Master 2

Report

Emergence of secondary facial expressions during human-robot interaction

by

Dorian IBERT

Artificial intelligence and robotics training student

ETIS UMR8051, CY Cergy Paris Université / ENSEA / CNRS
6 avenue du Ponceau, 95014 Cergy-Pontoise Cedex, France

Supervised by :

Pr. Sofiane Boucenna ENSEA

Supported on 31 mars 2021

In front of the jury composed of :

Pr. Lola Canamero CY University Reporter

I. Objectives and context

Human can produce a multitude of facial expressions other than Ekman's six primary emotions. They can express mixed expressions. In this project, we want to verify whether a perception-action architecture based on neural networks can perform this type of recognition. In contrast to existing models, we seek to bring out this ability through the neural architecture. We will show how the expressive head will be able to reproduce secondary expressions corresponding to slightly more sophisticated expressions (mixtures of primary expressions). In order to obtain a robot capable of producing any facial expression, rather than a finite set of facial expressions, our solution is to learn muscle groups of the face contracting together. They are these muscle groups that we will call motor primitives that will be learned separately by the robot.

The child's learning is the context of the project. Indeed, we are trying to reproduce the same learning as that of a young child through the perception of expressions made by others (usually someone close to the child) and the association of this expression with an internal emotional state. We will verify that the robot, like a child, is able to recognize new, more sophisticated emotions from a learning of a handful of basic emotions such as anger, sadness, joy or surprise, and then associating these with a new intermediate internal state.

II. Structure of the manuscript

In order to take the subject fully in hand, my work so far has been to carry out a complete state of the art of the research carried out on the theme of emotions in robotics. This state of the art also aims at investigating whether different works have been recently conducted on the capacity of emergence of emotions via neural networks. This state of the art includes:

- History, definition, and theories related to the origin of emotions
- The main works carried out today
- Work on the emergence of secondary expressions via neural networks

Following this state of the art, the project was structured in different steps:

- Getting to grips with the neural models used.
- Adaptation of the current neural model for the recognition of primary facial expressions
- Reflection on the encoding of expressions to constitute the model for the recognition of secondary facial expressions.
- Creation of different models for the recognition of secondary facial expressions
- Analysis of the results on 1 person and comparison of the models.
- Conclusion

III. History, definitions and theories [1].

Philosophers were the first to try to define the notion of emotions and their functioning. Plato hypothesised a tripartite structure of the soul (Dialogues, about 400 BC). Aristotle, in Rhetoric II, gave a first definition: "Emotions are all those feelings which change man in such a way as to affect his judgment and which are accompanied by suffering or pleasure".

Over the centuries, different philosophical currents have emerged based on the emotions, such as Stoicism "We must gain control over our emotions and accept what does not depend on us" and the body/soul dualism (Descartes). Different definitions are thus attributed to the emotions, but it is only in the 19th century that this branch of philosophy dedicated to the study of the soul becomes a science, called psychology, having recourse to the experimental method, to statistics and to mathematical models.

Charles Darwin (1809-1882) was the originator of this science and laid the foundations for the expression of emotions.

He will describe them as innate, universal and communicative. And his work will allow the development of principles called Darwin's principles. Darwin's first principle explains how a reaction that was initially voluntary becomes innate and reflexive over the generations. Darwin's last principle established the link between emotion and the nervous system **[2]**.

We will also mention some pioneers in the study of emotions since Darwin, William James (creation of the first psychology laboratory) and Carl Langen support the peripheralist theory (19th century) while Walter Cannon and Philip Bard study the centralist theory (20th century). These two theories oppose each other as shown in Figure 1.

James Paper proposes a brain circuit as an emotional process, shown in Figure 2, called the Papez circuit.

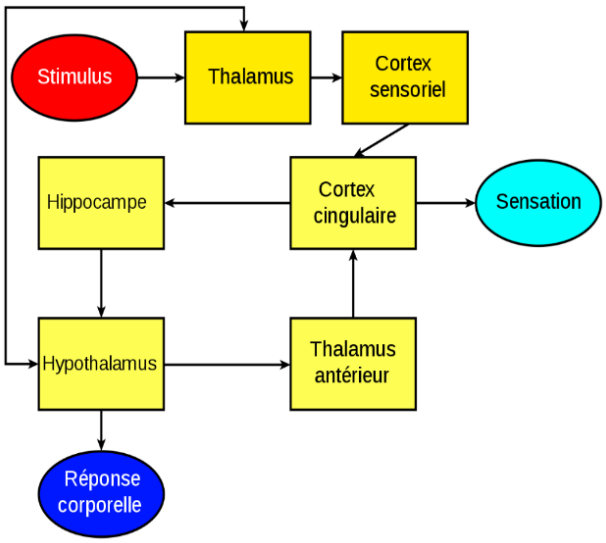
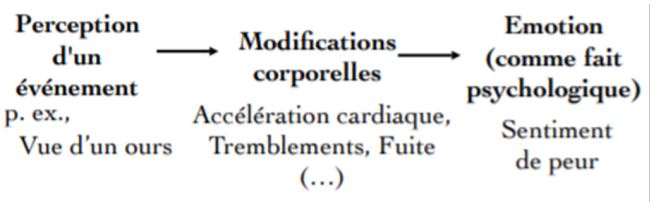


Fig. 2. The Papez circuit

Paul MacLean proposed an evolutionary theory, called the triune brain theory, in which the brain is the result of progressive evolution.

Various works during the 20th century have led to the emergence of new theories, which are essential for understanding the work done in affective computing today.

The theory of facial feedback, developed by Robert Soussignan, proposes that facial movements can modulate emotional feelings.



