Teaching a Robot to Read Faces: Incremental Emotion Learning with Selective Visual Attention

Raphaël d'Urso, Sofiane Boucenna, David Cohen, Alexandre Pitti

Teaching a Robot to Read Faces: Incremental Emotion Learning with Selective Visual Attention

Raphaël D'Urso [1], Sofiane Boucenna [1] , David Cohen [2 3], and Alexandre Pitti [1]

***Abstract*— In humans, changes in the visual field or obstructions influence the perception and interpretation of facial emotions. A partial or altered visual area can lead to a misunderstanding of the facial expressions of the interlocutor. This problem also arises in the field of social robotics, where the visual perception of robots is often constrained by technical and environmental factors. With the rise of robots intended for the general public, this is a question that must now be taken into account in many common situations.**

In this study, we explore the impact of restricting the visual field of a robot equipped with developmental AI on its ability to recognize human facial emotions. We analyze how different limitations of the field of vision affect its reactions and modify its ability to recognize facial emotions. For this, we developed

A architecture to explore this question, based on : (1) the recognition of primary emotions through an imitation game that enhances the robot's ability to recognize emotions, and (2) an attention-focusing mechanism. The integration of these two components allows the robot to modulate its areas of interest before performing facial emotion recognition. This approach enables the exploration of various recognition scenarios, ranging from full-face analysis to a more targeted focus on the mouth—an area particularly emphasized by elderly individuals and bilingual children.

Our results show that the restriction of the visual field leads to a decrease in the accuracy of emotion recognition, with variations depending on the type of expression and the degree of limitation of the visual field. These results are consistent with observations made in humans, suggesting the possibility of reproducing this mechanism in social robotics.
Correspondence : raphael.durso@cyu.fr

I. INTRODUCTION

Facial emotion recognition is a fundamental skill for social interactions. Our robots will be increasingly called upon to interact in everyday life situations. However, emotions have an influence on our decisions and actions [1]. Understanding and correctly interpreting the emotions expressed on faces facilitates interpersonal relationships and allows us to adjust our behavior according to the emotional reactions [2] [3]. But emotion recognition can be disrupted, particularly by an alteration of the visual field. Indeed, our perception of the emotions of others relies on two main areas: the eyes and the mouth. Directing the gaze towards these specific areas of the face could improve the accuracy of emotion recognition. Conversely, focusing only on one of these areas could reduce this accuracy [4].

Once we have learned to decode emotions, modifying the visual field, artificially or by a natural mechanism, modifies the ability to recognize the emotion on the face of others. Studies show that modifying the visual field leads to a decrease in the recognition of perceived emotions. This is the case with artificial modifications (wearing a mask or glasses) or by presenting images of faces with certain parts masked or invisible [4] [5] [6]. The modification of the visual field can also be linked to natural factors such as age or exposure to several languages from an early age in children [7] [8]. These studies show that attention is not constant on the face, but alternates between the eyes and the mouth. Various factors can therefore modify the visual field naturally and modify the recognition of emotions.

Most often, robotics uses deep learning models with CNNs such as ResNet, VGGNet Or other variations of CNNs [9] [10] [11], with RNN [12], Nested LSTM [13] or even Bayesian classifiers [14] for emotion recognition [15]. These networks are created to seek high performance and require complete databases as well as preprocessing. All these limitations are even stronger because we want to use the principle of developmental AI, which is inspired by humans for learning. Several studies utilize robotic heads for emotional facial recognition, similar to our approach. These models achieve diverse emotion recognition but primarily rely on visual information [16] [17] [18]. Interaction has also been shown to significantly enhance learning compared to simple camera or computer screen setups with non-interactive avatars [19] [20] [21]. Therefore, we chose a learning approach based on imitation games, which aligns with developmental robotics principles and diverges from deep learning models.

