First Model Proposal

Identifying Emotional States Considering Upper-body Kinematics

Elena Lechuga Redondo, *P.h.D student at IIT and UNIGE* ***15-01-2019***

Abstract - This document aims to describe the first approach of the three-year P.h.D thesis, which consists in building a system capable of identifying emotional states through the ICub's eyes. Essentially, the main decisions to highlight are: the corporal region to analyse, which will be focused on the upper zone of the body, including head and hands; the input features to extract from the person and send to the classifier, which will be related to the kinematics of the body rather than the body pose; the output emotion that the final system will predict, which will be centered on the level of conformity of the subject while interacting with the robot; the use of an external "detector" to help identifying certain parts of the body, in this case, the "Biological Motion Detector for Human–Robot Interaction" developed by Alessia Vignolo; the machine learning algorithm to train and predict from, which will be decided depending on its performance in Weka, the candidates are: Neural Network (NN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT) and Linear Regression (LR); and the database, which will be created specially for this project.

Additionally, future plans in which the study of the level of conformity will continue are also mentioned. This time, two more models will be implemented by taking into account the distance between the participants and the robot.

The main goal of the three-year thesis is to have a system capable of perceiving the level of conformity of the subjects in a natural interaction. Scenarios in which the faces are not clearly visible, the distance between them and the robot is variable and the interaction does not imply being face-to-face, are all included.

I. INTRODUCTION

Developing social intelligence is key to be integrated in society [10]. Taking into consideration the Human Robot Interaction (HRI) field, building robots with that capability is therefore essential. However, there is a lot of uncertainty regarding the way in which social intelligence works. That is the main reason why psychologists, neuro scientists and engineers focus their research in this topic.

While trying to study human-human or human-robot interaction without the intervention of language, several studies have shown relevant aspects to consider in order to extract the maximum amount of information concerning the emotion, affective state or intention of people. These aspects, also called modalities are: facial expressions [14], body poses [7] or movements [1,2,3,4,5,6], and speech-voice tones [12].

Initially, investigations were individually conducted in any of these modalities. The most demanded arquitectures were the ones analysing facial expressions, as Paul Ekman defined the Action Units (AUs) and the Facial Action Coding System (FACS), which bond perfectly single facial muscles with its associated micro expression and possible emotion [15,16]. Nevertheless, future research has proven that the best results are obtained when combining modalities, which enhanced the accuracy of the recognition task [2,8,9,11,13].

II. LITERATURE REVIEW

In order to approach this challenge in a contrasting manner, the information that will be used to identify and classify the emotional state of the subject, will be mainly obtained from the corporal region. Thus, the literature review will turn around this fact.

A. Body Regions

Assuming that the main region to study will be the body, there are still different places to extract information from.

In the investigation conducted by Glowinski [1], the data was selected from the *head and* the *two hands*. On the other side, Gunes and Piccardi [2] used multimodal data by merging information extracted from the face (*forehead, upper and lower eyebrows, upper and lower eyes, nose, upper right lip, lower right lip, upper left lip, lower left lip, and chin regions*) and the body (*hands, head, shoulders, or combinations of these*). The same as Psaltis [8] who also utilised the face (this

time concentrating the spot only in the *upper and lower regions*) and body.
Additionally, Vaufreydaz, Johal, Combe [9] and Kapur [3] focused on the full body skeletal, the same as Castellano, Villalba and Camurri [5], with the difference that the latter paid special attention to the *hands*. Similarly, Saha [6] gathered information from the *upper-body*.
Practically all of the studies mentioned above, dehumanize the human silhouette paying special attention to the kinematics rather than the human shape.
On the contrary, Kosti [7] focused solely on human figures, picking regions that could represent the *whole body, upper body, face* or even *not clearly visible zones*.

B. Input Features

Some studies, proved that humans can transmit and understand the emotional state of other people just by looking at their movements independently of their shape. Pollic [4], conducted an experiment in order to confirm this fact. In the experiment, 2 subjects were instructed to perform 10 different emotions while drinking and knocking. The subjects wore point-lights displays in the head, right shoulder, elbow, wrist, and the first and fourth metacarpal joints. The function of the point-lights was to transform the original video of the subject, into another in which just the six circles were seen while performing the action. Then, people were asked to classify both videos into one of the 10 possible emotions. The accuracy obtained was 88% for full-video and 63% for point-light displays.
That proves that a lot of information is contained in the movement itself. Hence, this section shows low-level features used in previous papers regarding emotion recognition based on kinematics.

In the case of Glowinski [1], they computed the following expressive features regarding the body regions mentioned before: *energy spent* (which implies operations with the velocity and acceleration); *spatial extent* of the bounding triangle comprehending the defined region; *smoothness/jerkiness* of the curvature; *symmetry* of the hands; and forward-backward *leaning* of the head.