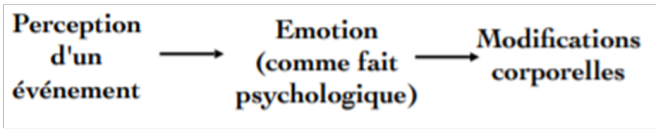


Fig. 1. Peripheralist vs. centralist theory

The bi-factorial theory, invented by Stanley Schachter, proposes that the triggering of a specific emotion is determined by two components, a physiological activation (translated by a state of alertness) and the awareness of the situation triggering this physiological activation. This physiological activation would determine the intensity of the emotion while cognition would determine the type of emotion. The theory of basic emotions (Paul Ekman) is based on the existence of a limited number of universal basic emotions, called primary emotions. And that the more complex emotions (boredom, frustration...) called secondary emotions, come from a mixture of these basic emotions. Paul Ekman presents six primary emotions: anger, fear, joy, sadness, surprise and disgust. These emotions are also located in distinct brain regions (amygdala, insula, etc.). This theory is consistent with James' theory that emotions can be differentiated according to bodily characteristics (heart rate, temperature) as shown in Figure 3. This would explain why, conversely, fear of a situation is intensified by the infrequency of its occurrence (conditioning).

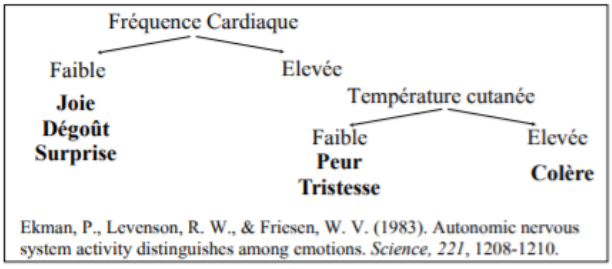


Fig. 3. Agreement of Paul Ekman's theory with Darwin's first principle

Dimensional theories are based on the idea that an emotion can be determined according to its value on several dimensions. Wilhelm Wundt suggests the existence of three dimensions: pleasure/ displeasure, exciting/depressing and stressing/relaxing. James Russell proposed a two-dimensional model whose axes would be valence (pleasure/pleasure) and activation (weak/strong). This model was taken up by Klaus Scherer in order to replace activation by intensity, allowing for a more complex model as shown in Figure 4 **[3]**. Finally, Charles Osgood presents a three-dimensional model with three dimensions: evaluation (positive/negative), activation (high/low) and power (high/low).

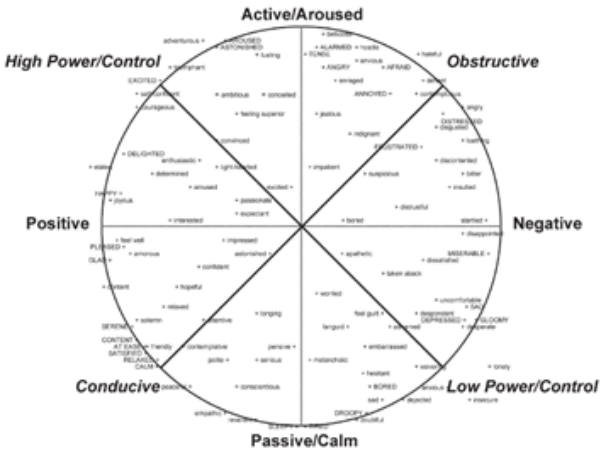


Fig. 4. Dimensional representation of emotions according to Klaus Scherer

Cognitive appraisal theory states that the emotion we feel about an event is determined by our cognitive assessment of its relevance to our well-being, and our ability to control the consequences of that event **[4]**. Together these theories form the psycho-evolutionary theory, named by Robert Plutchik. He also adds two emotions to the basic emotions: confidence and anticipation and thus establishes the wheel of emotions visible in Figure 5.

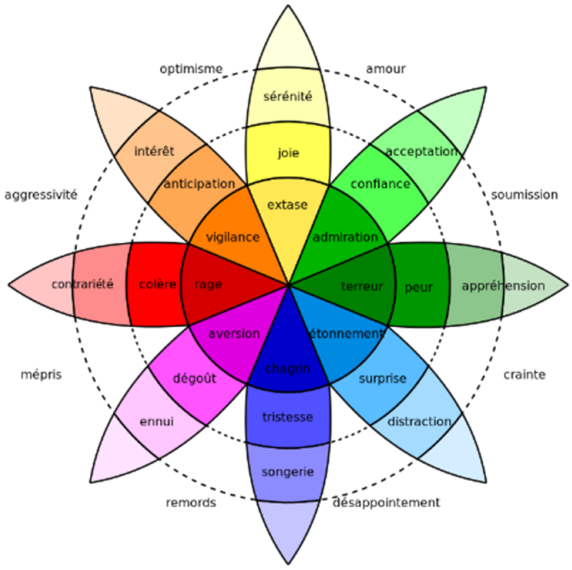


Fig. 5. Plutchik's flower of emotions

These theories highlight different ways of dealing with emotions in a robot. A system using only one source of emotion perception will be called unimodal, while a system considering different sources (physiological, behavioural and/or psychological) will be multimodal **[5][6]**.

IV. The main works carried out today

Depending on the study conducted, one or more theories will be favoured in robotics, if the system has a memory for example, the adaptationist theory should be included in the model **[7]**.

A multimodal system would a priori be the most realistic to reproduce the activation of emotions in adults. However, this project aims to reproduce the learning of a young child, i.e. an individual who does not have the same brain capacity as an adult (situational awareness, hindsight, etc.) and different theories can be dismissed from the subject.

During this state of the art, I have gone through different theses presenting the work done in recent years. This allowed me to target the usefulness of affective computing, the different creations and the state of progress of research today.