We build upon the Neural Model of Facial Expression Recognition (NMFER) proposed by [22], which adopts a developmental approach to facial emotion recognition. Additionally, we integrate the Multimodal Audio-Visual Attention (MAVA) model [23], which simulates attention allocation to the speaker's face during speech. Inspired by the mechanisms underlying a child's observation of an interlocutor, this model seeks to replicate the patterns of attention observed in natural interactions. A key question we aim to address is whether our model can predict the data observed in children [24], especially when the attention mechanism is modified as proposed in [25].

To develop our emotion recognition model with a visual field that can be modified with respect to the emotion recognition learning conditions, we will first train MAVA before integrating it into the emotion recognition model during the test part. This will simulate the capacity acquired in natural conditions by children. The visual field of the robot is then modified. We expect to find, with our model, results similar

[1]ETIS-UMR8051, CY Cergy Paris Universite, Cergy, France
[2]APHP, Pitié Salpétriere, GH Pitié-Salpêtrière, APHP, Paris, France
[3]UPMC, ISIR, Sorbonne Universite

to those of experimental psychology: a regression of emotion recognition, the more the visual field is disturbed, but a more or less strong regression depending on the emotion analyzed. In this article, we will describe the experimental setup, before presenting the MAVA Model. We will present the NFMER model, with the addition of MAVA allowing the modification of the field of attention, which distinguishes our approach from previous studies on these models. Finally, we will present our results.

II. Material and Methods

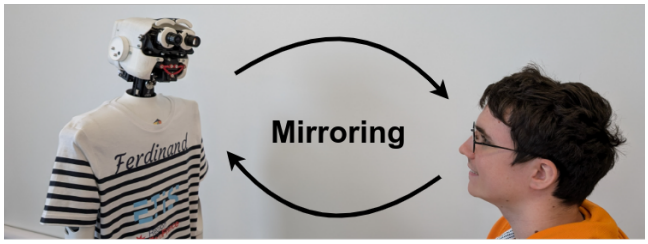


Fig. 1: Imitation game between human and robot.

A. Experimental setup

In our emotional interaction experiment, the participant positions himself in front of the robot Ferdinand, at an optimal distance allowing the robot's camera to clearly capture their face (Fig. 1). Thanks to its "expressive" head, Ferdinand can reproduce the primary facial expressions defined by Ekman, including "neutral", "happiness", "sadness","surprise" and "anger" [2]. These expressions are generated by adjusting the position of its motors (Fig. 2).

Ferdinand must acquire two cognitive skills. The first is visual attention, which enables Ferdinand to direct its gaze toward the most relevant areas of the human face. We use our neural model, MAVA (Multimodal Audio-Visual Attention), to identify the most active region [23].The Mouth is a key area for audio-visual vowel learning, facilitating lip reading. The mouth is also highly valuable for recognizing facial expressions. For this, we ask the participant to speak in front of the robot until our model converges. Before Ferdinand can begin learning emotions, it must first master the attention mechanism. To achieve this, it observes a speaker in front of it, and MAVA learns to focus on the most salient area (the mouth) through its Hebbian learning mechanism. Once learning is complete, MAVA perceives the mouth as the most active region, at which point the learning process is frozen for the remainder of the experiment. This final network state will then be used in the emotion recognition.

Then, the second cognitive skill that Ferdinand must learn is facial expression recognition. To achieve this, he engages in an imitation game with a human. During the learning phase, Humans imitate Ferdinand's facial expressions. This process allows him to associate his own facial actions (which he cannot see) with his visual perception (the facial expressions of his partner that he can observe). The main challenge

