Gaze Reveals Emotion Perception: Insights from Modelling Naturalistic Face Viewing

1st Meisam Jamshidi Seikavandi
IT University of Copenhagen
meis@itu.dk

2nd Maria Jung Barrett
IT University of Copenhagen
mbarrett@itu.dk

***Abstract*—Face Emotion Recognition (FER) is a fundamental human capability essential in social interactions and comprehension of others' mental states. Eye tracking emerges as an insightful tool to probe FER, shedding light on underlying cognitive processes. In this research, we adopted an instructionless paradigm, gathering eye movement data from 21 participants to probe two distinct FER processes: free viewing and grounded FER.**

During free viewing, participants observed faces without specific guidelines, revealing spontaneous attention allocation patterns. Grounded FER tasks, in contrast, had participants engage in emotion perception tasks driven by emotion-related words, enabling us to assess their performance and the influence of the grounding context. Importantly, we identified a predictive relationship between the success rate in grounded FER tasks and eye movement behavior during free viewing. Initial gaze patterns offered crucial cues for subsequent emotion perception processes. Moreover, we constructed machine learning models that accurately predicted gaze distribution based solely on the visual content of the stimulus in the FER task.

To boost scalability and comparability, we utilized features extracted from pre-trained deep-learning models for face recognition to model attention distribution during free viewing. This strategy facilitates the analysis of large-scale datasets and enables comparisons of emotion perception across various populations and settings. Our study enhances understanding of the complex relationship between eye movements and emotion perception, pushing the frontiers of FER research. The implications encompass psychology, human-computer interaction, and affective computing, with potential applications in developing precise emotion recognition systems.

***Index Terms*—Gaze, Emotion Perception, Face Emotion Recognition, Eye Tracking, Machine learning, Non-verbal Communication**

I. Introduction

Face Emotion Recognition (FER) plays a crucial role in human social interactions and non-verbal communication, as it involves interpreting emotions from facial expressions [1], [2]. Various methodologies, including brain imaging and physiological signals, have been employed to delve into this complex process [3]–[8].

Eye tracking, a non-invasive technology, offers profound insights into visual attention and emotional processing [9]. While less prevalent than brain imaging, it's often combined with other techniques to collect comprehensive data. Eye tracking imposes less burden on participants while providing abundant information.

Previous studies have exploited eye movements to understand emotion perception (EP) in adults [10], [11]. FER tasks have been instrumental in highlighting atypical EP linked to conditions like autism, ADHD, schizophrenia, and certain types of dementia [12]–[18].

Eye-tracking tasks used for diagnostic purposes can require high cognitive functioning. For instance, the antisaccade task is used to identify diseases like dementia and Alzheimer's [19]–[23].

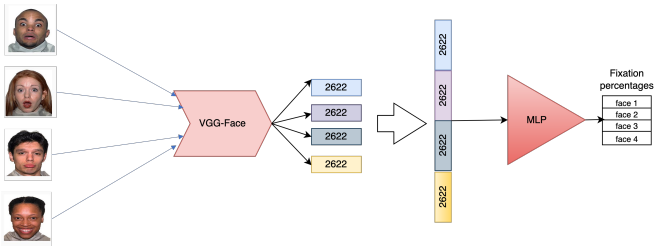

Fig. 1. The network structure designed for predicting fixations in Task 2

Tasks emulating real-life situations, known as naturalistic tasks, are gaining popularity due to their relaxed environment. These tasks prove beneficial for lightweight EP assessments and for identifying conditions like Alzheimer's [10], [19], [24].

The applications of eye-tracking research extend to clinical diagnosis and human-robot interaction (HRI) domains. Understanding gaze behavior in FER tasks can significantly augment the design of socially intelligent robots [25].

Our study builds upon the work of Russell et al. [18], exploring two FER processes and predicting FER success based on gaze features during face viewing. This contributes to the efficiency of EP assessments.

Fig. 2. The FER task unfolds in three steps: overlaying areas of interest, presenting faces with emotions, and randomizing locations to study gaze patterns.

II. Instructionless FER Task and Modifications

This section details the instructionless FER task developed by Russell et al. [18], along with our modifications to understand the FER process better.

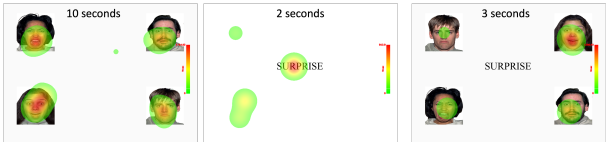

Fig. 3. The heatmaps display FER trial with no instructions. Different emotions have distinct fixation distributions.

