A Dataset of Gaze and Mouse Patterns in the Context of Facial Expression Recognition

Alexandre Bruckert, Lucie Lévêque, Matthieu Perreira Da Silva, Patrick Le Callet

A Dataset of Gaze and Mouse Patterns in the Context of Facial Expression Recognition

ALEXANDRE BRUCKERT, LUCIE LÉVÊQUE, MATTHIEU PERREIRA DA SILVA, and PATRICK LE CALLET, Nantes Université, École Centrale Nantes, LS2N, UMR 6004, France

Facial expression recognition is an important and challenging task for both the computer vision and affective computing communities, and even more specifically in the context of multimedia applications, where audience understanding is of particular interest. Recent data-oriented approaches have created the need for large-scale annotated datasets. However, most existing datasets present some weaknesses, because of the collecting methods used. In order to further highlight these issues, we investigate in this work how human visual attention is deployed when performing a facial expression recognition task. To do so, we carried out several complementary experiments, using the eye-tracking technology, as well as the BubbleView metaphor, both under laboratory and crowdsourcing settings. We show significant variations in gaze patterns depending on the emotion represented, but also on the difficulty of the task, i.e., whether the emotion is correctly recognised or not. Moreover, we use these results to propose recommendations on the ways to collect label data for facial expression recognition datasets.

CCS Concepts: • **Human-centered computing** → Human computer interaction (HCI); • **Applied computing** → Computers in other domains.

Additional Key Words and Phrases: emotions, facial expression recognition (FER), eye-tracking, BubbleView, dataset.

ACM Reference Format:
Alexandre Bruckert, Lucie Lévêque, Matthieu Perreira Da Silva, and, Patrick Le Callet. 2023. A Dataset of Gaze and Mouse Patterns in the Context of Facial Expression Recognition. In . ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As emotions play a crucial role in human communication, decisions, and perception, they have been under investigation in many different fields, including psychology, sociology, and neuroscience. Thanks to technological advances, new domains have arisen, such as affective computing [1]. The latter corresponds to systems able to recognise, analyse, model, and express human emotions.

In the framework of interactive media experiences, emotions can be of particular interest in a broad range of ways, and at many levels. Emotions and users engagement are indeed strongly linked, as audience engagement can be defined as the "cognitive, emotional, or affective experiences users have with media content" [2]. Several research works have been conducted with a view to study such relationship, e.g., to increase audience engagement in live interactive performance [3], or even to investigate the role of emotions in live streaming [4].

Automatic emotion recognition refers to the way a machine is able to identify the emotional states of human beings on the basis of various types of physiological or non-physiological signals. Amidst non-physiological measures, facial expressions are probably the most studied modality of non-verbal expression of emotions [5]. In 1969, Ekman identified six so-called “basic emotions", including happiness, anger, disgust, fear, sadness, and surprise – on top of these can be

Manuscript submitted to ACM

added the neutral emotion [6]. Automatic facial expression recognition (FER) therefore corresponds to classification algorithms used to put a name on each emotion by associating known characteristics on human faces. Such algorithms, made possible with advances in computer vision technologies, are trained on large amounts of data.

Facial expression recognition appears amongst the most demanding tasks in social communication [7]. As with many other tasks in machine learning, one of the main challenges remains the quality of datasets. Zeng *et al.* found annotations in public datasets (e.g., AffectNet [8]) to be inconsistent [9]. Similarly, Lévêque *et al.* showed in a previous study that the labels of widely used FER datasets (e.g., FER-2013 [10]) can be put into question, as they are mostly defined based on image search or by a single annotator [11]. Emotions can be perceived differently from one individual to another; annotating facial expressions can therefore be extremely tedious.

The study of human visual attention when performing a FER task can bring some insights on the analysis of such inconsistency. It has indeed been widely shown in the literature that both eye movements and positions can provide useful information on human perception and cognition [12]-[13]. The eye-tracking technology, i.e., the process of measuring where people look in a visual field, is commonly used to scrutinise how humans process visual information.

Due to the COVID-19 pandemic (amongst other factors), several metaphors have been proposed with a view to collect visual attention data without using eye-tracking devices. For instance, the BubbleView metaphor consists in blurring the image, except for the area around the mouse cursor. This metaphor has shown very good correlation with eye-tracking data in the literature [14]-[15]. Consequently, it allows for the collection of large visual saliency datasets, like the Saliency in Context (SALICON) dataset [16].