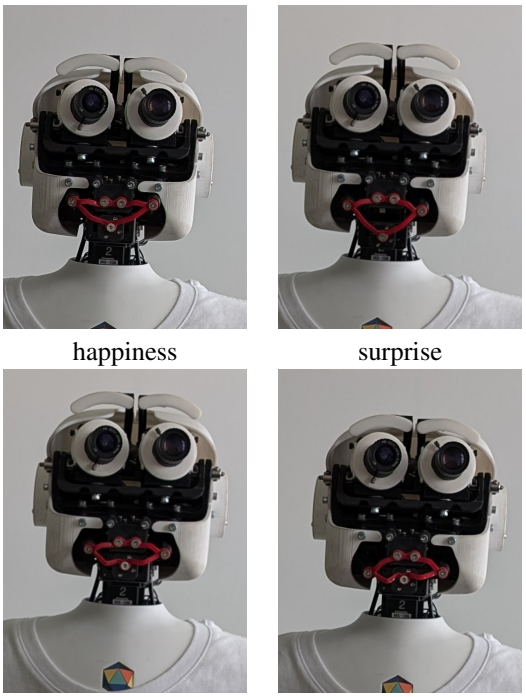


Fig. 2: Ferdinand's Expressive Capabilities.

is to enable Ferdinand to establish a connection between his own invisible expressions and those of the human, allowing him to recognize the emotions expressed by his interlocutor. Once this learning phase is completed, the roles are reversed : Ferdinand observes the human and attempts to recognize their expression while reproducing it. The robot is capable of producing five facial expressions and has interacted with 18 participants. Each person imitated each expression five times, resulting in a total of 250 recorded images (5 expressions × 5 repetitions × 18 participants). In this experiment, facial expression recognition may be influenced by the MAVA model, which restricts the robot's visual field to the participant's mouth. This limitation could impact the accuracy of emotion recognition, thereby reflecting perceptual biases related to visual attention.

B. Attention mechanism

MAVA is a bimodal attention mechanism (audio and visual) designed to focus attention on the speaker's mouth ([23] for more details). As illustrated in Fig. 3, its neural architecture relies on two streams of information : sound and image. A Hebbian learning rule allows for the association of image elements congruent with the sound, thereby facilitating the identification of the most relevant areas for audiovisual learning.

$$AE = \sqrt{\frac{\sum_{i=1}^{N_c} a_i^2}{N_c}} \tag{1}$$

a_i represents the audio data, N_c indicates the number contained in each data interval. The grayscale image is used

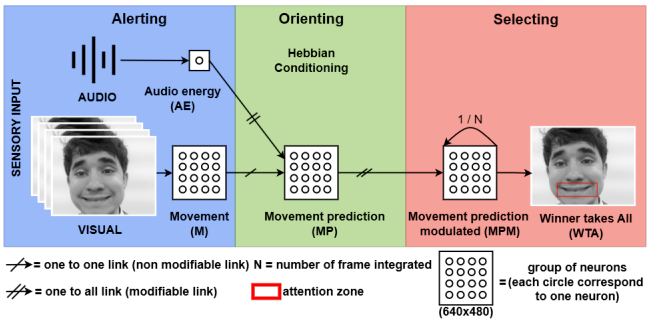


Fig. 3: Model of audio-visual attention (MAVA) will search for the most interesting zone, using bimodality between sound and image.

to form an optical flow (M : motion) with the preceding image. The matrix (MP : motion prediction) is obtained from the optical flow matrix using Hebbian enhancement. The Hebbian enhancement equation is

$$MP_i = w_i(t).AE \tag{2}$$

$$w_i(t+1) = w_i(t) + \varepsilon.AE.M_i \tag{3}$$

where w_i represents the weight of neuron i, ε denotes the learning rate, AE represents the audio energy, M_i is the movement associated with neuron i and MP_i is the movement prediction based on the output activity of neuron i.

The movement prediction (MP) is created using the activity of hebbian neurons. It is the result of optic flow modulation (MPM) : movement prediction modulated.

$$MPM_i(t+1) = M_i(t) * MP_i(t) + \frac{1}{n} * MPM_i(t) \tag{4}$$

Now that the points of the various maps have been learned, we can stop learning to finalize the network's training. The output of MAVA (the MPM map) will be used as input for NMFER. Its influence on emotion recognition will be moderated by a coefficient λ, allowing MAVA to contribute more or less to the final result.