The addition of emotions to the robot is a key point in psychology, for example to provide psychological support to people through conversation **[8]**, but also to study the social effect of a robot-child interaction **[9]** and the notion of anthropomorphism created associated with this interaction **[10][11]**.

I first read M. Boucenna's thesis **[12]** in order to know which models were at the origin of this project, but also the different methods used in signal processing. The objective being to get closer to the learning of the young child, only the theory of basic emotions was used at first, via the use of motor primitives **[13]**. Learning was achieved through mimicry. The robot, composed of different internal states associated with primary emotions, interacts with a human mimicking one of these primary emotions. A two-minute interaction allows the robot to link all the mimicked facial expressions with its corresponding internal states. The verification is done by performing the reverse experiment. In a second step, secondary emotions were added. Review of the theses associated with the most famous robots in affective robotics in order to compare the models and methods used: KISMET **[14][15]**, EDDIE **[16]**, EINSTEIN **[17]**, GRETA **[18]** The child's perception of emotion is produced through the senses of sight, touch, and hearing. Various studies merge these multimodal data to extract the perceived emotion **[14]**. The robot used for the project is equipped only with the sense of sight, and different models for detecting a face exist. There are three main families: models with knowledge of the structure of the human face (characteristic parts, relations between the different elements making up a face, classifiers), global models (using statistical methods, neural networks, eigenfaces) and hybrid models. Globally, the different research use the first family, with Gabor filters mounted in cascade allowing the detection of significant areas of the face characteristic of emotions, and, according to the theory, to detect an expressive intensity and its level of activation. **[17]**.

Cynthia Breazal's work, firstly on the Kismet robot designed for parent/infant interaction **[19][15]**, and the research on the interaction between a robot and a nurse **[20]** seem to be the most similar to this project, however the autonomous learning side has not been studied.

Subsequently, my research focused on work that aimed to generate unlearned emotions, i.e. to make unexpected behaviours emerge, but also to have the robot associate an internal state with that behaviour. I have selected the various most interesting works on this subject but the approach we have is different from all of them. **[21][22][23][24][25][26]**.

V. Methods and techniques used

The model is composed of two phases, a learning phase, during which the robot asks the user to perform the different primary facial expressions, and a testing phase aimed at validating the model, during which the user expresses any expression, and the robot displays the recognised expression. As the perception during these interactions is only visual, different image processing is performed. The OpenCV library under Python is used to capture the user's expressions with the camera, and then to visualise themselves if necessary. Using the SIFT[1] algorithm, also available in OpenCV, a fixed number of keypoints are selected, and for each keypoint a descriptor is calculated to extract unique points of the face **[27]**. These descriptors will constitute the input of a network of SAW[2] neurons thus activating a winner neuron. The activity of this winning neuron will activate one or more LMS[3] networks reflecting the internal emotional states of the robot or the muscle groups according to the model. STM[4] are used to accumulate the activities of the neurons of the LMS stages to facilitate the decision concerning the detected expression.

VI. Primary facial expression recognition

This model consists in distinguishing 5 primary emotions according to the method presented above, which is represented in figure 6 below.

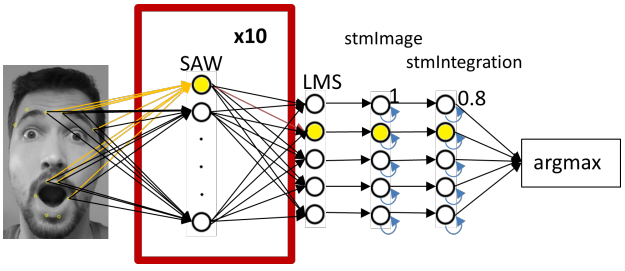


Fig. 6. Network structure

To obtain the most effective results, various parameters need to be set or optimised, including:

- Learning time
- The vigilance of the SAW

[1] scale-invariant feature transform
[2] self-adaptive winner (takes all)
[3] last mean squares
[4] short term memory

- The maximum number of SAW neurons
- The number of keypoints used in the SIFT

In order to achieve fast learning, the camera images are captured 10 by 10. Thus, for each facial expression presented by the user, 10 images are captured, from each of which descriptors are extracted and the network is updated.

The number of descriptors (created from keypoints) is initially set at 10 to limit the number of calculations, the aim being to have a real-time programme. Each descriptor of a captured image is the input to the SAW, so for each descriptor the activity of the SAW neurons is calculated according to the following formula:

$$a_j = \gamma_j.H(\gamma_j\ ) \quad where\ \gamma_j = 1 - \sum_{i=1}^{128} \frac{|w_{i_j} - desc_i|}{128}$$

wij corresponds to the weight of neuron j for input i, itself corresponding to component i of the descriptor and H the Heaviside function:

$$H(\gamma) = \begin{cases} 1 & \text{si } \gamma > max(vigilance, \mu + \sigma^2) \\ 0 & \text{sinon} \end{cases}$$

Where vigilance represents a user-determined threshold, μ and σ respectively the mean and variance of neuron activities.

If this descriptor induces sufficient activity, i.e. the neuron with the maximum activity has an activity above the set vigilance threshold, then the weights of this neuron are updated:

$$w_{ij} = w_{ij} - \varepsilon.(1 - a_j).(desc_i - w_{ij})$$

Where ε is the learning step, set at 0.01 here.

If this descriptor does not induce sufficient activity, then a new SAW neuron will be recruited, until the maximum limit is reached, at which point the weights will be updated.

The LMS links the activities of the SAW neurons to internal states by the following weight learning rule:

$$w_{ij} = w_{ij} - \varepsilon'.(a_j - a_j^d).x_i$$

Where ε' is the learning step, fixed at 0.1 here, xi is the input of the LMS i.e. the output of one of the SAW neurons, a_j is the activity of neuron j of the LMS, a_j^d is the desired activity of this neuron, reflecting the desired emotion.

