer Science, numerous research disciplines stand to benefit from the rich gaze data encapsulated in this dataset.

INDEX TERMS Human gazes, vision transformers, handcrafted features, human faces, visual attention.

I. INTRODUCTION

The human gaze efficiently captures the salient aspects of any scene [1]. Consequently, the field of computer vision has long been dedicated to studying human gaze behavior to understand attentive mechanisms and apply them in various applications [2], [3], [4]. Human fixations data play a crucial role in understanding the mechanisms of the efficient visual system [5]. Various datasets have been curated to collect data on human gaze behavior across a wide range of stimuli, encompassing landscapes, objects, animals, human faces, outdoor scenes, social scenes and more [6].

These datasets are not only valuable within the field of computer vision but also hold significant interest for other research domains. In the medical field, human gaze data can be harnessed for a variety of applications, including the assessment of the onset or progression of degenerative conditions [7], [8]. In experimental psychology, eye

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

TABLE 1. Recent human gaze datasets based on facial stimuli.

Dataset	Year	Stimuli	Observers	Tasks	Durations
FiFA data base [12]	2007	200 indoor and outdoor images	7	free viewing	2 sec
EyeCrowd data set [13]	2014	500 indoor and outdoor images with varying crowd densities	16	free viewing	5 sec
Coutrot Database [14]	2016	8 videos	405	free viewing	17 sec (avg)
Extensive dataset [15]					
Exp. Face Discrimination	2017	32 face images	29	discrimination	1.5 sec
Exp. Face Learning		8 face images	104	discrimination	1.5 sec
Uniss-FGD dataset	2023	120 face images (3 expressions)	20	free viewing	3 sec

movements serve as a powerful tool for investigating various psychological processes, including language processing, image processing, auditory processing, memory, social cognition, and decision-making, in an unobtrusive and accurate manner [9]. In the field of human-machine interaction, gaze tracking technology proves highly useful for predicting people's intentions [10], or for remotely controlling pointers or vehicles [11], thereby facilitating seamless collaboration between machines and humans.

Face detection and face recognition are involved in countless consumer applications and devices characterized by intelligent, vision-based, human-computer interaction. Deep learning-based models are the state of the art in face recognition and they now reach impressive performances even in the ''wild'', where faces are not captured in a controlled way and significant variations in pose, illumination, resolution might occur [16]. However, their performance comes at the price of massive amount of face data required to train them and a lack of explainability, so errors are hard to predict and prevent, which hampers their use in contexts where the consequences of misclassification are not acceptable, such as in forensics scenarios or the authentication to access sensitive data. They also lack built-in scale-invariance, an ability that humans are able to achieve after a single exposure to a novel object [17]. Deep Learning-based models (DLM) and the latest Vision Transformer (ViT) models seem to moderately correlate with human visual attention and, at least for DLM models, it seems that the higher the correlation, the better their performance at classification tasks [18], [19]. These results, combined with the fact that the human visual system is very efficient in detecting salient points and in driving its attention to them, entice the design of machine models that exploit human visual attention. Despite the possible benefits of embedding human gaze information into machine models for face detection and recognition, publicly available datasets of human gaze data on face images are scarce.

With the dataset presented in this paper, named Uniss-FGD (Facial Gaze Dataset), the authors intend to address this shortcoming by providing gaze data on good quality images of faces with three different expressions. The observers freely viewed each face image for 3 seconds, a time span that is long enough to study how human attention varies with expressions, sex of the image subject and sex of the observer. We use a subset of the presented dataset for a comparative study between ViT's attention, a type of handcrafted features and human attention on face images. Research conducted to compare machine visual attention with that of humans is currently limited to Convolutional Neural Networks (CNN) or general images [18], [20], [21], [22], we contribute to extend this investigation by analysing ViTs, handcrafted features and human fixations on face images.

The contributions of this paper are twofold:

1) we introduce Uniss-FGD, a novel and innovative dataset consisting of human gaze data collected from 20 observers who viewed images of human faces displaying three different expressions (happy, sad, neutral).
2) we provide two baseline experiments in which we evaluate the similarity between human attention and the Vision Transformer and human attention and handcrafted features, which are state of the art methods used in face recognition and face detection.

The paper is structured as follow: In section II, we briefly survey human gaze datasets present in the literature and related works; in section III we thoroughly describe the Uniss-FGD dataset by providing full details on the device used for data capture, the acquisition process, the data specifications and the data validation; in section IV we propose a set of baselines experiments to illustrate the dataset potential; in section V we presents the results of the experiments; and in section VI we draw conclusions, outlining some additional research directions for future investigations.

II. BACKGROUND

In the last two decades, a considerable number of datasets related to human gaze have been collected. This underscores the interest of the scientific community in this type of data. A comprehensive list of human gaze datasets can be found at the MIT/Tübingen Saliency Benchmark page [6].

Restricting our attention to face stimuli, in table 1 we report a summary of the most recent datasets.

In [12], seven subjects viewed 200 images that included frontal faces and 50 images that did not include faces but were otherwise identical. This dataset was used to establish that human faces are very attractive to observers and to test models of saliency that included face detectors.

Saliency in crowd scenes is the focus of the research in [13], 16 subjects viewed 500 images representing indoor and outdoor scenes with diverse crowd densities, from a few faces

FIGURE 1. Image from the Tobii pro studio eye tracking software settings window, showing a segment of the sequence containing three stimuli shown to the observers interleaved by black screens. The stimuli, from left to right, are the KDEF images AF09SAS.JPG, AM14NES.JPG and AM02NES.JPG.

(3-5) to hundreds (up to 268). The images were sourced from Flickr and Google Images. The authors identified key features that contribute to saliency in crowds and analyzed their roles with varying crowd densities.

In [14], 450 subjects viewed video clips featuring 8 different actors. The actors were positioned against a green background, with the point between their eyes aligned with the center of the screen. In the video clips, the actors moved their eyes up and down while keeping their head still and maintaining a neutral facial expression.

In [15], the authors present a dataset that aggregates data from 23 different studies. All studies allowed for free eye movements and differed in the age range of participants (7–80 years). Two studies are based on facial image stimuli: Face Discrimination (ID 18) and Face Learning (ID 19). Face Discrimination investigated eye movements during a face discrimination tasks using 32 faces. Face Learning tested the effect of aversive associative learning on the exploration of faces using 8 faces.

With respect with the previously surveyed datasets, our proposed dataset presents some additional important features:

- The stimuli are based on a large number of high-quality images of faces with different expressions, acquired without the presence of other objects or backgrounds.
- A large amount of data is extracted from each acquisition (see Appendix for a list of available data)

Investigations on the salient areas of images have become a critical topic in scientific research since Medathati et al. [23] showed the existence of a strict connection between visual attention and eye movements.

Switching from humans to machines, in Machine Learning attention is a mechanism that allow models to learn a relative importance of their inputs. In simple terms, it allows a model to focus on certain more informative parts of an input. Many models have been proposed which employ attention [24], [25], [26], perhaps the most well known is the Transformer [27] which computes attention exhaustively between sub-sections of the input.

In [28] the authors investigate if the self-attention modules in ViTs have similar effects to human attentive visual processing. The paper reveals a gap between human visual attention and the mechanisms implemented in ViT.