Previous studies have shown that human visual attention tends to prioritise emotional content (e.g., cute kittens) over non-emotional (i.e., neutral) stimuli [17]-[18]. Yet, to the best of our knowledge, little research has analysed the relationships between facial expression recognition (FER) tasks and visual attention.

In this article, we present three experiments which were conducted under different settings, i.e., eye-tracking in lab, BubbleView in lab, and BubbleView in crowdsourcing, in order to better understand where humans look when asked to perform a FER task. More specifically, our contributions are the following:

- We propose a new dataset dedicated to study visual attention in the context of facial expression recognition tasks. More specifically, this dataset includes both eye-tracking and mouse-tracking patterns. This dataset is made publicly available at *[link removed for anonymization]*.
- We highlight links between the way we visually explore human faces, and the emotions displayed.
- We discuss ways of leveraging visual attention data in order to improve the collection of large-scale facial expression recognition datasets, and more specifically their reliability.

2 MATERIAL AND METHODS

In this work, we conducted an eye-tracking experiment, as well as two BubbleView experiments, whose characteristics are summarised in Table 1 and further developed in the following subsections.

2.1 Stimuli

The source images used in our experiments consist of a total of 200 images of faces randomly taken in the AffectNet dataset [8]. AffectNet contains about one million of in-the-wild facial images collected using search engines. Half of the images were manually annotated (each image by a single annotator) using the seven discrete facial expressions model, while the other half were automatically annotated.

Table 1. Characteristics of the eye-tracking and BubbleView experiments.

Features	Eye-tracking	BubbleView in lab	BubbleView online
Number of images	120	80 (2 playlists of 40)	160 (4 playlists of 40)
Screen resolution	1920×1080	1920×1080	Depending on participants
Display time (sec)	4	7	7
Number of observers	50	60	240
Eye-tracker	Tobii Pro Fusion	/	/
Bubble size (px)	/	94	94
Blur sigma (px)	/	12	12

More specifically, we selected fifty images labelled as "happy", fifty as "sad", fifty as "surprised", and fifty as "angry". These four emotions were selected out of the seven basic ones as it was shown in the literature that they were less subject to recognition errors [11]. Amongst the fifty images of each emotion, half of them were tagged as "manually annotated" in the AffectNet dataset, whereas the other half were automatically annotated, using a ResNeXt neural network. We resized all images to 720×720 pixels, as their original resolution was different from each other.

2.2 Eye-tracking experiment

For the eye-tracking experiment, participants were presented 120 different stimuli (from the 200 previously selected) in a random order. he eye-tracking subset was still balanced in terms of emotions, i.e., it was composed of thirty images per emotion. Each stimulus was displayed for four seconds on a full HD 1920×1080 monitor screen with a grey background (as they were resized to 720×720), and followed by a one-second grey screen. After viewing a stimulus, the participants had to answer the following question: "Which facial expression was represented on the image?", by choosing one of the following four options on the screen: happiness, sadness, surprise, and anger. Once their choice was made and validated, they had to fix a target at the centre of the screen to move to the next image.

The experiment was conducted in a room with a low surface reflectance and constant ambient light. The screen luminance was set at 200 cd/m², and the luminance for the room's walls was measured at 30 cd/m². The viewing distance was maintained around 70 cm. The eye movements of the participants were recorded using a Tobii Pro Fusion eye-tracker, at a sampling rate of 120 Hz. At the beginning of a session, i.e., for each participant, the eye-tracking system was calibrated using a 9-point calibration protocol, ensuring a precision and accuracy under 0.5 degrees of visual angle. Before the start of the experiment, participants were given written instructions about the procedure, and a training session was provided, with a view to allow participants to familiarise themselves with the stimuli and the question asked. The four stimuli used in the training session were different from those used during the real experiment.

A total of fifty participants were involved in the eye-tracking experiment. Data from one participant were removed due to external issues during the procedure. Among the remaining participants were 31 females and 18 males, aged between 20 and 65 (M: 34.9, STD: 14.6). All participants had normal (or corrected-to-normal) vision. Note this experiment was approved by our local ethics committee.