As soon as the series of 10 images ends, a new expression is requested, corresponding to a new vector a_j^d and the learning process is repeated.

During the test phase, there is no more updating of weights, the activity of the LMS neurons is stored in a memory for each keypoint, for each image, the neuron in the memory with the highest activity designates the perceived emotion.

In order not to select disruptive keypoints and to obtain the best results, the experiment is carried out on a white background and only the user's face occupies the camera's capture area.

The results of this method are shown in Figure 7, the learned expressions are "joy", "surprise", "anger", "sadness" and "no emotion", a success occurs when the model detects the correct expression.

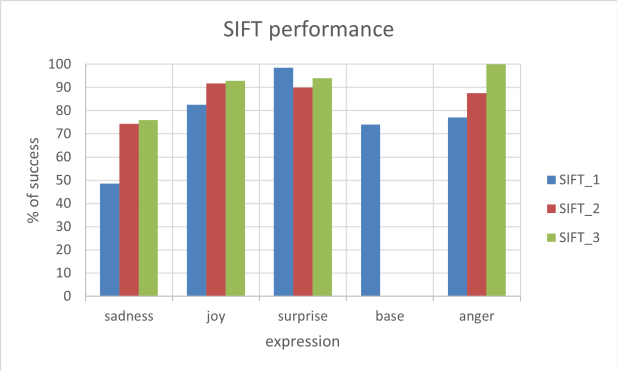


Fig. 7. Results from 3 experiments, in each of which 300 images are presented, i.e. approximately 100 per expression

However, there is a drawback: several keypoints detected by the SIFT algorithm are located in irrelevant places, such as the hair. These keypoints are therefore useless, they do not disturb the algorithm but lower its efficiency, especially for mixed expressions which require the location of muscle groups.

However, there is a drawback: several keypoints detected by the SIFT algorithm are located in irrelevant places, such as the hair. These keypoints are therefore useless, they do not disturb the algorithm but lower its efficiency, especially for mixed expressions which require the location of muscle groups.

There are several ways to improve the scores:

- Increasing the number of keypoints to increase the recognition percentage, however, increases the number of calculations and is not in favour of real-time.
- Use motion-detecting optic flow to target keypoints on moving facial areas such as the eyebrows, or the corners of the mouth. This is not possible because the learning is static here, the optical flow would only be useful in the test phase but would reveal points that have not been learned.
- The use of the dlib library **[28]**, which performs image processing to extract points located on the facial contours, as shown in Figure 8. The descriptors associated with these points are then calculated.

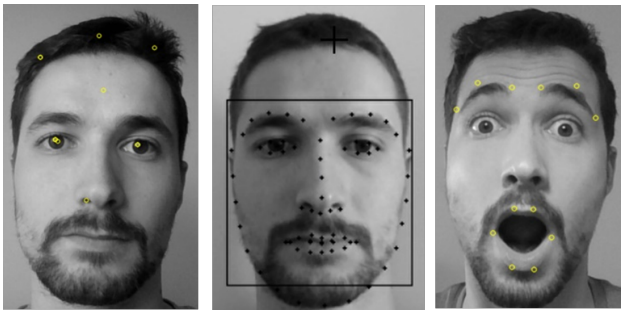


Fig. 8. Difference in the location of the keypoints between SIFT and DLIB, on the left the keypoints located by the SIFT algorithm, in the middle the points determined by the dlib library, on the right the selected points

The interest of dlib is that the upstream algorithm is similar to that of SIFT detection, while SIFT performs a DoG[5] to determine rotationally and scale invariant keypoints, dlib uses

[5]difference of gradients

face detectors based on HoG[6] **[29][30]** and linear SVM[7] **[31]** classifiers. Once a face is detected, a set of regression trees are used to quickly estimate the position of the face landmarks **[32]**.

One of the advantages of dlib is that a white background is no longer necessary to avoid detecting disturbing points during the experiments, however, in order to remain within the framework established at the beginning of the project, we will continue these experiments under the same conditions.

Thus, the dlib library provides a total of 68 points for the location of facial landmarks, 12 points deemed relevant for the detection of emotions are extracted, to which descriptors are calculated in the same way as for the SIFT algorithm.

An average of the performance of the methods using the dlib library and the SIFT algorithm to locate keypoints is shown in Figure 9.

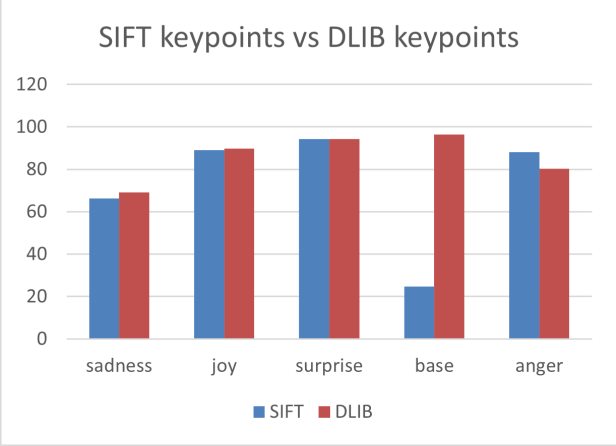


Fig. 9. Performance comparison of methods using SIFT and the dlib library

The results are globally similar, except for the expression "base" i.e. "without emotion", indeed, for this expression the keypoints of the SIFT are located at the same places as those for joy, and the program seems to distinguish joy when expressing "base" as shown in Figure 10, the dlib library thus makes it possible to deal with this problem