In [29], the authors designed a Transformer-based framework for Facial Expression Recognition (FER) based on Patch-Range-Attention (PRA) module to resolve the criticality of CNN-based methods in learning long-range biases to improve capacity in FER tasks. ViT is used to extract the picture patches that are too simple. Four FER datasets were considered to analyze the three different attention mechanisms in the proposed algorithm. The dataset shows disturbing elements that can influence the observer by shifting attention towards points of the image that cannot be traced back to the desired task.

In [30], a ViT has been optimized to extract salient features from images in order to improve speed and scalability of human activity recognition. The suitability of the proposed method has been verified in resource-constrained and real-time environments, but a comparative analysis with human behavior is missing.

Reference [31] presents the Domain-Adaptive ViT (DA-ViT) model, which merges human cognitive perspective to obtain domain generalization. Glance and gaze blocks are considered to initially capture general information from each block and subsequently acquire more detailed and focused information.

In [21] the authors investigate how Convolutional Neural Networks and Transformers ''look'' at general images compared with humans. Through metrics for examining error-consistency they showed that Transformers are more consistent with humans than CNNs.

III. Uniss-FGD DATASET

A. STIMULI

Stimuli consisted of facial images from the Karolinska Directed Emotional Faces (KDEF) [32], [33], [34] database. The KDEF DB consists of 4900 facial images of 70 individuals. Each individual was acquired while mimicing 7 different emotional expressions. Each one of the expressions was photographed (twice) from 5 different angles. All subjects were seated at a distance of approximately three meters from the camera. The lights were set to cast a soft indirect light evenly distributed at both sides of the face [34]. For our stimuli set, we curated a selection of 120 different images from the frontal images of the expressions ''neutral'',

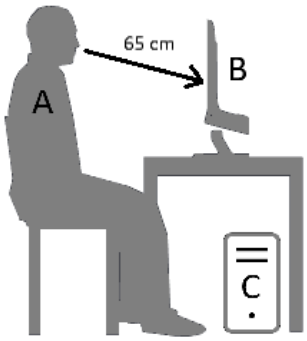

FIGURE 2. Graphical representation of the configuration employed for the data acquisition process: (A) denotes the observer, positioned in front of the screen at a distance of 65 cm. (B) represents the Tobii Pro TX300, which incorporates an eye tracker and an integrated screen. (C) indicates a workstation connected to the Tobii, with the eye tracker software running on it.

''happy'', and ''sad'' of the first 20 female and the first 20 male individuals in KDEF. A random function was used to generate a list of the 120 images which was therefore not ordered neither with respect to the subjects nor to the expressions. The sequence of faces was shown in the same order to all observers (see figure 1).

For full reproducibility of the experiments, a thorough list of the selected images has been made available at *https://github.com/CVLab-Uniss/Uniss-FGD*.

B. OBSERVERS

We collected eye gaze data from 20 Italian observers, 10 self-identifying as female and 10 self-identifying as male, recruited among university students or staff of two different degree courses. The gender labels of the observers is available at *https://github.com/CVLab-Uniss/Uniss-FGD*. They were verbally informed that their names would not be asked, that they would be enrolled in the database with a numerical ID and that the gaze data acquired from them could not lead to their identity. Only the sex and age of the participants were asked and associated with their numerical ID. The gathered data is therefore not classified as ''personal data'' by the EU General Data Protection Regulation and participant consent was not required. 18 of the observers were students from 19 to 24 years of age, 2 of observers were academic staff aged 30 and 50. All participants reported normal or corrected-to-normal vision. We must highlight that both culture and age play a role in human attention patterns when viewing faces [35]. The results obtained by exploiting the provided data cannot therefore generalize to non westerns subjects nor to old or very young age people.

C. ACQUISITION SETTING AND PROCEDURE

The acquisitions took place at the University of Sassari, in a dedicated 8 × 6 x 3.10 m (l x w x h) room equipped with a chair and a desk, a Tobii Pro TX300 Eye Tracker [36] connected to a PC running the Tobii Pro Studio Eye Tracking software [37], and a WiFi internet connection. A graphical representation of the acquisition set-up is depicted in figure 2.

The Tobii tracker performs a video-based pupil and corneal reflection eye tracking with dark and bright pupil illumination modes. Two cameras capture stereo images of both eyes for robust, accurate measurement of the eye gaze and eye position in 3D space, as well as pupils diameter. The sampling frequency is 300Hz. More detailed specifications and information about data quality can be found in section III-E, ''Technical Validation''. The acquisition environment, timing and duration of the experiment were determined based on recommendations from the ITU-T P.911 [38], in order to mitigate observer fatigue. Observers sat at 65 cm from the Tobii eye tracker, a distance equal to 2.5 times the height of the Tobii monitor.

The observers were informed that they would be shown a sequence of images of faces and that no particular task was specified, so they could freely look at the images. Each face image from the curated set was displyed on screen for three seconds and interleaved with two seconds of black screen (figure 1). Each capture session took approximately 10 minutes.

D. DATASET REPOSITORY

The Uniss-FGD dataset is stored in the following public repository:

- Repository name: CVLab-Uniss
- URL: *https://github.com/CVLab-Uniss/Uniss-FGD*

In the main directory have been provided:

- a ''readme.md'' file where users can find all the main information relating to the indexing and use of the data contained in the dataset;
- a ''security.md'' file where users can find all advices about dissemination and reuse of data;
- a folder called FGD containing the data, that is 120 files in csv format.

The fixations and saccades contained in the dataset are extracted from the row gaze data with the Tobii Pro Studio Eye Tracking software [37]. The acquired data was filtered by using the software Tobii Studio. In particular, Tobii Studio uses Stampe stage 2 algorithm [39] to remove noise from the raw gaze data.

All gaze data in Tobii Studio are mapped into a coordinate system. There are three available coordinate systems [37]:

- Active Display Coordinate System pixels (ADCSpx): Data types with the extension ADCSpx provide data mapped into a 2D coordinate system aligned with the Active Display Area, which is the Tobii screen area. The origin of the ''Active Display Coordinate System pixel'' (ADCSpx) is at the upper left corner of the Active Display Area.
- Active Display Coordinate System millimeters (ADCSmm): Data types with the extension ADCSmm provide data mapped into a 3D coordinate system aligned with the Active Display Area which is the screen area. The origin of the ''Active Display Coordinate System millimeter'' (ADCSmm) is at the bottom left corner of the Active Display Area.

TABLE 2. Validity codes.

(VLC, VRC)	No samples (%)	Description
$(0, 0)$	2,073,679 (99.11)	Found two eyes
$(0, 4)$	2,475 (0.12)	Found one eye. Most probably the left eye
$(1, 3)$	10,424 (0.50)	Found one eye. Probably the left eye
$(2, 2)$	758 (0.04)	Found one eye. The tracker cannot with any certainty determine which eye it is
$(3, 1)$	2,398 (0.11)	Found one eye. Probably the right eye
$(4, 0)$	2,607 (0.12)	Found one eye. Most probably the right eye

- Media Coordinate System pixels (MCSpx): Data types with the extension MCSpx provide data mapped into a 2D coordinate system aligned with the media. The origin of the coordinate system is at the top left of the media shown to the participant being eye tracked.

Each file (one for each of the 120 images showed to the observers during the test) contains a list of gaze events where each row is a single gaze event (i.e., fixation or saccade) and the columns summarise different information, as reported in Appendix.