2.3 BubbleView experiments

As introduced previously, the BubbleView metaphor is a "mouse-contingent, moving-window interface in which participants are presented with a series of blurred images and click to reveal bubbles" [14]. The Bubbleview methodology "replaces" eye-tracking with mouse clicks. We conducted two BubbleView experiments, one in laboratory conditions,

and the other one on a crowdsourcing platform. Such platforms allow conducting large-scale experiments with reduced costs and efforts [19]. A total of 160 images, selected from the original 200 stimuli, were used for the BubbleView experiments. Amongst these, eighty are common to the images used for the eye-tracking experiment. These 160 images were carefully divided into four playlists of forty images, balanced in terms of emotions. Two playlists (i.e., eighty images) were randomly presented to in-lab participants, while only one playlist (i.e., forty images) was shown to the online participants, as crowdsourcing sessions need to last less than fifteen minutes. Participants were asked to answer the same question, i.e., to recognise the facial expression represented.

In order to set up the BubbleView metaphor, several parameters need to be defined, i.e., the size of the bubble, the blur sigma, and the display time of the stimulus. The size of the bubble shall approximate the size of the fovea [14]. Therefore, it was set to 1.5 degrees of visual angle, i.e., 94 pixels. The blur sigma needs to be derived to approximate the visual acuity, i.e., 12 pixels in our case [20]. As for the viewing time, it was set at seven seconds, as some research works showed that a longer presentation time was need for BubbleView compared to eye-tracking experiments [15].

Sixty subjects participated in the laboratory experiment, including 34 females and 26 males. They all had normal, or corrected-to-normal vision. As for the online experiment, a total of 240 participants (note one was further discarded) were recruited using the Prolific platform [21]. Contrary to other platforms, Prolific takes into account researchers' needs by maintaining a recruitment process which is close to that of a laboratory experiment [22]-[23]. Like for the eye-tracking experiment, both BubbleView experiments were approved by the local ethics committee.

3 EYE-TRACKING EXPERIMENT: RESULTS

With this first experiment, using the eye-tracking technology, we aimed to study the way people look at human faces when trying to understand their emotional state.

3.1 Emotion recognition labels

After viewing each image, participants were asked to label it depending on the emotion they recognised (i.e., happiness, sadness, anger, or surprise).

As defining which emotion is displayed on a face can be a difficult task [11], we first evaluated the amount of agreement amongst observers watching the same image. The amount of agreement P, based on Fleiss' kappa, was computed as such:

$$P = \frac{1}{N(N-1)} \sum_{i=1}^{4} n_i(n_i - 1)$$

where N is the total number of observers (i.e., $N = 49$), and n_i the number of observers who assigned the emotion i among the four possible choices to a given image.

Figure 1 (a) represents examples of faces ordered by their agreement scores. Over the 120 images, only 29 of them (24.2%) had an agreement value over 0.9, indicating that observers strongly agreed on the label, while 31 images (25.8%) presented an agreement value under 0.5, indicating that the emotion displayed on these images was not easily recognised by the participants.

We also compared the annotations we gathered with the original labels of the AffectNet dataset. Figure 1 (b) illustrates a few examples of images falling in this configuration. The emotion selected by the highest amount of observers did not match the original labels in 18 cases (15%), which is coherent with other studies on the reliability of facial expression recognition ground truth labels [11]. The mean degree of agreement over these 18 images is equal to 0.48, showing a