TABLE I
PARTICIPANT CHARACTERISTICS AND THE R(READING) M(INDS IN THE) E(YES) T(EST) SCORE.

	AGE	RMET SCORE	DRIFT ERROR
COUNT	20	15	20
RANGE	23–44	17–32	0.01–1.21
MEAN	29.3	29.4	0.41
SD	5.3	3.6	0.28

A. Instructionless FER Task

Russell et al.'s instructionless FER task was designed to detect early-stage frontotemporal dementia. The task comprised three steps: showing four faces displaying distinct emotions for 10 seconds; presenting an emotion word for 2 seconds, then displaying both for 5 seconds. In this experiment, the positions of the faces remained constant, simulating a retrieval-like task. This setup facilitated the analysis of the free-viewing and retrieval phases while recognizing working memory differences as a potential confounding factor.

B. Modifications to the Instructionless FER Task

To gain a deeper understanding of the FER process and counteract the possible confounding effect of working memory, we revised Russell et al.'s task. In our version, the positions of the faces were randomized in Step 3. This modification required participants to recognize emotions rather than recall positions. In turn, this change separated the influence of memory from the task and allowed us to differentiate between the free-viewing (Step 1) and grounded FER (Step 3) phases.

By examining the gaze behavior and performance differences between these phases, we aimed to extract more quantitative data on FER cognitive processes. Our modifications simulate real-world FER scenarios, aiding our exploration of the relationship between eye movements, emotion perception, and attention allocation during FER tasks.

III. DATA COLLECTION

a) Participants: We collected data from twenty-one volunteers, of whom one was excluded due to incomplete data. The 20 remaining participants, five female, had educational backgrounds ranging from high school to Ph.D., most holding MSc degrees. Detailed demographics are presented in Table I. Participants provided informed consent in accordance with the protocol approved by our institution's Legal Department.

b) Apparatus and Stimuli : We used the Eyelink 1000 Plus eye-tracker to record eye movements during the FER tasks in a darkened room. The eye tracker was calibrated prior to the experiment and recalibrated as necessary to maintain data accuracy. We utilized the NimStim face emotion dataset [26] for 60 trials, ensuring balance in the facial images displaying different emotions and the associated emotion words. The trials were balanced for face diversity, similarity, and dissimilarity to the target emotion. Figure 2 depicts the positioning of the emotion and target faces.

TABLE II
DWELL TIME % PER STEP WRT. MAIN AREAS OF INTEREST (TARGET FACE, NON-TARGET FACE, AND WORD) AND D(WELL) T(IME) C(HANGE) ACROSS EMOTIONS.

	STEP 1	STEP 2			STEP 3			DTC
		TARGET		WORD	TARGET		WORD	
		NO	YES		NO	YES		
ANGRY	23.8	7.3	10.9	68.6	15.8	42.2	11.9	27.8
DISGUST	23.5	7.7	11.1	69.8	14.9	45.3	11.4	31.7
FEAR	25.3	7.5	8.6	69.0	17.8	34.9	10.8	16.0
HAPPY	21.4	6.7	11.0	68.2	14.4	43.8	12.1	34.2
SAD	23.0	6.2	11.4	65.8	14.8	38.5	11.3	25.1
SURPRISE	23.9	7.3	8.4	68.9	16.8	40.9	11.5	26.1
AVG.	23.5	7.1	10.2	68.4	15.7	40.9	11.5	26.8

c) Areas of Interest: We defined interest areas for each trial to include all four faces and the corresponding word. We also identified sub-areas within each face, specifically the eye, nose, and mouth regions, as shown in Figure 2.

d) Experiment Protocol: Participants completed six trials for task familiarization, followed by two rounds of 27 trials, with a short break in between. After the experiments, participants optionally completed the Reading the Mind in the Eyes test[1], the results of which are shown in Table I.

e) Preprocessing and Cleaning: Data from one eye were analyzed for consistency and reliability. The first six trials were excluded to eliminate the effect of initial familiarization. Fixation events were assigned to the nearest area of interest to facilitate data interpretation.