E. TECHNICAL VALIDATION

The Tobii Pro TX300 Eye Tracker, with a sampling rate of 300Hz, allows robust tracking and compensation for large head movements. This ensure a very high precision and accuracy of the captured data. Regarding the instrument's sensitivity, according to the official documentation the mean accuracy is 0.55° while the mean precision 0.13° at 65 cm. Gaze accuracy is the angular average distance from the actual gaze point to the one measured by the eye tracker, while gaze precision is the spatial variation between individual gaze samples. Both are typically measured in degrees of visual angle, where one degree accuracy corresponds to an average error of 11 mm on a screen at a distance of 65 cm. These values have been calculated through extensive tests to measure and report performance and data quality [40].

In human populations, there exists natural variation in the shape and geometry of the eyes. To address this variation, the calibration procedure provided with the eye tracker has been utilized to optimize the gaze estimation algorithms. The calibration procedure was supervised by experienced researches before the recording of each participant. During calibration, participants are instructed to focus on calibration targets appearing at multiple locations on the display monitor where the stimulus is located. The speed of the calibration was set to medium, a number of nine calibration locations were selected and and the full screen was used.

The calibration procedure consists of three distinct phases: (i) Data collection phase: participants are directed to fixate on a predefined number of targets sequentially displayed on the screen. (ii) Optimization phase: continuous recalibration of the distance and distribution between the mapped data and the actual location of calibration targets to refine the 3D eye model. The model includes information about shapes, light refraction and reflection properties of the different parts of the eyes (e.g., cornea and placement of the fovea) [37]. (iii) validation phase: New targets are presented to validate the updated 3D eye model configuration, and data quality measures are reported. Calibration is performed only once before data collection begins and does not require adjustments during recording.

At the end of the calibration procedure, after validation, if for one or more of the nine calibration locations the supposed point of gaze was far from the measured point of gaze, the calibration was repeated. Particular care was taken for the points in the central area, since face images were shown at the center of the screen, so common miscalibrations that tipically occur in the lower area of the screen did not affect our samples. If the calibration was repeatedly unsuccessful (i.e., the calibration result "Not Enough Calibration Data" was displayed), the participant was checked for any factors that could interfere with pupil detection (e.g., an infrared light source that is directed at the eye tracker sensor or the participant's eyes, dirty or scratched glasses or droopy eyelids, make-up) before starting a new calibration procedure.

Since each acquisition took approximately 10 minutes, for each observer around 500.000 samples have been gathered. To ensure good quality of the data, samples with low level of confidence (see discussion below about validity code) have been excluded. Thus for each observer the number of valid samples (and excluding samples corresponding to black screens between the stimuli) is about 104.600, which amounts to 2.092.341 valid samples for the whole dataset (all observers on all stimuli).

To estimate the quality of the samples, for each one of them we considered the Tobii validity code, which estimates the probability that each one of the two eyes was detected during that sampling:

- VLC - Validity Left Code - indicates the confidence level that the left eye has been correctly identified. Code values are integer number from 0 (high confidence) to 4 (eye not found);
- VRC - Validity Right Code - indicates the confidence level that the right eye has been correctly identified. Code values are integer numbers from 0 (high confidence) to 4 (eye not found).

The validity codes are paired as (VLC, VLR) and can assume values in the set $\{(0, 0), (0, 4), (4, 0), (1, 3), (3, 1), (2, 2)\}$ whose meanings are explained in table 2. The Tobii automatically discards the (4, 4) pair codes, so samples for which no eyes were located are not present in the dataset. For

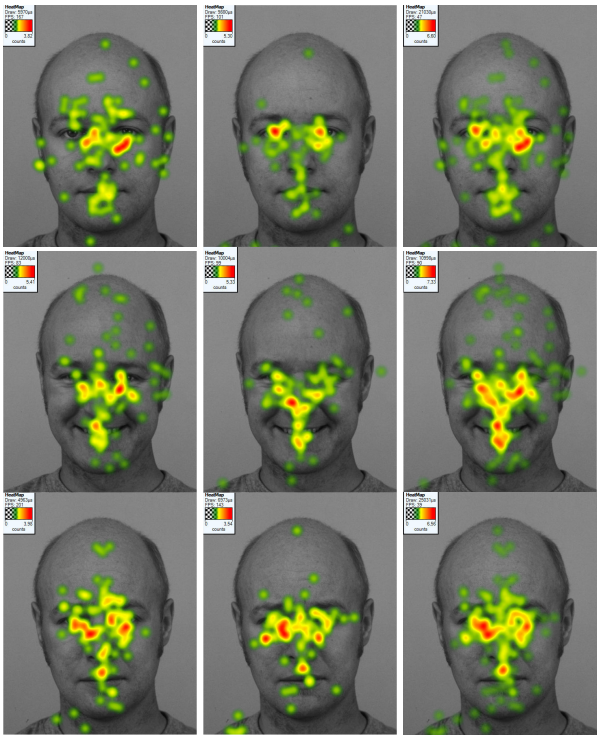

FIGURE 3. Fixation densities on neutral expression (first row), happy (second row) and sad (third row) filteres by male observers (left column), female observers (central column) and all observers (right column). The KDEF stimuli are AM09NES.JPG (first row), AM09HAS.JPG (second row) and AM09SAS.JPG (third row).

each observer, the validity code pair of each sample relative to the simuli (excluding the black screens) was checked. The cumulative results are reported in the second column of table 2, as the total number of code pairs and in percentage. For over 99% of samples both eyes were correctly identified, which gives a measure of the high quality of the data.

Figures 3 and 4 show some examples of fixation data visualization:

- Figure 3, first column depicts the density of the cumulative fixations of all male observers on a face image from the stimuli set in a neutral expression (first row), a happy expression (second row) and a sad expression (third row). The second column is relative to all female observers and the third on all 20 observers (male and female).
- Figure 4, first row, shows the fixation sequences of the same observer over images of the expressions neutral, happy and sad of the same subject of the stimuli set. The second row, shows the fixation sequences of three different observers when looking at the same stimulus. Notice that the radius of the fixation disk is proportional to the time the observer fixated that image point.

F. DATA AVAILABILITY AND USABILITY

The authors defined a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other

FIGURE 4. First row: fixation paths of the same participant observing three different emotions of the same KDEF subject (left to right, the KDEF stimuli are AM14NES.JPG, AM14HAS.JPG and AM14SAS.JPG). Second row: fixations paths of three different observers on the AF09SAS.JPG image of the KDEF.

dissemination tools, official publications and other published documents, can be reused without any payment or written licence provided that:

- the source is indicated as CVLab-Uniss;
- manuscripts that present work that makes use of the dataset cite this paper;
- when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information.

The distributed csv files were generated from the row gaze data with the Tobii Pro Studio Eye Tracking software [37]. Data may be extracted from the repository into the target computing environment with traditional csv import functions.

The Uniss-FGD dataset, given its richness and variety of eye-tracking data provided (see Appendix for a complete list), holds significant potential to contribute to research efforts in numerous scientific domains. Some possible application which can benefit from the Uniss-FGD are briefly discussed below.

By leveraging data on eye movements, including fixation locations, saccades directions, and gaze duration, it becomes feasible to instruct a social robot to observe humans in a natural and human-like manner [41], [42]. This data can be utilized to train the robot on how to gaze at a person during verbal interaction, while also responding to various expressions. The same approach can also be employed for AI-driven non-player characters (NPCs) that will populate the Metaverse [43].