C. Facial recognition

Also known as the "NMFER network", facial recognition enables the recognition of primary emotions. Here, we're interested only in the five emotions considered primary by Ekman : anger, happiness, surprise, sadness, and an emotion considered "neutral". As already mentioned, the network identifies the most likely facial expression expressed by the subject, using characteristic points on the subject's face. This detail selection process is carried out in three stages, presented in Fig. 4 from the grayscale image: convolution with a DOG kernel, then its addition with MAVA's MPM map according to the λ coefficient, which modulates the focus on the mouth and reduces the visual field to be used.

We will utilize MAVA's MPM map to dynamically adjust the visual field, progressively focusing it on the mouth.

The remaining components of NMFER will now be explained.

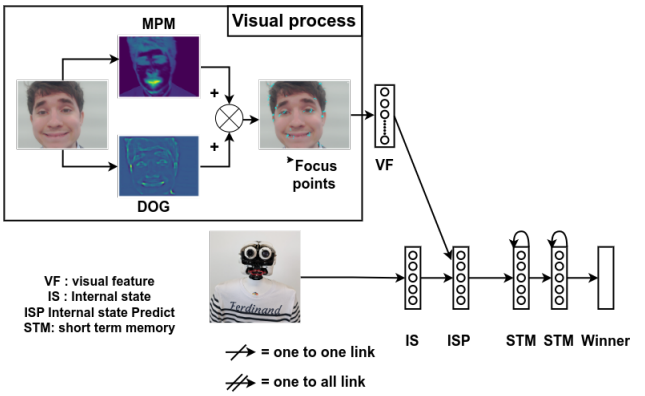


Fig. 4: Neural Model of Facial Expression Recognition (NMFER) is a architecture for facial emotional recognition, give the prediction based on the face associated with the most likely emotion among those learned.

a) SAW (self adaptative winner take all) : The extracted local view around each focus point is learned and categorized by a group of neurons V F (visual features) using a k-means variant that allows online learning and real-time functions [27] called SAW (Self Adaptive Winner takes all):

$$VF_j = net_j H_y(net_j) \tag{5}$$

$$net_j = 1 - \frac{1}{N}\sum_{i=1}^{N} |W_{ij} - I_i| \tag{6}$$

$H_y(x)$ is the Heaviside function :

$$H_y(x) = \begin{cases} x \ if \ y < x \\ 0 \ otherwise \end{cases} \tag{7}$$

I_i in Eq. 5 corresponds to the input signal VF_j is the activity of neuron j in group VF, net_j is the pattern recognition measure, y is the threshold for recognizing new information. H corresponds to the Heaviside function.

b) ISP: LMS This layer predicts the emotion perceived by the robot. It uses the LMS rule to update the weights according to the robot's internal state.

$$\Delta w_{ij} = \varepsilon \cdot VF_i \cdot (IS_j - ISP_j) \tag{8}$$

Δw_{ij} is the synaptic weight update, ε is the learning rate VF_i the visual features $(IS_j - ISP_j)$ is the error between the obtained state and the desired state.

In this study, the states represent the various emotions the robot must recognize on the participant's face. During the learning process, IS will provide the robot with the target state, while ISP represents the state the robot perceives on the participant's face.

c) STM layer (short term memory): This layer of neurons preserves information by using two of the parameters α and β corresponding to the coefficient of the input values and the coefficient of the values contained in the STM layer. Here, two STMs are used : one during training, and the other

during the test phase. This will ensure better prediction of the output emotion using multiple images.

$$stm(t+1) = \alpha * \ stm(t) + \beta * \ entry \tag{9}$$

III. Results

In this section, we present the results obtained from the experiments described above. Fig. 5 illustrates the dynamics of MAVA and the activity levels of three key facial regions : the mouth, eyes, and other areas. Notably, only the mouth and eyes exhibit a rapid increase in activity during the initial iterations, indicating that the MAVA model focuses primarily on the most dynamic regions of the face. Through the integration of auditory and visual inputs, the model is able to pinpoint the most active facial areas : the mouth and eyes.