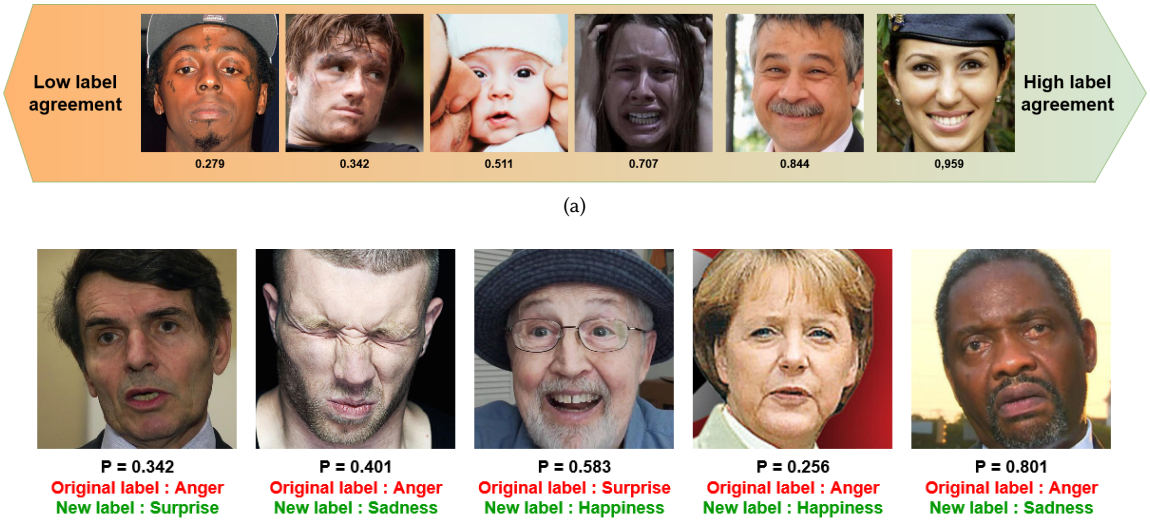


Fig. 1. Illustration of a sample of images based on the degree of agreement between participants in terms of labels. More particularly: (a) examples ordered by agreement score, (b) examples of disagreement between the majority vote and original labels.

medium agreement between participants. It is also interesting to note that, amongst these 18 images, 7 were originally annotated manually by a single observer, while 11 were annotated by an automatic model [8].

3.2 Eye-tracking data processing

During the experiment, gaze samples were recorded at 120 Hz. First, we discarded all gaze samples where the eye was not detected, due to tracking errors or blinks. We applied a first filter to eliminate data from an observer on an image if more than 10% of gaze samples had been discarded. This way, we eliminated 14 scanpaths, over 12 images.

Eye fixations were then extracted using a thresholding algorithm, relying on motion, velocity, and acceleration [24]. We also removed fixations lasting less than 80 ms, i.e., roughly the minimal time required to process foveal information [25]. Gaze points detected as part as the same fixation were aggregated in a single fixation point, located at the barycentre of all said gaze points.

We defined a saccade as the motion occurring between two successive fixations, provided that they last less than 150 ms, and that no tracking error or blink happened in between.

Visual saliency maps were obtained by aggregating the fixation points of all observers on a binary map, and applying a 2D gaussian kernel which standard deviation was set to represent one degree of visual angle, i.e. 94px, to account for both the size of the fovea and the precision of the eye-tracking device.

3.3 Fixation and saccade patterns

Over the 120 images and 49 observers, we gathered a total of 58 273 valid fixations, and 44 321 valid saccades. On average, a fixation lasted 315.5 ms (*STD*: 251.0 ms), and 9.55 fixations per stimulus per observer were recorded (*STD*:

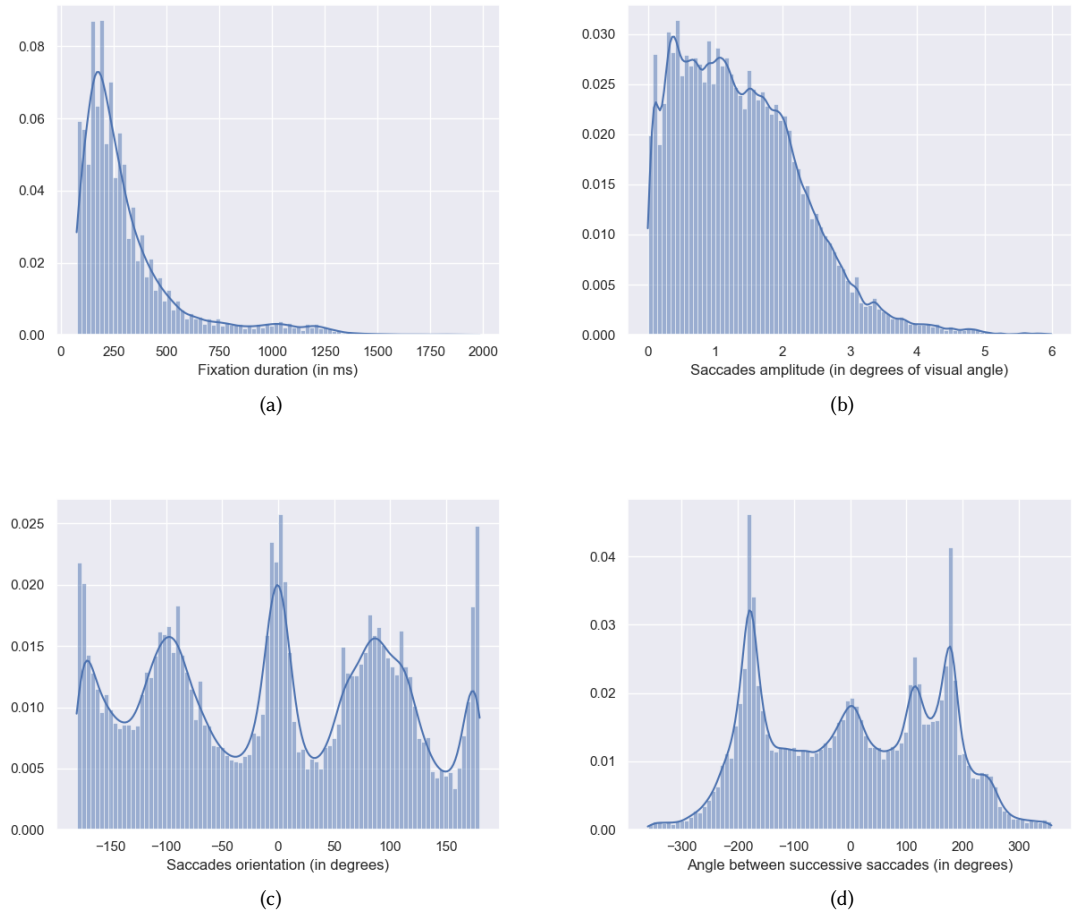


Fig. 2. Illustration of distributions of (a) fixation durations, (b) saccade amplitudes, (c) saccade orientations, and (d) angle between successive saccades over the 120 images of the eye-tracking dataset.

0.58). We found an inverse correlation ($\rho = -0.48$) between the number of fixations recorded and the agreement score P, indicating that participants tended to explore a stimulus more when the emotion was not easy to recognise.

Figure 2 shows the distributions of fixation duration (a), saccade amplitudes (b), saccade orientations (c), and angle between two successive saccades (d). The average saccade amplitude is 1.38 degree of visual angle, which is quite low compared to what can be usually observed in natural scenes exploration [26], but can be explained by the nature of the stimuli. Indeed, the distance between the features of the presented faces (eyes, mouth, nose, and eyebrows) is roughly in the order of two degrees of visual angle. We can also observe a relatively high amount of vertical saccades (Figure 2 (c)), indicating a high number of saccades going from mouth to eyes, and vice-versa).

Fig. 3. Illustration of examples of face semantic segmentation.

Table 2. Amount (in percent) of gaze fixations located on the area of each facial feature, for each emotion.

	Left eye	Right eye	Left eyebrow	Right eyebrow	Nose	Mouth
Happiness	21.43	21.08	2.71	2.62	7.55	19.78
Anger	18.26	18.17	6.80	6.76	5.23	19.07
Sadness	20.59	20.35	3.26	3.38	6.17	20.70
Surprise	23.87	24.01	4.97	4.93	4.28	26.32

3.4 Facial features and areas of interest

In order to distinguish specific gaze patterns relatively to the displayed emotion, we first segmented each image into areas of interest based on six facial features, i.e., left and right eyes, left and right eyebrows, nose, and mouth. To do so, we used the BiSeNet semantic segmentation model [27], fine-tuned on the CelebAMask-HQ dataset [28], using the implementation and weights in [29]. Figure 3 shows visual examples of such segmented faces.

We then counted the amount of gaze fixations falling into each region of interest. As shown in Table 2, we can observe significant differences in the distribution of fixations amongst the different facial features areas. For instance, the amount of fixations happening on the eyebrows is significantly higher in images displaying anger or surprise (11.73% on average for both left and right eyebrows) compared with faces displaying happiness (5.33%). Similarly, the amount of fixations located on the mouth are higher on surprised faces (26.32%) than on any other faces. These results are coherent with the general idea we have on facial expressions (e.g., surprise is usually linked to an open mouth).

4 BUBBLEVIEW EXPERIMENT: RESULTS

With this experiment, we aimed to evaluate whether or not the BubbleView metaphor could be reliably used to collect data that would be analogous to eye-tracking scanpaths in the context of facial expression recognition. To this aim, 160 images were used, in two different settings, i.e., in lab and through crowdsourcing.

4.1 Emotion recognition labels

Similarly to the eye-tracking experiment, participants were asked to label the image they were shown, with the same four emotion categories.

In laboratory settings, the average P agreement was 0.710 (*STD*: 0.235). Over the 160 images, 49 (30.62%) had an agreement score over 0.9, and 45 (28.12%) had an agreement under 0.5, which remains consistent with the scores obtained during the eye-tracking experiment.