In neuroscience, eye-tracking methods and techniques are extensively employed to unobtrusively investigate alterations in eye movement or oculomotor problems, which are considered evidence of neurodegenerative diseases such as Parkinson's. This analysis involves examining pupil size, eye position, fixation duration and locations [7], [44].

Eye gaze has demonstrated relevance in the security and privacy domain as well [45]. Eye-tracking data, for instance, can be utilized for purposes such as authentication, privacy protection, and gaze monitoring during security-critical tasks. The eye movement data provided by the Uniss-FGD dataset is particularly suitable for the development of implicit authentication algorithms. Research in the field mainly focuses on assessing unique eye movements, analyzing data such as fixation density map, angular saccade velocity or scan-paths, while individuals perform activities with varying visual stimuli and types [46]. These systems could potentially be employed in AR/VR headsets equipped with eye-tracking systems to verify the identity of a human who wishes to embody an avatar within future immersive digital worlds of the Metaverse [47], [48].

IV. BASELINE EXPERIMENTS

The dataset's utility has been validated through baseline experiments, designed to compare the fixation densities in Uniss-FGD dataset to two state-of-the-art techniques for face detection and recognition: Visual Transformers (ViT) and multi-scale handcrafted features. A selection of 120 images extracted from the KDEF database (i.e., the ones used to build the Uniss-FGD dataset) was used to produce the outputs with the two techniques. The final human fixations, ViT's attention maps or multi-scale densities are cropped in an area around the face, as the extremities are noisy and contain only background.

A. VISION TRANSFORMERS AND ATTENTION MAP EXTRACTION

The Transformer, originally designed as a state-of-the-art architecture for Natural Language Processing, has demonstrated remarkable efficacy in the computer vision domain, including tasks such as face recognition [49]. ViTs [50], [51] are sequence-based models that process input by splitting it into distinct *tokens*, which are then embedded. Self-attention is a key mechanism employed, where the relationships between token embeddings determine their relative importance. This entire sequence of token embeddings undergoes simultaneous processing. The output is aggregated through feed-forward networks and a non-linearity, forming a single Encoder. Initially, positional information is incorporated into the input. For classification tasks, a *CLS* token can be added, attending to all tokens and creating a comprehensive description of the input. However, the quadratic complexity of attention computations between all inputs limits pixel-level processing. Therefore, input images are typically *tokenized* into patches (e.g., 16×16 or 32×32 pixels), which are then flattened through an embedding layer. It's noteworthy that various alternative strategies are explored in the field [52], [53], [54]. The general architecture of a ViT is illustrated in figure 5.

The Transformer, like any Machine Learning model, requires a clearly defined task or goal for it to learn its representation. We have arbitrarily chosen recognition, which is a comparable task to the free-viewed human fixations in Uniss-FGD. Moreover, Transformers are known for their strong general representative capacity. For example, BERT trained to reconstruct randomly masked sentences has been shown to be an excellent sentiment classifier with only very limited extra training [55]. We believe this general representative capacity to be key in allowing for meaningful comparison between the two domains. The main characteristics of the ViT initially used to maximise the resolution of the output attention maps are summarised in table 3.

TABLE 3. ViT setting.

Setting	Description
Encoders	12
Attention Heads	12
Input Image Size	384x384
Patch Size	16x16
PreTrained model	ImageNet21k
Fine-tuned model	ImageNet-1k

The model is fine-tuned using controlled images from the Face Recognition Grand Challenge (FRGC) database [56]. During training, faces are tightly cropped based on the distance between the eyes to minimize background interference while preserving facial details and maintaining roughly equal scale. Despite the small dataset size, the aim is to focus the Transformer's internal representation to extract meaningful attention maps. The images for a single identity are randomly split into 90% for training and 10% for validation. A batch size of 16, a learning rate of $1e^{-4}$, and a weight decay of 0.01 are employed. Cross Entropy loss is utilized with the AdamW optimizer. The model achieves perfect classification convergence within 15 epochs on an RTX 3090 paired with an Intel Core i9-10980xe.

Attention maps are extracted with Attention Rollout [57]. This technique not only considers the attention at the final layer, but also the attention as it flows through the model. Additionally, each Transformer layer has multiple attention *heads*, we elect to average the attention across these. Finally, we only consider attention flowing to the *CLS* token. The final attention maps have a resolution of 24×24.

B. MULTI-SCALE HANDCRAFTED FEATURES

In order to extract significant handcrafted features we apply the well-known scale-space theory developed by Lindeberg [58]. Indeed, it was shown that persistent points that characterize some kind of visual information like a face naturally emerge at different scale levels [59], with no need to pre-determine the number of features or the spatial scales that better represent the image information in a bottom down fashion.

Given an image, a Gaussian scale-space representation is defined to be a map:

$$L(x, y; \sigma) = \int_{(u,v)\in\mathbb{R}^2} f(x-u, y-v)g(u, v; \sigma)dudv \quad (1)$$

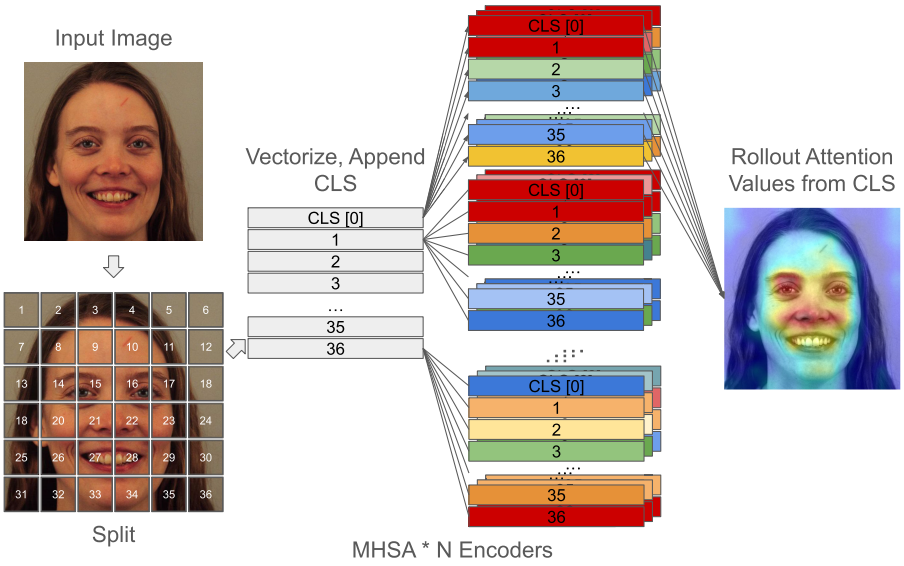

FIGURE 5. Attention extraction from a ViT.

where $g(u, v; \sigma) = \frac{1}{2\pi t} e^{-(u^2+y^2)/2\sigma}$ is the Gaussian kernel of variance σ, which represents the scale parameter. Starting from this representation, given a scale σ, Lindeberg defines four spacial differential operators based on the Hessian matrix H_L of L, each leading to a particular type of feature points. Among the operators we chose the Laplacian, defined as:

$$\nabla^2 L = L_{xx}^2 + L_{yy}^2 = \lambda_1 + \lambda_2 \tag{2}$$

where λ_1 and λ_2 are the eigenvalues of the Hessian matrix, or the principal curvatures of $L(, , \sigma)$.