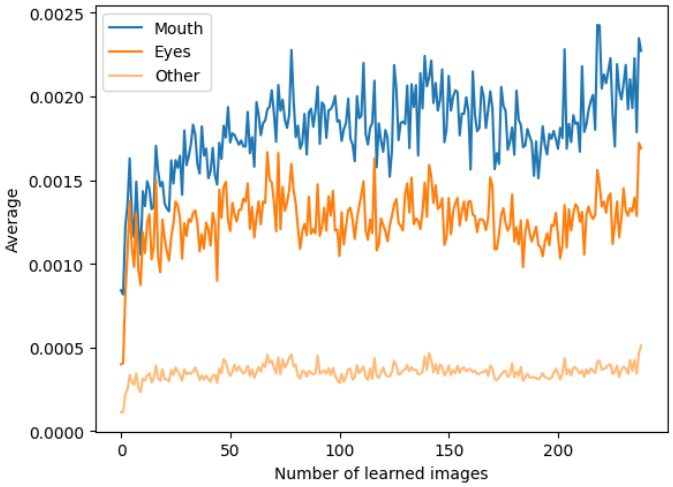


Fig. 5: Intensity of each zone during MAVA training, we record the values of all neurons in the target area and divide this area by the area of neurons in the matrix throughout the learning process.

Fig. 5 shows the different activities of three important facial zones : mouth, eyes and other. It can be seen that only mouth and eyes show a rapid progression during the first iterations. This shows that the MAVA model goes straight to the moving areas.Using sound and image, it identifies the most active areas of the face : mouth and eyes. The average value is obtained by summing the values on the surface of each area. The curve visually represents the progression of MAVA's output activity, highlighting its attention on the mouth. These results demonstrate that our bimodal attention model is particularly effective in focusing on the mouth and eyes regions.

Fig. 6 illustrates the effect of MAVA on focal point selection. For this analysis, we tested different values of λ to evaluate the impact of MAVA on focal point selection. The results show the evolution of MAVA's output activity, highlighting an increased focus on the mouth. These observations confirm that MAVA is effective in directing its attention toward key facial regions, particularly the mouth and eyes.

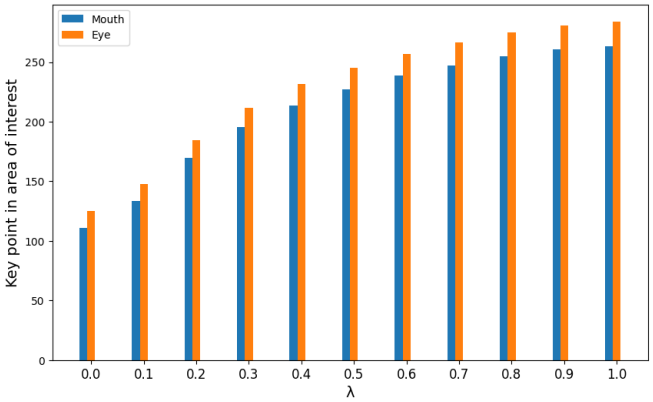
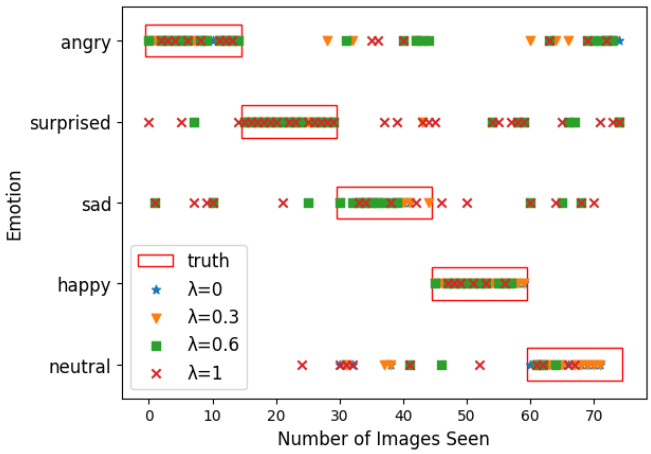


Fig. 6: Integration of MAVA with the MPM card for NFMER emotion recognition.