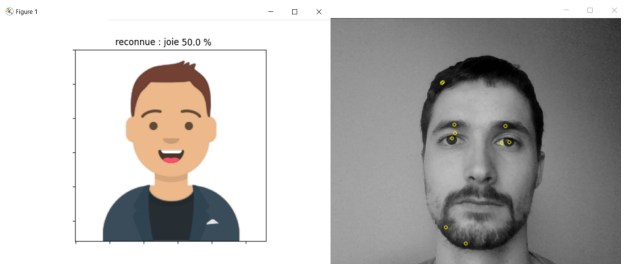


Fig. 10. Recognition error via the SIFT algorithm, while I made no-emotion face, the model detects joy

In order to guarantee a smoother convergence towards an emotion and to avoid aberrant uncertainty errors, a second memory with a so-called "slip" coefficient fixed at 0.8 is added. At time t+1, the activity of neuron j in this STM is:

$$a_j(t+1) = 0.8.a_j(t) + b_j(t)$$

Where $b_j(t)$ is the activity of neuron j at time t of the STM output from the LMS.

To make the experience more vivid during the test phase, an avatar based on the pyavataaaars library **[33]** (Figure 11) was created to reflect the perceived emotion.

Fig. 11. Avatars to represent perceived expression

VII. Secondary facial expression recognition

The models dealing with secondary emotions are very similar to the previous one.

The difference is in the characteristics to be learned. In the previous model the internal states were emotions themselves, in this one the internal states represent muscle groups, and it is the arrangement of these muscle groups that characterises the emotions. In order to establish a simple model beforehand, we select emotions with orthogonal muscle groupings.

So we start with

$$joy : \begin{cases} normaleyebrows \\ stretchedlips \\ littleopenedmouth \end{cases} \quad surprise : \begin{cases} raisedeyebrows \\ normalstretching \\ openedmouth \end{cases}$$

$$anger : \begin{cases} frownedeyebrows \\ pinchedlips \\ closedmouth \end{cases}$$

A vector is assigned to each muscle grouping, so we obtain 3 vectors forming an orthogonal basis, each of which is linked by a LMS to the SAW:

$$\left\{ \begin{pmatrix} \text{eyebrows raised} \\ \text{lips stretched} \\ \text{mouth opened} \end{pmatrix}, \begin{pmatrix} \text{eyebrows normal} \\ \text{lips normal} \\ \text{mouth little opened} \end{pmatrix}, \begin{pmatrix} \text{eyebrows frowned} \\ \text{lips pinched} \\ \text{mouth closed} \end{pmatrix} \right\}$$

Thus we have:

$$\text{joy: } \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\}, \quad \text{surprise} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\},$$

$$\text{colère} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$$

[6]Histogram of gradients
[7]Support Vector Machines

During the learning phase, the primary emotions ”joy”, ”surprise” and ”anger” are presented to the robot and to each the robot associates 3 internal states as shown in Figure 12.

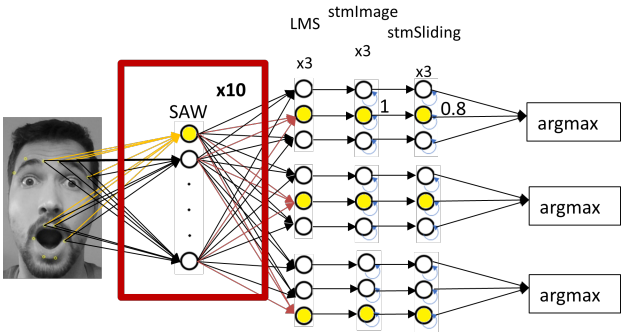


Fig. 12. Network structure for secondary expression recognition

The objective is to analyse whether, following training, the network can segment the SAW according to muscle groupings, i.e. each LMS, receiving SAW activity as input and according to facial descriptors, generates sufficient activity only for the descriptors associated with its muscle grouping. Such flexibility allows the identification of mixed expressions such as frowning and smiling.

To carry out the experiments, we select 6 keypoints on the eyebrows, and 6 others on the mouth as shown in Figure 8.

As the avatar created does not have enough facial expressions to display mixed expressions during the test phase, the display is done via prints.

However, short tests with mixed emotions reveal that the algorithm does not distinguish between individual muscle groups, the 3 muscle groups always come out together depending on the primary emotion as shown in Figure 13.

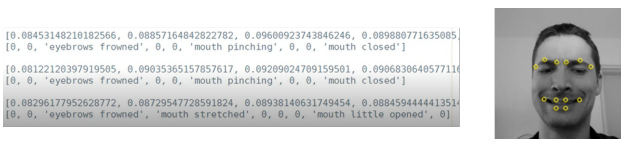


Fig. 13. Mixed expression of anger and joy

Different settings of vigilance and learning time led to the same intuition. In order to be able to target the points on which to work, it is necessary to collect statistics on the performance of expression recognition locally and globally. Locally means analysing whether a keypoint on an eyebrow responds to the correct expression expressed by an eyebrow, and globally means analysing whether all keypoints return the correct expression for each muscle group.

Since learning several times is very time-consuming, an image database is created to speed up the process. In addition, this allows for a more stochastic process that promotes network flexibility.

The database consists of :

- 850 images per primary expression (so x3) for the learning base

- 20 * (3 primary expressions + 18 secondary expressions) for the test base

For these tests, we count the number of images for which the model has recognised the 3 correct positions of the muscle groups.

According to the results in Figure 14, the model seems to fit at least for the primary emotions, with a lowest result of 73% recognition on 200 images per expression.