Extrema of $\nabla^2 L$ correspond to dark or bright blobs, according to whether the Hessian is positive or negative definite. Edges will also be detected, but they are discarded to improve the repeatability of points detection.

For the Laplacian operator a pyramid of 10 layers was built, one for each scale, starting from the original image, and halving the image every two steps. For each scale, local extrema were calculated with respect to the image coordinates. Most of these extrema are likely to persist across two or more scales. Scale linking as described in Lindeberg has been carried out to select their strongest response across scales. The Laplacian features on a given face image are the resulting Laplacian extrema extracted from the 10 layers pyramid.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Here we compare human attention to ViT and Laplacian features on the face images from the stimuli set of Uniss-FGD. The attention maps of a ViT on each image of a subset of the stimuli are extracted with Attention Rollout as described in section IV-A. These attention maps are then summed and normalized. The Laplacian features over a subset of face images are the union of the features on each of the face images in the subset. To estimate a probability density function of the union of them, they are fed into a kernel density estimation based on diffusion [60]. As for the human attention, we cumulate all fixations from all observers from Uniss-FGD on a subset of the stimuli set. Cumulating observers is a necessary step, since humans have different strategies to observe faces leading to idiosyncratic gaze paths whose fixations cluster around a few facial areas (one of the two eyes or between them, or the mouth [14]). To estimate the variation of the cumulative fixations across observers, we do a statistical analysis of the location of the center of their distributions. The 95% confidence intervals of the distribution centres' coordinates are reported in table 4 for the distributions obtained by cumulating fixations on neutral, happy, sad and on the whole set of stimuli.

TABLE 4. 95% confidence intervals for the *x* and *y* coordinates of the centers of the distributions on Neutral, Happy, Sad and All stimuli image faces.

	Neutral	Happy	Sad	All
x	285.4 ± 5.7	285.5 ± 5.6	286.5 ± 5.9	285.8 ± 5.7
y	423.7 ± 7.9	423.6 ± 7.9	425.0 ± 8.3	427.7 ± 7.9

Each of the four cumulative fixations set is fed into the kernel density estimation in [60] to estimate a probability density function.

To compare the density functions we employ three different metrics:

1) Jensen Shannon similarity which is defined for two probability distributions P and Q as:

$$JSD(P\|Q)_{sim} = 1 - \frac{1}{2}(KLD(P\|M) + KLD(Q\|M)) \tag{3}$$

where KLD refers to the Kullback-Leibler Divergence, defined as:

$$KLD = \sum_x P(x) \ln \frac{P(x)}{Q(x)} \tag{4}$$

FIGURE 6. Fixation densities, attention maps and handcrafted features over happy, neutral, sad and all images (from left to right). On the first row are human fixation densities, on the second: ViT attention maps, on the third: ∇^2 densities.

and

$$M = \frac{1}{2}(P + Q) \tag{5}$$

2) χ^2 similarity:

$$\chi^2_{sim} = 1 - \sum_x \frac{(P(x) - Q(x))^2}{(P(x) + Q(x))} \tag{6}$$

3) and the Pearson correlation coefficient.

Each of the three metrics sheds light on the similarity between two density functions. The Jensen Shannon similarity is widely used to compare density functions and looks at how different the two densities are from their average, while the χ^2 similarity can be seen as a weighted Euclidean distance between probability values. The Pearson correlation estimates the covariance between the two probability distributions and, unlike the previous metrics, it establishes the linear relationship between them.

We designed two experimental protocols, one to compare the attention of humans to ViT and Laplacian features on human faces with different expressions, the other to compare humans to ViT and Laplacian features on faces showing one of the three expression happy, neutral and sad.

For the first protocol, we consider all image stimuli of Uniss-FGD, and cumulate all fixations and Laplacian features to estimate the respective density maps. On the same stimuli set, we extract the ViT attention maps. We then compare the human density function to the Laplacian one and the ViT attention maps with the three chosen metrics. For the second protocol, we consider three sets of stimuli, each corresponding to one of the three expressions. On each set, we estimate the density functions of human fixations and Laplacian features and the ViT attention map and we compare them with the three metrics.

The density functions generated by all expressions are plotted in the last column of figure 6. The attention of humans and ViT are quite similar. Humans tend to be less interested by lateral areas such as the ears, where the ViT attention falls, although in a marginal way. The similarity measures between the fixations density and the ViT attention map are reported in the bar plot on the left of figure 7. They confirm the visual similarity of the two attentions, with a correlation of over 60% and the JS and χ^2 over 80%. The same similarities are found for the stimuli subsets of neutral and sad expressions, while in the case of happy expressions there are some differences in the mouth area, where the ViT seem to attend to the whole lip/teeth arch, compared to humans who just scan the area in a vertical direction down to the center of the mouth. This reflects to the lower similarity measures (especially the correlation) reported in the bar plot.

The results from the two experimental protocols are shown in figure 7 and figure 6, where the various densities are superimposed over an image from the stimuli set. A generally higher similarity is evident for the Transformers. This outcome is observed both for the densities of individual expressions and for the density over the whole stimuli set.

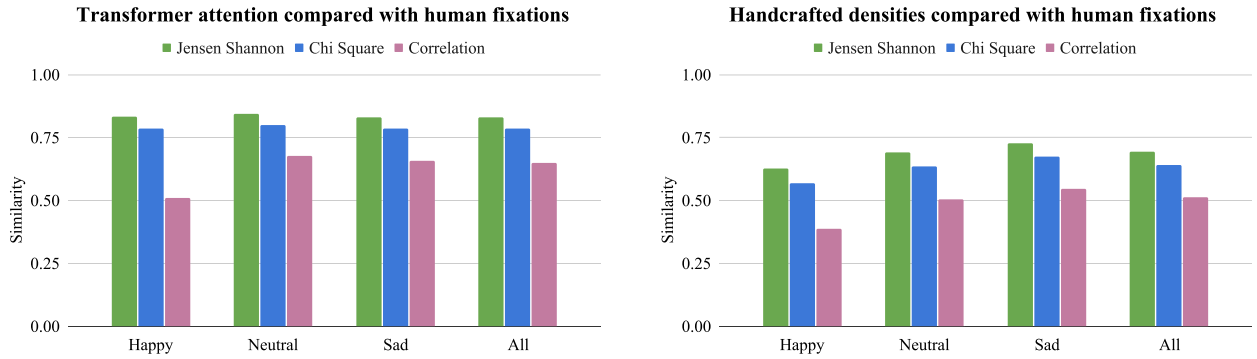

FIGURE 7. Human fixation densities similarity measurements from Transformer (left) and handcrafted features (right), split by expression and cumulative (all).

The densities of the Laplacian features shown in the last row of figure 6, reflect the fact that the ∇^2 points extracted are clustered around smaller areas around the eyes, nose and mouth with respect to the ViT and the fixations. This is because the ∇^2 stronger extrema selected by the process correspond to blobs that are persistent at multiple scales, so flat areas are not attended. The densities exhibit more variations with expressions. The bar plot reveals they are most similar to fixation densities on the sad stimuli, while they differ the most on the happy stimuli, where most of the ∇^2 attention falls on the open mouth.