IV. STATISTICAL ANALYSIS

Following the recommendations of Skaramagkas et al.'s review [9], we adopted the dwell time percentage, or dwell time %, as our primary measure of visual attention. This metric represents the total focus duration on a specific area of interest (AOI) as a fraction of the total time spent on a given step. Additionally, we used the change in dwell time for target faces to measure emotion perception (EP) performance, a strategy proposed by Russell et al. [18].

$$dwell\ time\ change = dwell\ time\ \%\ step\ 3 - dwell\ time\ \%\ step\ 1$$

Table II supports the theory that participants, when operating without specific instructions, naturally pay more attention to the target face after the presentation of the emotion word (as indicated by a positive dwell time change score for the target). Performance ratios for a range of emotions align well with findings from previous FER studies, such as those conducted by Tottenham et al. [26], Russell et al. [18], and Polet et al. [27].

[1]https://s3.amazonaws.com/he-assets-prod/interactives/233_reading_the_mind_through_eyes/Launch.html

Table II also provides insight into the distribution of dwell time % across different emotions and for both target and non-target faces. In Step 1, we observed varied fixation distributions for different emotions. Fearful ($M = 25.3, SD = 11.9$) and surprised faces ($M = 23.9, SD = 12$) garnered more attention, while happy faces drew less focus ($M = 21.4, SD = 11.9$). Independent t-tests revealed significant differences associated with surprise ($t(1438) = 6.63, p < 0.0001$) and fear ($t(1458) = 4.06, p < 0.0001$). our null hypothesis (H0) stated that there is no significant difference in dwell time % between different emotions.

In Step 2, participants naturally sought to match the emotion word to the corresponding face. The emotion word and the position of the target face attracted the most attention, indicative of a memory effect. Specifically, the position of the target face received more focus, particularly for emotions like sadness, fear, and surprise, as depicted in Figure 3.

In Step 3, we noticed a new FER process where target faces ($M = 40.9, SD = 20.2$) received significantly more attention than non-target faces ($M = 15.7, SD = 11.8$) ($t(4318) = 64.2, p < 0.0001$). Non-target faces showing fear and surprise still attracted more attention than other non-target faces. Our results align well with previous studies that found participants tend to fixate longer on emotional faces, especially fearful and surprised ones, during daily communication.

V. Modeling

To mitigate bias from the initial learning phase of the task, we disregarded data from the first six trials. Consequently, our analysis used data from 20 participants from the remaining 54 trials. We obtained 54 sets of input and output data by averaging fixation events across all participants. Due to the dataset's limited size, we applied a leave-one-out cross-validation strategy at the trial level and reported the mean squared error (MSE) rates as our primary evaluation metric

A. Task 1

Task 1 endeavored to predict the dwell time for each face, both target and non-target, in Step 3, based on the averaged fixation events across participants. Given that Step 1 is more naturalistic compared to Step 3, accurate predictions of the fixation distribution in Step 3 and subsequent emotion perception performance based solely on Step 1 fixation events could facilitate a more authentic, instruction-free task. This development might stimulate modeling emotion perception during everyday interpersonal communication.

For prediction, we employed a set of features including spatial aspects like the percentage of dwell time and the number of fixations per face, along with temporal aspects like the duration of the first and last fixation and the start time. We also incorporated one-hot encoded features for emotions and whether the face is the target, resulting in a total of 15 features per face. These features, when concatenated for the four faces, were fed as input to a 3-layer Multilayer Perceptron (MLP) with 32-16-4 nodes to predict the dwell time percentage for each corresponding face. After 500 epochs and a learning rate of 0.001, the model achieved convergence. Separate models were trained on spatial, temporal, and spatiotemporal features. Though the dwell time for the target face is of primary interest, we also sought to understand the dwell time for non-target faces. To accommodate this, we assigned a higher weight to the prediction of the target face in the loss function.

TABLE III
Average MSE results for predicting the dwell time of Tasks 1 and 2

Task 1		
Features	All	Target
Baseline	0.0164	0.0060
Spatial	0.0134	0.0066
Temporal	0.0053	0.0030
Spatiotemporal	**0.0046**	**0.0024**
Task 2		
Features	Step 1	Step 3
Baseline	0.0152	0.0164
Face embedings	**0.0065**	**0.0077**

B. Task 2

Task 2 aimed to predict the fixation dwell time for each face in Steps 1 and 3, relying exclusively on their visual features. This methodology enabled us to generate an average fixation distribution for healthy individuals in a new trial and assess the trial's difficulty by calculating the fixation dwell change score. To this end, we utilized a pre-trained VGG-Face model [28] to extract face features. The embedding array, of size 2622, was obtained from the network's final feature layer, trained for face recognition. After concatenation, these arrays were input to a 3-layer MLP with 100-16-4 nodes to predict the dwell time, as depicted in Figure 1. With a learning rate of 0.001, the model achieved convergence after 1000 epochs. The network was trained separately for Steps 1 and 3, with the target face embeddings positioned first to distinguish between target and non-target faces in Step 3.