We observed very similar results in crowdsourcing settings, where the average P agreement was 0.671 (*STD*: 0.222), 33 images (20.6%) had an agreement score over 0.9, and 46 (28.75%) had an agreement under 0.5. We believe that the lower amount of very high agreement images is just due to the higher number of annotators, and the fact that they might have different cultural backgrounds, meaning that they could interpret facial expression slightly differently.

Only 4 images (2.5%) were labeled differently by the majority of participants between laboratory and crowdsourcing settings. All of those images had an agreement score under 0.3 (both in-lab and in crowdsourcing), indicating that these cases were particularly difficult to classify, most likely because they did not fit within one of the four proposed categories (i.e., happiness, sadness, surprise, and anger). We also observe the exact same labeling by the majority of in-lab and crowdsourcing participants compared with eye-tracking experiment participants on the 80 common stimuli, showing that the collected labels are independent of the conditions in which they were obtained.

4.2 BubbleView data processing

For this study, we chose to rely on the continuous BubbleView metaphor, i.e., using solely continuous mouse tracking and not relying on clicks, in order to account for both bottom-up and top-down visual processes. In this way, the whole image was blurred, except for a circular area located around the position of the mouse cursor on screen. For each image and each observer, we removed the mouse tracks from the study when less than 50 points were recorded, indicating that the mouse was not moved during extended periods of time.

To infer visual saliency maps from mouse tracking data, we relied on the work presented in [15]. The value of each pixel of the map was set to be the total amount of time spent with the mouse cursor at this location, aggregated over all observers. This map was then convolved with the same 2D gaussian kernel used in the eye-tracking experiment, representing one degree of visual angle.

With a view to validate the soundness of the BubbleView metaphor in this context, we compared the saliency maps obtained with this method to ground-truth eye-tracking data, using three metrics commonly used to evaluate visual saliency maps, i.e., Pearson's correlation coefficient (CC), Normalized Scanpath Saliency (NSS), and Borji's Area Under Curve (AUC-B) [30]. Over the 80 images that are common to the BubbleView and eye-tracking experiments, we can observe a CC value of 0.84, a NSS value of 2.43, and an AUC-B value of 0.87, indicating a very high similarity between both ways of collecting data. We also observe no significant difference between the maps collected in laboratory and in crowdsourcing settings (CC = 0.98). Consequently, we merge these two sources of data in the following.

Figure 4 however highlights one of the main differences between eye-tracking and BubbleView saliency maps: eye-tracking maps seem to focus much more on the areas between the eyes, as well as on the nose, which might be indicative of intermediate fixations happening during the visual path from one eye to the other, or from the eyes to the mouth.

Fig. 4. Illustration of examples of eye-tracking (middle line) and BubbleView (bottom line) saliency maps.

4.3 Mouse track patterns

During the BubbleView experiments, the position of the mouse was recorded every time the mouse was moving, with a maximum sampling rate of 250 Hz (i.e., corresponding to one point every 4ms). On average, one point was recorded every 20.64ms, and we gathered a total of 3 834 077 points over the 160 images.

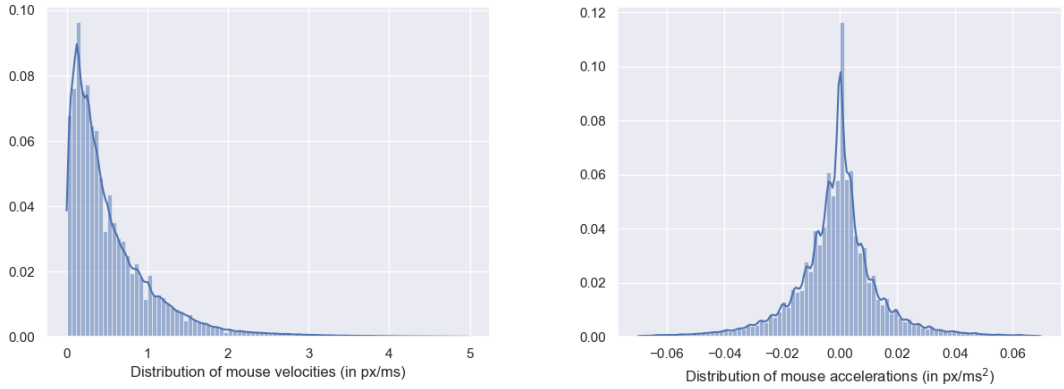
Figure 5 shows the distribution of mouse velocities and accelerations. We observe that, overall, mouse velocity did not vary much, indicating that participants explored the stimuli in a relatively smooth and continuous way, which is also supported by low values of mouse acceleration.