All in all, the experimental results show there is some intersection between the facial areas where the attention of the ViT and of the Laplacian features falls and the areas where the human attention falls. The intersection is stronger between humans and the ViT, which highlights the fact that a system that is allowed to optimize its features, at the end of the training process it produces features that are similar to human fixations. However, the ViT seems to get distracted by uninformative areas of the face, such as around the ears, so the idea to include human attention in the ViT training process could lead to a more efficient model.

Regarding the Laplacian features, they seem to be less distracted by uninformative areas. However, due to their nature, they are located on areas of large variation between pixels so that in happy faces, when all 10 scales are considered as in this experiment, they tend to concentrate on the mouth. By guiding the scales selection through human fixations, in [59] it is shown that a face detection method based on Laplacian features can perform better and more efficiently than it would if all scales were considered, which proves the potential of exploiting human attention for face detection.

VI. CONCLUSION

This paper introduces Uniss-FGD, a novel dataset that collects human gaze data on face images. The gathered data can be a precious resource for research investigations across different scientific fields, from Neuroscience to Physiology, Psychology, Human-Robot Interaction, and Computer Science. Since humans exhibit an extraordinary efficiency at tasks such as recognition, reaching levels still unattainable by automatic computer systems, human gaze tracking can be immensely useful for training and optimizing automatic recognition systems, such as those based on neural networks.

The data collected in Uniss-FGD were acquired, using a professional eye tracking tool, from 20 observers who viewed 120 images, each for 3 seconds. The avaliable data includes fixations, saccades and a variety of raw data. The dataset's utility has been validated through some baseline experiments which compare fixation densities in the Uniss-FGD dataset with two state-of-the-art techniques for face detection and recognition: Visual Transformers (ViT) and multi-scale handcrafted features based on the Laplacian operator. We conducted a comprehensive comparison of Laplacian features and Transformers to human observers by using three similarity measures. Furthermore, we compared the attentions mechanisms on each of the facial expressions of the image stimuli. The results reveal a stronger similarity of ViT and human attentions, which holds true for each expression.

Future work will involve assessing methodologies for integrating Uniss-FGD data into the attentive mechanisms of advanced face detection and recognition systems, especially those based on Deep Learning techniques. These methods, by leveraging the effectiveness of human vision, aim to enhance performance by prioritizing salient aspects of faces, thus reducing the computational resources required for training.

The Uniss-FGD dataset is publicly available for academic research purposes. Full details on how to download the Uniss-FGD database can be found on the project website: *https://github.com/CVLab-Uniss/Uniss-FGD*.

APPENDIX

UNISS-FGD STRUCTURE AND RECORDS

The Uniss-FGD dataset contains a collection of gaze data relative to fixations and saccades. The dataset is structured with a main folder containing 120 CSV files. The size of the full dataset is about 90 MB.

Each file (one for each of the 120 images showed to the observers during the test) contains a list of gaze events where

each row is a single gaze event (fixation or saccade) and the columns, in order from left to right, correspond to:

- **ParticipantName**: unique anonymous identification number associated to each observer.
- **RecordingDate**: Date when the recording was performed (Year, Month, Day).
- **RecordingDuration**: The duration of the recording (Milliseconds).
- **RecordingResolution**: The resolution of the screen or of the video capture device used during the recording.
- **RecordingTimestamp**: Timestamp counted from the start of the recording (t0=0). This timestamp is based on the internal computer clock of the computer running Tobii Studio. This clock is regularly synchronized with the eye tracker clock in order to ensure that the timestamps of the gaze data is accurate in relation to other events such as when media is shown or participant generated events such as mouse clicks (Milliseconds).
- **MediaPosX (ADCSpx)**: Horizontal coordinate of the left edge of the eye tracked media (pixels).
- **MediaPosY (ADCSpx)**: Vertical coordinate of the top edge of the eye tracked media (Pixels).
- **MediaWidth**: Horizontal size of the eye tracked media (Pixels).
- **MediaHeight**: Vertical size of the eye tracked media (Pixels).
- **FixationIndex**: Represents the order in which a fixation event was recorded. The index is an auto-increment number starting with 1 (first gaze event detected).
- **SaccadeIndex**: Represents the order in which a saccade event was recorded. The index is an auto-increment number starting with 1 (first gaze event detected).
- **GazeEventType**: Type of eye movement event classified by the fixation filter settings applied during the gaze data export (Fixation; Saccade; Unclassified).
- **GazeEventDuration**: Duration of an eye movement event (Milliseconds).
- **FixationPointX (MCSpx)**: Horizontal coordinate of the fixation point on the media. Column empty if: Fixation is outside media, Media is covered, No media is displayed. (Pixels)
- **FixationPointY (MCSpx)**: Vertical coordinate of the fixation point on the media. Column empty if: Fixation is outside media, Media is covered, No media is displayed. (Pixels)
- **SaccadicAmplitude**: Distance in visual degrees between the previous fixation location and the current fixation location as defined by the fixation filter (Degrees).
- **AbsoluteSaccadicDirection**: Offset in degrees between the horizontal axis and the current fixation location where the previous fixation location is set as the origin (Degrees).
- **RelativeSaccadicDirection**: The difference between the absolute saccadic direction of the current and previous saccade where the current saccade is between the current and previous fixation (Degrees).
- **GazePointIndex**: Represents the order in which the gaze sample was acquired by Tobii Studio from an eye tracker. The index is an auto-increment number starting with 1 (first gaze sample)
- **GazePointLeftX (ADCSpx)**: Horizontal coordinate of the unprocessed gaze point for the left eye on the screen (Pixels).
- **GazePointLeftY (ADCSpx)**: Vertical coordinate of the unprocessed gaze point for the left eye on the screen (Pixels).
- **GazePointRightX (ADCSpx)**: Horizontal coordinate of the unprocessed gaze point for the right eye on the screen (Pixels).
- **GazePointRightY (ADCSpx)**: Vertical coordinate of the unprocessed gaze point for the right eye on the screen (Pixels).
- **GazePointX (ADCSpx)**: Horizontal coordinate of the averaged left and right eye gaze point on the screen (Pixels).
- **GazePointY (ADCSpx)**: Vertical coordinate of the averaged left and right eye gaze point on the screen (Pixels).
- **GazePointX (MCSpx)**: Horizontal coordinate of the averaged left and right eye gaze point on the media element. Column empty if: Fixation is outside media, Media is covered, No media is displayed (Pixels).
- **GazePointY (MCSpx)**: Vertical coordinate of the averaged left and right eye gaze point on the media element. Column empty if: Fixation is outside media, Media is covered, No media is displayed (Pixels).
- **GazePointLeftX (ADCSmm)**: Horizontal coordinate of the unprocessed gaze point for the left eye on the screen (Millimeters)
- **GazePointLeftY (ADCSmm)**: Vertical coordinate of the unprocessed gaze point for the left eye on the screen (Millimeters).
- **GazePointRightX (ADCSmm)**: Horizontal coordinate of the unprocessed gaze point for the right eye on the screen (Millimeters).
- **GazePointRightY (ADCSmm)**: Vertical coordinate of the unprocessed gaze point for the right eye on the screen (Millimeters).
- **StrictAverageGazePointX (ADCSmm)**: Horizontal coordinate of the averaged gaze point for both eyes on the screen. ''average'' function similar to the one used for Eye selection (Millimeters).
- **StrictAverageGazePointY (ADCSmm)**: Vertical coordinate of the averaged gaze point for both eyes on the screen. ''average'' function similar to the one used for Eye selection (Millimeters).
- **EyePosLeftX (ADCSmm)**: Horizontal coordinate of the 3D position of the left eye. (Millimeters).
- **EyePosLeftY (ADCSmm)**: Vertical coordinate of the 3D position of the left eye (Millimeters).