VI. Baseline Model

Our baseline model presumes that the target face garners the most attention. Hence, we designated the longest continual viewing time to the target face, equally dividing the remaining time among the other faces. This strategy produced optimal results, with a dwell time of 0.50 for the target face and 0.1666 for non-target faces in Step 3 of Task 1 and Step 2 of Task 2. In Step 1 of Task 2, where there is no specified target face, we allocated an equal dwell time to all four faces.

VII. Results

Table III presents the modeling results for both tasks, demonstrating accurate prediction of dwell times with low MSE rates. In Task 1, the best performance is achieved by utilizing both spatial and temporal feature sets, with the temporal features proving to be more effective than the spatial ones. This result highlights the potential of temporal gaze event distribution for modeling complex emotion perception tasks, even with eye information collected during natural, unrestricted emotion perception. Task 2 involves predicting fixation times

and modeling the identification of relevant emotion tasks in Step 3, which is, as expected, more challenging than free emotion perception in Step 1.

VIII. Discussion and Conclusion

We adapted Russell et al.'s (2021) [23] instructionless FER task for a deeper understanding of the FER process. The modifications allowed for extensive statistical analysis and revealed key differences in processing various emotions.

Our results suggest that gaze events, particularly temporal features, can predict FER performance by merely observing faces. We also predicted the fixation duration of FER tasks based on face visual features, aiding the assessment of trial difficulty. Uniquely, we predicted emotion perception accuracy from free face viewing, marking a step towards lightweight emotion recognition assessments not reliant on language skills.

Moreover, we introduced a standardized tool for FER datasets, enhancing result comparability. Overall, our work offers insights for FER research and could influence the development of more naturalistic emotion recognition assessments.

In conclusion, our work advances FER by exploring new paradigms and models. Predicting FER performance from free-viewing eye movements offers a path for efficient and ecologically valid emotion perception assessments. We hope our work will spur further research and foster improved tools and methodologies for studying human emotion perception.

Acknowledgements

Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN. Data Collection was supported by the DANGER project (Demens ANalyse i ALS Gennem Eye Tracking).

References

[1] A. S. Walker-Andrews, "Emotions and social development: Infants' recognition of emotions in others," *Pediatrics*, vol. 102, no. 5 Suppl E, pp. 1268–71, 1998.

[2] M. L. Smith, G. W. Cottrell, F. Gosselin, and P. G. Schyns, "Transmitting and decoding facial expressions," *Psychological science*, vol. 16, no. 3, pp. 184–189, 2005.

[3] L. Collin, J. Bindra, M. Raju, C. Gillberg, and H. Minnis, "Facial emotion recognition in child psychiatry: a systematic review," *Research in developmental disabilities*, vol. 34, no. 5, pp. 1505–1520, 2013.

[4] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2020.

[5] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics," *Scientific reports*, vol. 4, no. 1, pp. 1–13, 2014.

[6] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[7] M. L. Schroeter, S. Pawelke, S. Bisenius, J. Kynast, K. Schuemberg, M. Polyakova, S. Anderl-Straub, A. Danek, K. Fassbender, H. Jahn, *et al.*, "A modified reading the mind in the eyes test predicts behavioral variant frontotemporal dementia better than executive function tests," *Frontiers in aging neuroscience*, vol. 10, p. 11, 2018.

[8] B. Montagne, R. P. Kessels, E. H. De Haan, and D. I. Perrett, "The emotion recognition task: A paradigm to measure the perception of facial emotional expressions at different intensities," *Perceptual and motor skills*, vol. 104, no. 2, pp. 589–598, 2007.

[9] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. S. Tachos, E. Tripoliti, K. Marias, D. I. Fotiadis, and M. Tsiknakis, "Review of eye tracking metrics involved in emotional and cognitive processes," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 260–277, 2021.

[10] C. Aracena, S. Basterrech, V. Snáel, and J. Velásquez, "Neural networks for emotion recognition based on eye tracking data," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, (Hong Kong), pp. 2632–2637, IEEE, 2015.

[11] L. Chaby, I. Hupont, M. Avril, V. Luherne-du Boullay, and M. Chetouani, "Gaze behavior consistency among older and younger adults when looking at emotional faces," *Frontiers in Psychology*, vol. 8, p. 548, 2017.