We can see with this figure Fig. 7 the impact of MAVA for the selection of characteristic points. Fig. 7 a) allows us to see how, through the impact of MAVA and λ, the different emotions will be less and less perceptible by the robot. And in b), we can observe an attraction of key points towards the bottom of the face, and especially the mouth.

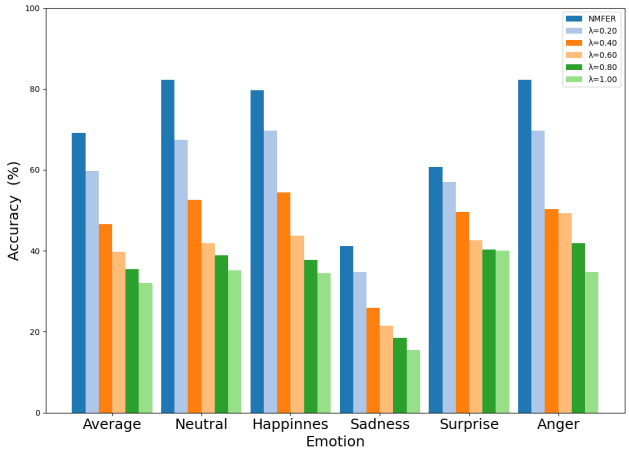
Fig. 8 illustrates the performance of the NMFER model. First, the results indicate that the best performance is achieved without MAVA (blue curve), meaning that facial expressions are better recognized when the robot analyzes the entire face of its partner. In this case, recognition rates range between 65% and 85%. Moreover, the model reaches these performance levels with a small number of iterations (25), demonstrating a fast, online, and real-time learning process. Next, the results highlight the impact of MAVA on facial expression recognition. We use the parameter λ to adjust MAVA's influence in the recognition process. The findings show that the more MAVA is considered, the lower the performance. Fig. 8 a) illustrates the impact of MAVA when it is activated only during the validation phase, while Fig. 8 b) shows its effect when the mechanism is used during both the training and validation phases.
Fig. 8 b) illustrates the impact of MAVA during both the learning and validation phases of emotion recognition. Initially, recognition performance starts at around 70% on average and gradually declines to 45%, but at a much slower and more progressive rate. Unlike previous cases, all expressions show a similar decrease. In conclusion, Fig. 8 demonstrates a decline in emotion recognition performance compared to the scenario where the system operates without MAVA. With these two results, Fig. 8 a) and Fig. 8 b), we can see a decrease in emotion recognition compared to the case where our system recognizes emotions in a classical way. We can see that using MAVA in the learning phase reduces the loss of recognition for different emotions. This indicates that the less abrupt the shift in attention (meaning the model focuses its attention on the mouth from the start of learning), the less impact it will have on emotion recognition.



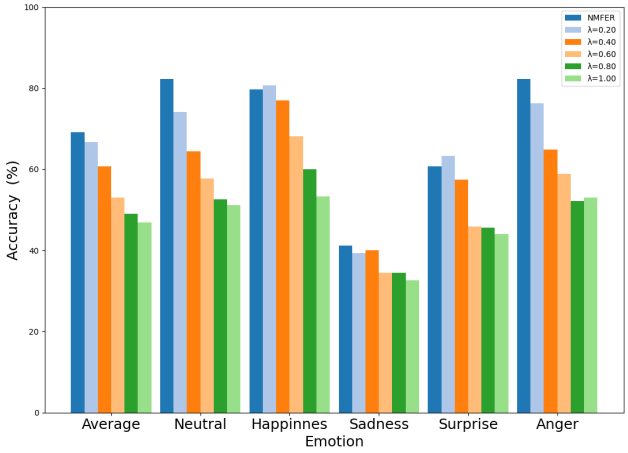
(a) Facial Emotional Recognition with MAVA during testing

(b) Key point found for ”happiness” emotion for different MAVAœ values

Fig. 7: (a) Emotion recognition with MAVA during testing. Colors indicate different values of λ (attention directed towards the mouth). (b) Key points for the ”joy” emotion and different values of λ. The model's attention (blue points) focuses on the mouth as λ increases.