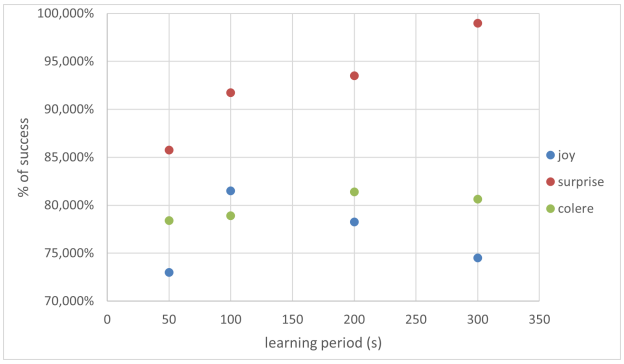


Fig. 14. Model performance on primary expressions

To collect data, we determine the number of times that the keypoints responded with the correct expression in the same 20 secondary expression images.

20 images, 12 keypoints per image :
- 6 keypoints for eyebrows → 6*20 = 120 outputs for eyebrows
- 2 keypoints for lip stretching → 40 outputs
- 4 keypoints for lip opening → 80 outputs

A total of 240 outputs for 1 mixed expression (20 pictures)

A picture of the debugging is shown in Figure 15.

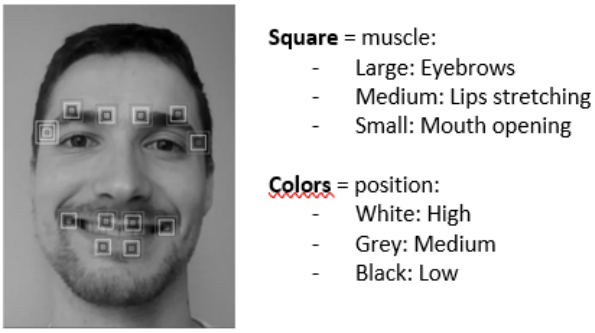


Fig. 15. Detail for data analysis

A first set of tests allowed us to set the vigilance to 0.97 for a learning time of less than 300 seconds (5 minutes) in order to obtain the best performances (Figures 16 to 19). For a fixed number of neurons, a learning time that is too long tends to average out the weights and the network no longer evolves.

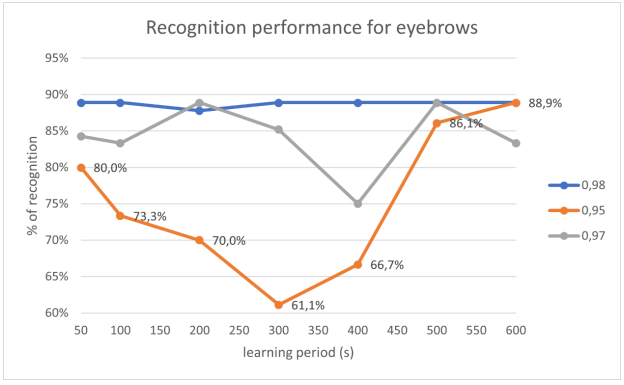


Fig. 16. Number of times the eyebrows detected the correct position of their muscle grouping according to vigilance and learning time

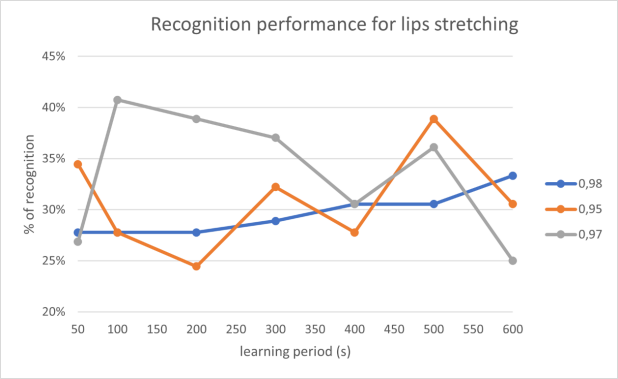


Fig. 17. Number of times the keypoints associated with the lip stretch provided the correct position of their muscle grouping according to vigilance and learning time

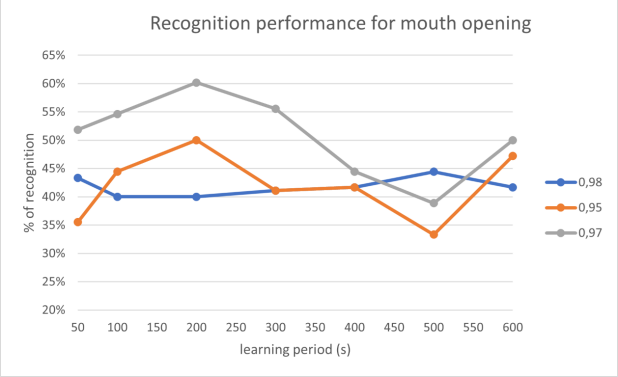


Fig. 18. Number of times the keypoints associated with opening the mouth provided the correct position of their muscle grouping according to vigilance and learning time

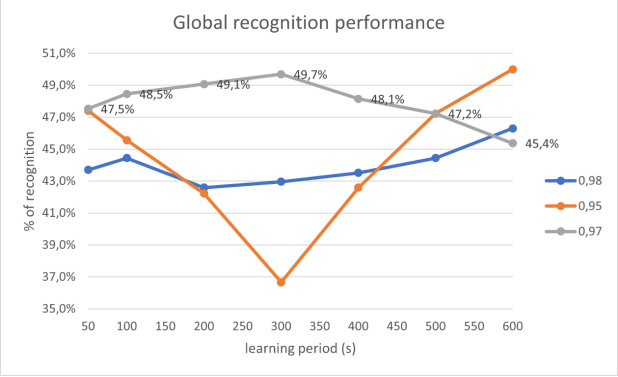


Fig. 19. Number of times the 3 correct muscle group positions were detected

Locally, the eyebrows provide the response associated with the image perceived during the test. Indeed, out of 20 images, with 6 keypoints for the eyebrows, 120 responses are obtained and more than 80% of the responses given are good. The stretch is poorly perceived, but it should be remembered that there are only 2 keypoints for this muscle group, and the opening is around 60

However, when we look at the 1 by 1 cases for a mixed expression (Figure 20), we can see that for most expressions, the model constantly hesitates between 2 primary emotions: it recognises two emotions more or less strongly but cannot bring itself to combine these two expressions to make one.