Similarly to the eye-tracking patterns, we fond an inverse correlation ($\rho = -0.45$) between the total distance of the mouse path on a stimulus and the agreement score, meaning that an ambiguous stimulus caused observers to explore a face more thoroughly in order to discriminate between the four possible labels.

4.4 Facial features and areas of interest

As for the eye-tracking experiment, we used the semantic segmentation of the stimuli to evaluate the importance of facial features in the classification task. In this case, instead of counting the number of recorded points falling in the different regions, we rather counted the amount of time during which the mouse was located on each facial feature area. Similarly to the saliency maps created from mouse-tracking, this was done to avoid an over-representation of transition points relatively to "fixations" where the mouse could stay still at a given location, and thus creating only a single point.

Table 3 shows the distribution of time spent on each of the considered facial feature. It should be noted that, as illustrated in Figure 4, most of the time was spent on these given features rather than on the rest of the face: 95.4% on average, compared to 77.6% of eye fixation points falling in these areas.



(a) (b)

Fig. 5. Illustration of the distributions of (a) mouse velocities and (b) mouse accelerations over the 160 images in the BubbleView dataset.

Table 3. Amount (in percent) of time spend by the mouse cursor on the area of each facial feature, for each emotion.

	Left eye	Right eye	Left eyebrow	Right eyebrow	Nose	Mouth
Happiness	28.21	27.03	3.64	3.16	9.83	22.31
Anger	20.98	21.09	9.71	9.90	7.74	24.50
Sadness	25.27	24.88	4.13	5.82	6.26	25.87
Surprise	27.07	27.18	5.24	5.10	5.27	31.38

We observe a distribution of attention between the feature areas that is quite similar to the one obtained in the eye-tracking experiments: "Happiness" stimuli show a high proportion of fixations on the eyes, "Anger" stimuli draw more attention on the eyebrows, and "Surprise" stimuli focus more on mouths and eyes. This supports the hypothesis that different emotions are linked to different distinguishable attention patterns, highlighting the interest of such a multi-modal dataset.

5 DISCUSSION

With this new dataset, we have proposed to consider visual attention and facial expression recognition (FER) as a whole, showing links between attention patterns obtained through eye-tracking or the BubbleView metaphor, and the annotation task of assigning the right emotion label to pictures of faces. In the following, we discuss several ways to make use of the existence of such relationship, and propose perspectives for future works.

5.1 Reliability of FER annotations

As discussed in several previous works, e.g., [9, 11], existing FER datasets often present inconsistent, or unreliable labels. For instance, over the 200 images that we used for this work, 32 (16%) presented discrepancies between the original label of the AffectNet set and the most-voted label in our data collection.

This can be explained by various factors, related to the way of collecting such data, but also to the intrinsic nature of the very definition of ground-truth in the context of facial expressions. It can indeed be noted that the AffectNet dataset was annotated with a single human annotator per image on half of the set, and a combination of predictions of deep neural networks for the remaining half. This way of collecting data, while enabling the large scale of the dataset, allows for numerous cases of misclassification.

As highlighted previously, we showed the existence of a relationship between human visual attention, through gaze and mouse patterns, and the emotion recognised, as well as with the difficulty of the FER task - that is to say on whether an emotion was easily recognised or not. Consequently, such patterns could be used as a mean to evaluate the confidence level of human annotators when associating an emotion to a given portrait.

In practice, the annotation of images of faces would include, on top of the FER task, recording the gaze patterns of annotators using the eye-tracking technology, or their mouse patterns using the BubbleView metaphor. This way, FER datasets could be more reliable and would include objective levels of confidence for each image.

5.2 Human and machine attention in FER tasks

Recently, the computer vision and machine learning fields have drawn inspiration on the human mechanism of visual attention, i.e., the selection of the most task-relevant areas of an image before further cognitive process. By implementing or learning such mechanisms, some models offer an additional layer of explicability, by providing saliency maps indicating the areas of interest the most responsible for the final decision [31, 32].