[12] V. Tsang, "Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders," *Autism*, vol. 22, no. 2, pp. 161–170, 2018.

[13] J. N. Airdrie, K. Langley, A. Thapar, and S. H. van Goozen, "Facial emotion recognition and eye gaze in attention-deficit/hyperactivity disorder with and without comorbid conduct disorder," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 57, no. 8, pp. 561–570, 2018.

[14] V. J. Serrano, J. S. Owens, and B. Hallowell, "Where children with adhd direct visual attention during emotion knowledge tasks: Relationships to accuracy, response time, and adhd symptoms," *Journal of attention disorders*, vol. 22, no. 8, pp. 752–763, 2018.

[15] M. Asgharpour, M. Tehrani-Doost, M. Ahmadi, and H. Moshki, "Visual attention to emotional face in schizophrenia: an eye tracking study," *Iranian journal of psychiatry*, vol. 10, no. 1, p. 13, 2015.

[16] S. M. Shdo, C. L. Brown, J. Yuan, and R. W. Levenson, "Diminished visual attention to emotional faces is associated with poor emotional valence perception in frontotemporal dementia," *Dementia and Geriatric Cognitive Disorders*, vol. 51, no. 4, pp. 331–339, 2022.

[17] R. Hutchings, R. Palermo, J. Bruggemann, J. R. Hodges, O. Piguet, and F. Kumfor, "Looking but not seeing: Increased eye fixations in behavioural-variant frontotemporal dementia," *Cortex*, vol. 103, pp. 71–81, 2018.

[18] L. L. Russell, C. V. Greaves, R. S. Convery, J. Nicholas, J. D. Warren, D. Kaski, and J. D. Rohrer, "Novel instructionless eye tracking tasks identify emotion recognition deficits in frontotemporal dementia," *Alzheimer's research & therapy*, vol. 13, no. 1, pp. 1–11, 2021.

[19] M. R. Readman, M. Polden, M. C. Gibbs, L. Wareing, and T. J. Crawford, "The potential of naturalistic eye movement tasks in the diagnosis of alzheimer's disease: A review," *Brain sciences*, vol. 11, no. 11, p. 1503, 2021.

[20] C. Meyniel, S. Rivaud-Péchoux, P. Damier, and B. Gaymard, "Saccade impairments in patients with fronto-temporal dementia," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 11, pp. 1581–1584, 2005.

[21] J. Currie, B. Ramsden, C. McArthur, and P. Maruff, "Validation of a clinical antisaccadic eye movement test in the assessment of dementia," *Archives of neurology*, vol. 48, no. 6, pp. 644–648, 1991.

[22] R. J. Leigh, S. A. Newman, S. E. Folstein, A. G. Lasker, and B. A. Jensen, "Abnormal ocular motor control in huntington's disease," *Neurology*, vol. 33, no. 10, pp. 1268–1268, 1983.

[23] L. L. Russell, C. V. Greaves, R. S. Convery, M. Bocchetta, J. D. Warren, D. Kaski, and J. D. Rohrer, "Eye movements in frontotemporal dementia: Abnormalities of fixation, saccades and anti-saccades," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 7, no. 1, p. e12218, 2021.

[24] S. Primativo, C. Clark, K. X. Yong, N. C. Firth, J. Nicholas, D. Alexander, J. D. Warren, J. D. Rohrer, and S. J. Crutch, "Eyetracking metrics reveal impaired spatial anticipation in behavioural variant frontotemporal dementia," *Neuropsychologia*, vol. 106, pp. 328–340, 2017.

[25] C. Fu, Q. Deng, J. Shen, H. Mahzoon, and H. Ishiguro, "A preliminary study on realizing human–robot mental comforting dialogue via sharing experience emotionally," *Sensors*, vol. 22, no. 3, p. 991, 2022.

[26] N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, and C. Nelson, "The nimstim set of facial expressions: Judgments from untrained research participants," *Psychiatry research*, vol. 168, no. 3, pp. 242–249, 2009.

[27] K. Polet, S. Hesse, A. Morisot, B. Kullmann, S. L. de la Chapelle, A. Pesce, and G. Iakimova, "Eye-gaze strategies during facial emotion recognition in neurodegenerative diseases and links with neuropsychiatric disorders," *Cognitive and Behavioral Neurology*, vol. 35, no. 1, pp. 14–31, 2022.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)* (X. Xie, M. W. Jones, and G. K. L. Tam, eds.), (Swansea, United Kingdom), pp. 41.1–41.12, BMVA Press, September 2015.