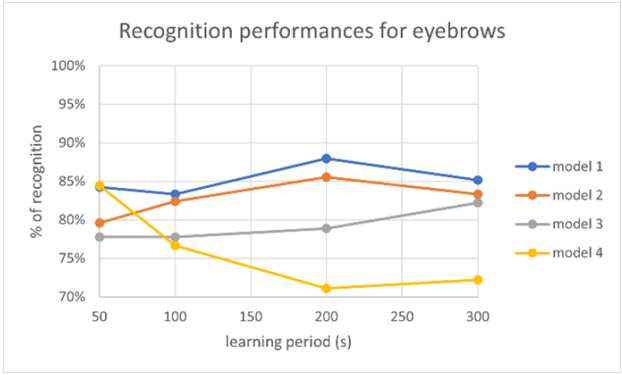
image	eyebrows	eyebrows	eyebrows	mouth stre	mouth nor	mouth pin	mouth ope	mouth littl	mouth close
joie	5	181	54	181	4	55	4	179	57
surprise	216	2	22	4	213	23	214	4	22
colere	30	7	203	8	29	203	29	8	203
mix_0	109	90	41	90	104	46	111	88	41
mix_1	64	106	70	114	57	69	56	109	75
mix_2	58	85	97	93	54	93	58	90	92
mix_3	124	15	101	23	118	99	125	13	102

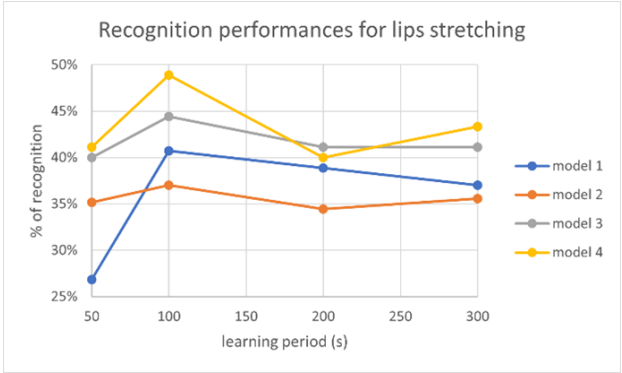
Fig. 20. For example, for the easy mixed expression 3, the expression was: ”eyebrows raised, mouth pinched and closed”, translating into surprise + anger

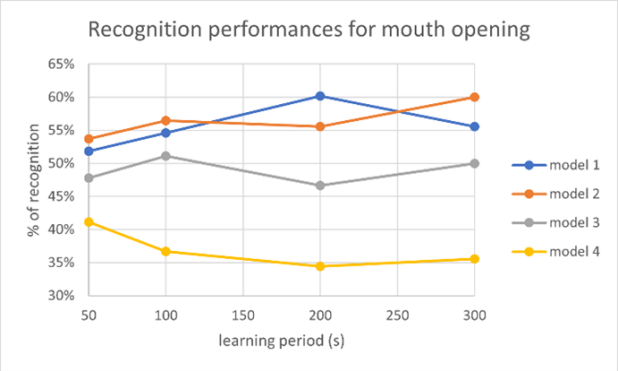
The current model is therefore unable to segment the SAW but manages to distinguish that there are several emotions.

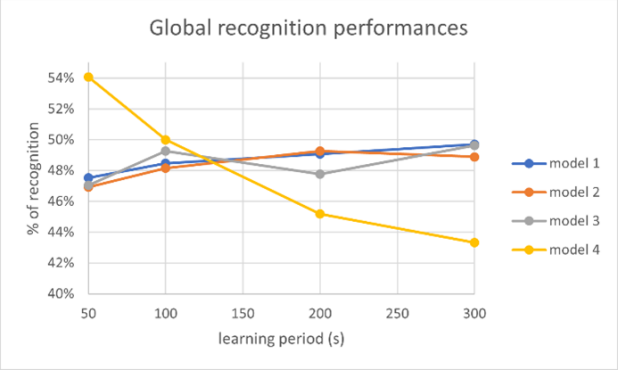
However, the aim is to be able to distinguish the different positions of these muscle groups, and several experiments are being carried out to try to obtain better results.

Three others, somewhat different, programmes have been designed to try to achieve better results:
- Version 2 where the learning is random. At each frame, each LMS has a 1 in 2 chance of learning, so the internal states are not learned at the same time.
- Version 3 in which the output of the SAW during learning contains a winner takes all additionally, setting all other neuron activities to 0 and the most active neuron to 1.
- Version 4 where there is a competition between the LMSs, only the LMS with the highest activity is updated.





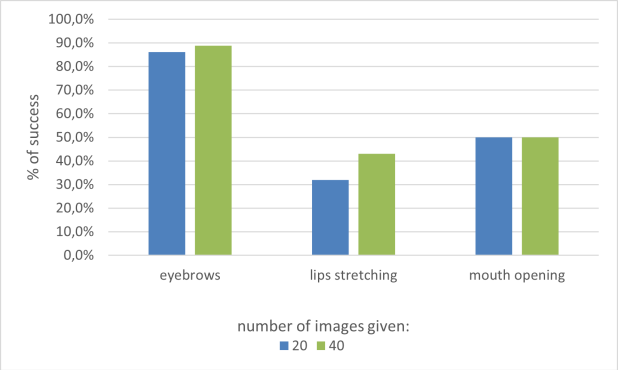




We can note the first model was from the best for a learning period of 100 and 200 seconds.