(a) Facial Emotional Recognition with MAVA during testing



(b) Facial Emotional Recognition with MAVA during learning and testing

Fig. 8: NMFER recognition with MAVA This graph shows the facial recognition accuracy of each of the emotions we have chosen to study, as well as the average of all emotions (far left of both graph). The possible emotions are “neutral” happiness”, ‘sadness’, ‘surprise’, ‘anger’. The MAVA synaptic link value starts at zero (this is equivalent to NMFER) and ends at 1. a) shows the result when MAVA is used only for the test part. b) shows the result when MAVA is used for the whole experiment.

IV. Conclusion

This research explores the interaction between emotion recognition and attention, showing that attention modulates the ability for emotional recognition. The integration of attention into our model has allowed us to generate predictions about the dynamics of the recognition of different emotions, in accordance with observations from experimental psychology.

This study highlights two major findings. First, MAVA helps to focus attention on the most relevant areas of the face during face-to-face interactions. The results reveal a particularly strong focus on the mouth, while the rest of the face generates less activity. The second finding concerns the impact of MAVA on facial emotion recognition. Without MAVA, the model achieves an accuracy of 70%. However, as the selection of key facial points is progressively reduced, accuracy declines to approximately 35% when the selection is heavily restricted to the mouth. In contrast, when MAVA is used during both training and validation phases, the drop in accuracy is less pronounced, reaching a threshold just below 50%. These observations emphasize MAVA's role in optimizing facial emotion recognition. The increased interest of the mouth in the visual field is in line with the study showing that there will be a decrease in emotion recognition [24]. This study demonstrates that the way in which faces are analyzed influences both the recognition and learning of

facial expressions.

In the case of Fig. 8 a). We can consider that the restriction of the visual field only at the time of the test corresponds to the mechanism highlighted in the two articles [4] [5]. Where participants developed emotion recognition on complete faces but are disturbed by face modifications. The result showing a decrease down to 35% recognition is coherent. Not all emotions experience a similar drop in recognition. Surprise and sadness were found to decline more slowly than other emotions when MAVA was used.

A decrease in the accuracy of emotion recognition was observed in our model, which is consistent with the results presented in [24].

In future work on developmental robotics, we aim to continue using facial recognition to support the robot's development [26] [22] [27]. Additionally, we plan to further evaluate our model in terms of its predictive capabilities to determine whether it can predict conditions such as autism or attention deficits [28] [29]. Finally, we would like to explore dynamic visual fields that replicate how an infant's attention evolves during the first year of life.

ACKNOWLEDGMENT

This work was funded by CY Initiative (grant "Investissements d'Avenir" ANR-16-IDEX- 0008).

REFERENCES

[1] J. M. George and E. Dane, "Affect, emotion, and decision making," *Organizational Behavior and Human Decision Processes*, vol. 136, p. 47–55, Sep. 2016.

[2] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, p. 169–200, May 1992.

[3] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[4] M. Grahlow, C. I. Rupp, and B. Derntl, "The impact of face masks on emotion recognition performance and perception of threat," *PLoS ONE*, vol. 17, no. 2, p. e0262840, Feb. 2022.

[5] G. Kim, S. H. Seong, S.-S. Hong, and E. Choi, "Impact of face masks and sunglasses on emotion recognition in south koreans," *PLOS ONE*, vol. 17, no. 2, p. e0263466, Feb. 2022.

[6] J. N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of Personality and Social Psychology*, vol. 37, p. 2049–2058, 1979.

[7] L. Abbruzzese, N. Magnani, I. H. Robertson, and M. Mancuso, "Age and gender differences in emotion recognition," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02371

[8] F. Pons, L. Bosch, and D. J. Lewkowicz, "Bilingualism modulates infants' selective attention to the mouth of a talking face," *Psychological science*, vol. 26, no. 4, p. 490–498, Apr. 2015.