Such saliency maps can then be compared to human visual attention ground-truth when performing similar tasks. Previous works showed that, for high-level tasks, the closeness between human and machine attention leads to better performances of the model [31]. We would argue that facial expression recognition is an excellent example of such high-level vision tasks. Therefore, our dataset could be used to improve FER models relying on attention mechanisms, e.g., [33, 34], for instance by adding a regularisation term to the loss function to penalise the generated attention map should it be too far away from human ground truth.

Moreover, considering the relationship between attention patterns and facial expressions, we believe that deep learning models could rely on a common set of extracted features for both attention prediction and FER tasks. Consequently, a multi-task learning framework, where a model would learn to predict the displayed emotion as well as visual saliency or scanpaths using the same feature extractor, seems to be an interesting approach, for which our dataset - or extensions of it - would prove particularly useful.

REFERENCES

[1] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[2] Marcel Broersma. Audience engagement. *The international encyclopedia of journalism studies*, pages 1–6, 2019.

[3] Rossana Damiano, Vincenzo Lombardo, Giulia Monticone, and Antonio Pizzo. Studying and designing emotions in live interactions with the audience. *Multimedia Tools and Applications*, 80:6711–6736, 2021.

[4] Yan Lin, Dai Yao, and Xingyu Chen. Happiness begets money: Emotion and engagement in live streaming. *Journal of Marketing Research*, 58(3): 417–438, 2021.

[5] Albert Mehrabian. Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation*, 1(6):203–207, 1968.

[6] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.

[7] I Michael Revina and WR Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, 33(6):619–628, 2021.

[8] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. doi: 10.1109/TAFFC.2017.2740923.

[9] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[10] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, and Aaron Courville. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing*, pages 117–124, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-42051-1.

[11] Lucie Lévêque, François Villoteau, Emmanuel VB Sampaio, Matthieu Perreira Da Silva, and Patrick Le Callet. Comparing the robustness of humans and deep neural networks on facial expression recognition. *Electronics*, 11(23):4030, 2022.

[12] Jeremy M Wolfe. Visual attention. *Seeing*, pages 335–386, 2000.

[13] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694, 2007.

[14] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–40, 2017.

[15] Waqas Ellahi, Toinon Vigier, and Patrick Le Callet. Evaluation of the bubble view metaphor for the crowdsourcing study of visual attention deployment in tone-mapped images. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, 2021. doi: 10.1109/EUVIP50544.2021.9483985.

[16] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[17] Tobias Brosch, Gilles Pourtois, and David Sander. *The perception and categorisation of emotional stimuli: A review*. Psychology Press, 2010.

[18] Jessica Gall Myrick. Emotion regulation, procrastination, and watching cat videos online: Who watches internet cats, why, and to what effect? *Computers in human behavior*, 52:168–176, 2015.

[19] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[20] GERALD Westheimer. Visual acuity and hyperacuity: resolution, localization, form. *American journal of optometry and physiological optics*, 64(8): 567–574, 1987.

[21] Stefan Palan and Christian Schitter. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[22] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[23] Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, and Frédéric Dufaux. Rv-tmo: Large-scale dataset for subjective quality assessment of tone mapped images. *IEEE Transactions on Multimedia*, 2022.

[24] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *ACM ETRA*, page 71–78, 2000. doi: 10.1145/355017.355028.

[25] Barry R. Manor and Evian Gordon. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods*, 128(1):85–93, 2003.

[26] Roman von Wartburg, Pascal Wurtz, Tobias Pflugshaupt, Thomas Nyffeler, Mathias Lüthi, and René M Müri. Size matters: Saccades during scene perception. *Perception*, 36(3):355–365, 2007. doi: 10.1068/p5552.

[27] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Computer Vision – ECCV 2018*, pages 334–349, Cham, 2018.

[28] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Zllrunning. face-parsing.pytorch, 2019. URL https://github.com/zllrunning/face-parsing.PyTorch.

[30] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[31] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2021. doi: 10.1109/TMM.2020.3007321.

[32] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad S Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *arXiv preprint arXiv:2204.07756*, 2022.

[33] Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, and Zhaojie Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.06.014.

[34] Xiaoliang Zhu, Zili He, Liang Zhao, Zhicheng Dai, and Qiaolai Yang. A cascade attention based facial expression recognition network by fusing multi-scale spatio-temporal features. *Sensors*, 22(4), 2022.

Received February 15, 2023; revised March 20, 2023; accepted March 31, 2023