- **EyePosLeftZ (ADCSmm)**: Distance/depth coordinate of the 3D position of the left eye (Millimeters).
- **EyePosRightX (ADCSmm)**: Horizontal coordinate of the 3D position of the right eye (Millimeters).
- **EyePosRightY (ADCSmm)**: Vertical coordinate of the 3D position of the right eye (Millimeters).
- **EyePosRightZ (ADCSmm)**: Distance/depth coordinate of the 3D position of the right eye (Millimeters).
- **PupilLeft**: Estimated size of the left eye pupil. The Tobii Eye Trackers aim to measure the true pupil size, i.e. the algorithms take into account the magnification effect given by the spherical cornea as well as the distance to the eye (Millimeters).
- **PupilRight**: Estimated size of the right eye pupil. The Tobii Eye Trackers aim to measure the true pupil size, i.e. the algorithms take into account the magnification effect given by the spherical cornea as well as the distance to the eye (Millimeters).
- **ParticipantGender**: information about the sex of each (anonymous) observer.

REFERENCES

[1] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognition*, vol. 7, nos. 1–3, pp. 17–42, Jan. 2000.

[2] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[3] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 541–554, Mar. 2013.

[4] P. Ruiu, L. Mascia, and E. Grosso, "Saliency-guided point cloud compression for 3D live reconstruction," *Multimodal Technol. Interact.*, vol. 8, no. 5, p. 36, 2024.

[5] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," 2015, *arXiv:1505.03581*.

[6] MIT. (2023). *MIT Tübingen Saliency Benchmark Datasets*. [Online]. Available: https://saliency.tuebingen.ai/datasets.html

[7] M. Gorges, E. H. Pinkhardt, and J. Kassubek, "Alterations of eye movement control in neurodegenerative movement disorders," *J. Ophthalmol.*, vol. 2014, pp. 1–11, Jan. 2014.

[8] M. S. Ekker, S. Janssen, K. Seppi, W. Poewe, N. M. de Vries, T. Theelen, J. Nonnekes, and B. R. Bloem, "Ocular and visual disorders in Parkinson's disease: Common but frequently overlooked," *Parkinsonism Rel. Disorders*, vol. 40, pp. 1–10, Jul. 2017.

[9] M. L. Mele and S. Federici, "Gaze and eye-tracking solutions for psychological research," *Cognit. Process.*, vol. 13, no. S1, pp. 261–265, Aug. 2012.

[10] A. Belardinelli, "Gaze-based intention estimation: Principles, methodologies, and applications in HRI," *ACM Trans. Hum.-Robot Interact.*, pp. 1–29, Apr. 2024, doi: 10.1145/3656376.

[11] M. Dirik, O. Castillo, and A. F. Kocamaz, "Gaze-guided control of an autonomous mobile robot using Type-2 fuzzy logic," *Appl. Syst. Innov.*, vol. 2, no. 2, p. 14, Apr. 2019.

[12] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2007, pp. 1–8.

[13] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8695, 2014, pp. 17–32.

[14] A. Coutrot, N. Binetti, C. Harrison, I. Mareschal, and A. Johnston, "Face exploration dynamics differentiate men and women," *J. Vis.*, vol. 16, no. 14, p. 16, Nov. 2016.

[15] N. Wilming, S. Onat, J. P. Ossandón, A. Açik, T. C. Kietzmann, K. Kaspar, R. R. Gameiro, A. Vormberg, and P. König, "An extensive dataset of eye movements during viewing of complex images," *Sci. Data*, vol. 4, no. 1, pp. 1–11, Jan. 2017.

[16] P. Ruiu, A. Lagorio, M. Cadoni, and E. Grosso, "Enhancing eID card mobile-based authentication through 3D facial reconstruction," *J. Inf. Secur. Appl.*, vol. 77, Sep. 2023, Art. no. 103577.

[17] Y. Han, G. Roig, G. Geiger, and T. Poggio, "Scale and translation-invariance for novel objects in human vision," *Sci. Rep.*, vol. 10, no. 1, p. 1411, Jan. 2020.

[18] M. I. Cadoni, A. Lagorio, E. Grosso, T. J. Huei, and C. C. Seng, "From early biological models to CNNs: Do they look where humans look?" in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6313–6320.

[19] M. Cadoni, S. Nixon, A. Lagorio, and M. Fadda, "Exploring attention on faces: Similarities between humans and transformers," in *Proc. 18th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2022, pp. 1–8.

[20] M. Cadoni, A. Lagorio, and E. Grosso, "Do CNN's features correlate with human fixations?" in *Proc. 3rd Int. Conf. Appl. Intell. Syst.* New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 1–6.

[21] S. Tuli, I. Dasgupta, E. Grant, and T. Griffiths, "Are convolutional neural networks or transformers more like human vision," in *Proc. Annu. Meeting Cognit. Sci. Soc.*, 2021, vol. 43, no. 43, pp. 439–455.

[22] M. Cadoni, A. Lagorio, S. Khellat-Kihel, and E. Grosso, "On the correlation between human fixations, handcrafted and CNN features," *Neural Comput. Appl.*, vol. 33, no. 18, pp. 11905–11922, Sep. 2021.

[23] N. V. K. Medathati, H. Neumann, G. S. Masson, and P. Kornprobst, "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision," *Comput. Vis. Image Understand.*, vol. 150, pp. 1–30, Sep. 2016.

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[25] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3104–3112.

[27] 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, pp. 5998–6008.

[28] P. Mehrani and J. K. Tsotsos, "Self-attention in vision transformers performs perceptual grouping, not attention," 2023, *arXiv:2303.01542*.

[29] Z. Lu and X. Tang, "Patch-range attention and visual transformer for facial expression recognition," in *Proc. 2nd Int. Conf. Electron. Inf. Eng. Comput. Technol. (EIECT)*, Oct. 2022, pp. 196–201.

[30] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 72227–72249, 2023.

[31] Y. Cho, J. Yun, J. Kwon, and Y. Kim, "Domain-adaptive vision transformers for generalizing across visual domains," *IEEE Access*, vol. 11, pp. 115644–115653, 2023.

[32] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The Karolinska directed emotional faces: A validation study," *Cognition Emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[33] D. Lundqvist, A. Flykt, and A. Ohman, "Karolinska directed emotional faces (KDEF)," *Database Records*, Jan. 1998, doi: 10.1037/t27732-000.

[34] D. Lundqvist, A. Flykt, and A. Ohman. (1998). *The Karolinska Directed Emotional Faces*. Figshare. [Online]. Available: https://www.kdef.se/home/aboutKDEF.html

[35] R. Caldara, "Culture reveals a flexible system for face processing," *Current Directions Psychol. Sci.*, vol. 26, no. 3, pp. 249–255, Jun. 2017.