[9] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial emotion recognition (fer) using convolutional neural network (cnn)," *Procedia Computer Science*, vol. 235, p. 2079–2089, Jan. 2024.

[10] S. Ramalingam and F. Garzia, "Facial expression recognition using transfer learning," in *2018 International Carnahan Conference on Security Technology (ICCST)*, Oct. 2018, p. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8585504

[11] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, vol. 411, p. 340–350, Oct. 2020.

[12] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, *Recurrent Neural Networks for Emotion Recognition in Video*, Nov. 2015.

[13] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested lstm for facial expression recognition," *Neurocomputing*, vol. 317, Aug. 2018.

[14] M. Singh, A. Majumder, and L. Behera, "Facial expressions recognition system using bayesian inference," in *2014 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2014, p. 1502–1509. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6889754

[15] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: a comprehensive review," *Neural Computing and Applications*, vol. 36, no. 16, p. 8901–8947, Jun. 2024.

[16] L. Boccanfuso, Q. Wang, I. Leite, B. Li, C. Torres, L. Chen, N. Salomons, C. Foster, E. Barney, Y. A. Ahn, B. Scassellati, and F. Shic, "A thermal emotion classifier for improved human-robot interaction," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2016, p. 718–723. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7745198

[17] T. Horii, Y. Nagai, and M. Asada, "Imitation of human expressions based on emotion estimation by mental simulation," *Paladyn, Journal of Behavioral Robotics*, vol. 7, no. 1, Dec. 2016. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/pjbr-2016-0004/html

[18] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, p. 119–155, Jul. 2003.

[19] A. Aubret, C. Teulière, and J. Triesch, "Self-supervised visual learning from interactions with objects," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, p. 54–71.

[20] H. Kim, E. Murphy-Chutorian, and J. Triesch, "Semi-autonomous learning of objects," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, Jun. 2006, p. 145–145. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1640591

[21] J. Leitner, S. Harding, P. Chandrashekhariah, M. Frank, A. Förster, J. Triesch, and J. Schmidhuber, "Learning visual object detection and localisation using icvision," *Biologically Inspired Cognitive Architectures*, vol. 5, p. 29–41, Jul. 2013.

[22] S. Boucenna, P. Gaussier, P. Andry, and L. Hafemeister, "A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game," *International Journal of Social Robotics*, vol. 6, no. 4, p. 633–652, Nov. 2014.

[23] R. Bergoin, S. Boucenna, R. D'Urso, D. Cohen, and A. Pitti, "A developmental model of audio-visual attention (mava) for bimodal language learning in infants and robots," *Scientific Reports*, vol. 14, no. 1, p. 20492, Sep. 2024.

[24] M. Guarnera, Z. Hichy, M. I. Cascio, and S. Carrubba, "Facial expressions and ability to recognize emotions from eyes or mouth in children," *Europe's Journal of Psychology*, vol. 11, no. 2, p. 183–196, May 2015.

[25] A. Ayneto and N. Sebastian-Galles, "The influence of bilingualism on the preference for the mouth region of dynamic faces," *Developmental Science*, vol. 20, no. 1, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12446

[26] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 5504, pp. 599–600, 2001. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.291.5504.599

[27] C. Hasson, P. Gaussier, and S. Boucenna, "Emotions as a dynamical system: the interplay between the meta-control and communication function of emotions," *Paladyn*, vol. 2, Mar. 2012.

[28] S. Boucenna, S. Anzalone, E. Tilmont, D. Cohen, and M. Chetouani, "Learning of social signatures through imitation game between a robot and a human partner," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 3, p. 213–225, Sep. 2014.

[29] S. Boucenna, D. Cohen, A. N. Meltzoff, P. Gaussier, and M. Chetouani, "Robots learn to recognize individuals from imitative encounters with people and avatars," *Scientific Reports*, vol. 6, no. 1, p. 19908, Feb. 2016.