Because the stretch is not satisfactory, perhaps we are too restrictive with the model, trials removing the muscle grouping for the stretch revealed surprisingly poorer results for mouth distinction with a maximum success percentage of only 35.9% for 300 seconds of training.

A final hypothesis to be tested is that it might be necessary to facilitate the learning of the robot. Indeed, the robot only sees exaggerated images during training, for example for surprise: raised eyebrows and wide-open mouth do not exactly reflect what a human display when surprised. Adding to the database a few images with less pronounced expressions of surprise and training these images on the LMS corresponding to the right muscle group, in order to give the robot a first input, could facilitate the recognition of mixed emotions later on.



We obtained almost 50% success for the lip muscle group with 40 images provided for each emotion, which represented 20% of the images visualised for the experiment.

Beyond that, it would make the task of the robot too easy and short-circuit the notion of emergence that we want to obtain. It is therefore possible to conclude that despite encouraging results, the proposed global model is not flexible enough to allow the robot to distinguish different secondary facial expressions efficiently, and that a new model in the same spirit should be defined.

VIII. Acknowledges

I would like to thank my project supervisor, Mr. Boucenna S., professor and researcher of the neurocybernetics team of the University of Cergy-Pontoise, for his ideas, advice, follow-up and motivation.

References

[1] Sander, S. D. (2008). *Psychology of Emotion*. unige.fr
[2] Turbiaux, M. (2009). *Charles Darwin (1809-1882) and psychology*. Psychology Bulletin, Issue 502
[3] Scherer, K. (2005). *What are emotions ? And how can they be measured?*
[4] GARCIA-PRIETO, P., TRAN, V., WRANIK T., (2005). *Theories of emotional appraisal and differentiation: a key to understanding the emotional experience of individuals at work.*
[5] DANG, T. (2012). *EMOTION AND GRACE - TOWARDS A UNIFIED COMPUTATIONAL MODEL OF EMOTION - APPLICATION TO MUSIC LISTENING IN A DANCING ROBOT.*
[6] Khalfi, F. (2018). *Automatic emotion recognition using multimodal data: facial expressions and physiological signals.*
[7] Velasquez, J. D. (1998). *When Robots Weep : Emotional memories and Decision-Making.*
[8] Laban G., George J-N., Morrison V., Cross E. (2020). *Tell Me More ! Assessing Interactions with Social Robots From Speech.*
[9] Kory-Westlund J M., Breazal C. (2019). *Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning.*
[10] Boucenna, S. (2011). *From facial expression recognition to shared visual perception: a sensory-motor architecture for initiating social referencing of objects, places or behaviors.*
[11] Tian Y., Kanade T., Cohn J. F. (2001). *Recognizing Action Units for Facial Expression Analysis.*
[12] Breazal, C. (2004). *Robot Emotion: A Functional Perspective.*
[13] Breazal, C. (2002). *Emotion and sociable humanoid robots.*
[14] Sosnowski S., Sosnowski S., Bittermann A., Kühnlenz K., Buss M. (2006). *Design and evaluation of emotion-display EDDIE.*
[15] Wu T., Butko N. J., Ruvulo P., Bartlett M. S., Movellan J. R. (2009). *Learning to Make Facial Expressions.*
[16] Poggi I., Pelachaud C., Rosi F de., Carolis B De. (2005). Greta. A *Believable Embodied Conversational Agent.*
[17] Breazal, C. (2000). *Sociable Machines : Expressive Social Exchange Between Humans and Robots.*
[18] Breazal C., Scassellati B. (2000). *Infant-like Social Interactions between a Robot and a Human Caregiver.*
[19] Billard A., Dautenhahn K., Nehaniv C. L. (1999). *Imitation : a means to enhance learning of a synthetic proto-language in an autonomous robot.*
[20] Kanoh M., Iwata S., Kato S., Itoh H. (2005). *EMOTIVE FACIAL EXPRESSIONS OF SENSITIVITY COMMUNICATION ROBOT "IFBOT".*
[21] Berthelon, F. (2014). *Modeling and detecting emotions from expressive and contextual data.*
[22] Becker-Asano C., Wachsmuth I. (2008). *Affect Simulation with Primary and Secondary Emotions.*
[23] Bui T. D., Heylen D., Poel M., Nijholt A. (2001). *Generation of Facial Expressions from Emotion Using a Fuzzy Rule Based System.*
[24] Scherer K., Mortillaro M., Mehu M. (2013). *Understanding the Mechanisms Underlying the Production of Facial Expression of Emotion : A Componential Perspective.*
[25] L. Younes, B. Romaniuk, E. Bittar (2011). *Comprendre et paramétrer l'algorithme SIFT.*
[26] Adrian Rosebrock (2017). *Detect eyes, nose, lips, and jaw with dlib, OpenCV, and Python.*
[27] Adrian Rosebrock (2014). *Histogram of oriented Gradients and Object Detection*
[28] Navneet Dalal and Bill Triggs (2005). *Histograms of Oriented Gradients for Human Detection*
[29] Michel Crucianu, Marin Ferecatu, Nicolas Thome, Nicolas Audebert (2016). *Cours – SVM linéaires*
[30] Vahid Kazemi and Josephine Sullivan (2014). *One Millisecond Face Alignment with an Ensemble of Regression Trees*
[31] Fang-Pen Lin (2020). *py-avataaars – Python component for Avataaars*