[36] Tobii AB (Publ). (2013). *White Paper Tobii Eye Tracking: An Introduction to Eye Tracking and Tobii Eye Trackers*. Figshare. [Online]. Available: http://www.123seminarsonly.com/Seminar-Reports/2013-11/25907389-Tobii-Eye-Tracking.pdf

[37] Tobii AB (Publ). (2021). *Tobii Studio User's Manual*. figshare. [Online]. Available: https://stemedhub.org/resources/3374/download/TobiiStudio3.3Manual.pdf

[38] International Telecommunications Union (ITU). (2001). *Recommendation P.911 Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*. figshare. [Online]. Available: https://www.itu.int/rec/T-REC-P.911

[39] D. M. Stampe, "Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems," *Behav. Res. Methods, Instrum., Comput.*, vol. 25, no. 2, pp. 137–142, Jun. 1993.
[40] Tobii AB (Publ). (2020). *Field Metrics Test Report Accuracy, Precision, and Detected Gaze During Normal Usage With 400+ Particpants*. [Online]. Available: https://www.tobii.com/resource-center/data-quality#cta-section
[41] E. B. Onyeulo and V. Gandhi, "What makes a social robot good at interacting with humans?" *Information*, vol. 11, no. 1, p. 43, Jan. 2020.
[42] J. Urakami and K. Seaborn, "Nonverbal cues in human–robot interaction: A communication studies perspective," *ACM Trans. Hum.-Robot Interact.*, vol. 12, no. 2, pp. 1–21, Jun. 2023.
[43] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim, "Artificial intelligence for the metaverse: A survey," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105581.
[44] P. Tsitsi, M. N. Benfatto, G. Ö. Seimyr, O. Larsson, P. Svenningsson, and I. Markaki, "Fixation duration and pupil size as diagnostic tools in Parkinson's disease," *J. Parkinson's Disease*, vol. 11, no. 2, pp. 865–875, Apr. 2021.
[45] C. Katsini, Y. Abdrabou, G. E. Raptis, M. Khamis, and F. Alt, "The role of eye gaze in security and privacy applications: Survey and future HCI research directions," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–21.
[46] I. Sluganovic, M. Roeschlin, K. B. Rasmussen, and I. Martinovic, "Analysis of reflexive eye movements for fast replay-resistant biometric authentication," *ACM Trans. Privacy Secur.*, vol. 22, no. 1, pp. 1–30, Feb. 2019.
[47] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, "SoK: Authentication in augmented and virtual reality," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 267–284.
[48] Y. Zhang, W. Hu, W. Xu, C. T. Chou, and J. Hu, "Continuous authentication using eye movement response of implicit visual stimuli," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–22, Jan. 2018.
[49] S. Nixon, P. Ruiu, M. Cadoni, A. Lagorio, and M. Tistarelli, "Exploiting face recognizability with early exit vision transformers," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2023, pp. 1–7.
[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[51] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.
[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[53] Y. Zhong and W. Deng, "Face transformer for recognition," 2021, *arXiv:2103.14803*.
[54] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.
[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[56] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2005, pp. 947–954.
[57] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.
[58] T. Lindeberg, "Image matching using generalized scale-space interest points," *J. Math. Imag. Vis.*, vol. 52, no. 1, pp. 3–36, May 2015.
[59] M. Cadoni, A. Lagorio, and E. Grosso, "Face detection based on a human attention guided multi-scale model," *Biol. Cybern.*, vol. 117, no. 6, pp. 453–466, Dec. 2023.
[60] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010.

PIETRO RUIU received the master's degree in telecommunication engineering and the Ph.D. degree in electric, electronic, and communication engineering from the Polytechnic University of Turin, in 2006 and 2018, respectively. From 2013 to 2018, he held the position of the Head of the Infrastructure and Systems for Advanced Computing (IS4AC) Research Unit, Istituto Superiore Mario Boella (ISMB). His research interests include computer vision, face recognition, machine learning, computing infrastructures, datacenter networks and architecture, and automatic resource provisioning. He has actively contributed to the academic community by serving as a technical program committee (TPC) member for international conferences and as a reviewer for esteemed international journals.

MAURO FADDA (Senior Member, IEEE) received the Ph.D. degree in electronic and computer engineering from the University of Cagliari, in 2013. He is currently an Assistant Professor (RTD-B) with the University of Sassari, Italy. In March 2020, he was the elected Chair of Italian Chapter of the Broadcast Technology Society, Institute of Electrical and Electronic Engineering (IEEE). He served as the chair for various international conferences and workshops. He is an Associate Editor of IEEE Access and the Topic Editor of *Sensors*.

ANDREA LAGORIO received the degree in electronic engineering from the University of Genova, Italy, in 1999. Since 2010, he has been an Assistant Professor with the University of Sassari, Italy. He is currently an Assistant Professor of computer science. He is the author and coauthor of many publications in peer-reviewed journals, international conferences, and workshops. He participated in many national and international research projects. His research interests include biometric, face recognition, pattern recognition, and machine learning. In 2014, he received the Highest Impact Award by CVPR Biometrics Workshop Organizer. He served as a Reviewer for international conferences and international journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence and *Image and Vision Computing*.

SETH NIXON received the Ph.D. degree in computer science from the University of East Anglia, U.K., with a focus on image analysis using classical methods. He is currently a Research Fellow (RTD-A) with the University of Sassari, Italy. After, he took a postdoctoral research position, he is involved on the Secure Passwordless Authentication for Digital Identities (SPADA) Project with the University of Sassari. His research interests include computer vision, machine learning, deep learning, classical image analysis, biometrics, and medical imaging.

MATTEO ANEDDA (Senior Member, IEEE) received the M.Sc. degree (summa cum laude) in telecommunication engineering and the Ph.D. degree in electronic and computer engineering from the University of Cagliari, in 2012 and 2017, respectively. He has been a Research Fellow with the Department of Electrical and Electronic Engineering, University of Cagliari, since 2017. His research interests include real-time applications, 5G networks and network selection, the IoT and smart cities, adaptive multimedia streaming, and heterogeneous radio access environment. He is a Senior Member of IEEE Broadcast Technology, IEEE Communications, and IEEE Vehicular Technology societies.

ENRICO GROSSO received the degree in electronic engineering and the Ph.D. degree in electronic and computer engineering from the University of Genoa. He was a Researcher with the Faculty of Engineering, University of Genoa, from 1995 to 2002, then an Associate Professor with the Faculty of Economics, University of Sassari, in that period he directed numerous national and international research projects, mainly concerning automation, robotics, and artificial vision. Since January 2005, he has been a Full Professor of information processing systems. He was the Dean of the Faculty of Economics, from 2009 to 2012.

MARINELLA IOLE CADONI received the Ph.D. degree in mathematics from the University of Warwick, U.K., in 2004. From 2004 to 2008, she was a Postdoctoral Researcher with IEIIT-CNR, Turin, where she carried out research in computer vision. She then won a three year grant from Sardinian Region to finance a research project on biometrics with the University of Sassari. From 2011 to 2017, she carried out research activities both at the University of Sassari and for private companies. Since 2017, she has been a Lecturer of computer science with the University of Sassari. She is the coauthor of several articles published in international journals and acted as a reviewer for some of the major international conferences and journals on computer vision. She has been involved in several nationally funded research projects over the years. Her research interests include computer vision, in particular biometrics, 3D object reconstruction and retrieval, convolutional neural networks attention, and human attention.

Open Access funding provided by 'Università degli Studi di Cagliari' within the CRUI CARE Agreement