Meanwhile, Gunes and Piccardi [2] considered the variation of the *centroid*, *rotation*, *length*, *width*, and *area* of the detected regions. Along with its *texture* and *quantity of the optical flow* with respect to a neutral frame. Once they acquired the features, they defined specific phases (temporal segments) for the body (*preparation - hold - stroke - hold - retraction*) and the face (*neutral - onset - apex - offset - neutral*) to gather the features of both modalities in a timeline.
Stroke and apex are the temporal segments which show the maximum level of intensity of the emotion performed. In order to merge both modalities, they first identified the frames related to the *stroke* and *apex* of both regions and then, they built two different descriptors (one per modality) containing the features detected in those frames. Consequently, depending on the selected option, they could simply fuse those descriptors into a single one or train individually two classifiers and use some decision-level criterion to decide the bimodal affect.

Kapur researchers [3] opted for simplistic measures, taking into account their interest in representing the dynamics of the motion over larger time scales (10 seconds). Thus, they considered the *mean values of velocity and acceleration* and the *standard deviation* values of position, *velocity* and *acceleration*.

On the contrary, Castellano, Villalba and Camurri [5] defined their own non-propositional movement qualities, which are: the *quantity of motion* of the silhouette, *contraction index* (degree of contraction and expansion of the body), *velocity* and *acceleration* of the hands and general fluidity.

The study carried out by Saha [6] returned to the extraction of basic symmetry measures, they studied the *Euclidean distance* of both hands with respect to the spine, maximum *acceleration* of hand and elbow with respect to spine, *angle* between head, shoulder center and spine, and *angle* between shoulder, elbow and wrist.

Psaltis and his team [8] defined a broad range of features, split into the following categories: kinematics (*kinetic energy, velocity* and *acceleration*), spatial extent (*bounding box, density* and *index of contraction*), smoothness (*curvature* and *smoothness index*), symmetry (*wrists, elbows, knees* and *feet symmetry*), leaning (*forward and backward leaning of a torso* and *head*, as well as *right and left leaning*) and distances (*distances between hands, hand* and *head* as well as *hand and torso*).

Finally, the last of the experiments contemplated, performed by Vaufreydaz, Johal and Combe [9] was the one including the highest number of features. They started by training the classifier with 90 features and empirically, they proved that keeping 32 and discarding the rest, provided the same results. The measurements which were kept, are the following: spatial/proxemic features (*feet detection, speed and acceleration estimation* and *pedestrian tracking considering the space between legs*), acoustic features (*pitch, intensity, speech rate, pitch contours, voice quality* and *silence*) and body features (body pose: *stance for hips, feet, torso and shoulders* and *relative torque angle for hips, torso and shoulders orientation*, and *skeleton distance*; face detection: *orientation of the face toward the robot*).

C. Output Emotion

Regarding the emotional state which can be learnt and predicted by a classifier, there are different ways of approaching.

On the one hand, a finite number of emotions could be previously defined, as the following researchers did: Gunes and Piccardi [2] defined *anger, disgust, fear, happiness, sadness, anxiety, boredom, uncertainty, puzzlement, and neutral/negative/positive surprise*; Kapur [3] described *sadness, joy, anger, and fear*; Saha [6] established *anger, fear, happiness, sadness and relaxation*; and Psaltis [8] distinguished among *anger, fear, happiness, sadness, surprise and neutral* (the frames which information does not match with any of the previous labels are classified as *neutral*).

Additionally, apart from defining a specific number of emotions, every emotion can be constituted by continuous values. For instance, Castellano, Villalba and Camurri [5] considered *anger, joy, pleasure and sadness* in the space of valance (positive and negative) and arousal (high and low). Also, Kosti [7] defined 26 emotional categories together with their continuous dimensions of valence, arousal, and dominance: *peace, affection, steem, anticipation, engagement, confidence, happiness, pleasure, excitement, surprise, sympathy, confusion, disconnection, fatigue, embarrassment, yearning, disapproval, aversion, annoyance, anger, sensitivity, sadness, disquietment, fear and pain suffering*.

On the other hand, there are scenarios in which the possible emotions are not defined. Glowinski [1] played with the valance (pleasantness/unpleasantness dimension) and arousal levels of the emotional states, to cluster them in 4 possible groups. The first cluster (*high positive*) considered videos with high amount of activity and high movement excursion and at the same time, asymmetric and slightly discontinuous type of movements. The second cluster (*high negative*), contemplated videos with also high amount of activity and high movement excursion and at the same time, high score of jerkiness (movement to unfold in a discontinuous way). The third cluster (*low positive*) included videos with low motor activity and a reduced spatio-temporal excursion movements. Last, the fourth cluster (*low negative*) gathered videos with low activity, spatio-temporal excursion and jerkiness in movement execution, yet displaying a symmetry score similar to the one observed.

Alternatively, Vaufreydaz, Johal and Combe [9] instead of approaching different affective states simultaneously, decided to focus on one. The level of *engagement*, which was represented by five classes in the first experiment (*will interact, interact, leave interact, no one, someone around*) and three classes in the second experiment (*will interact, no one, someone around*).

D. External Detector

Investigations are rarely performed without utilizing external libraries or tools that could help solving secondary tasks of the research.

For example, machine learning classifiers are usually built intrinsically by programmers dedicated to that task, and then used by external scientists. The same happens with software which detects specific regions of the image, such as the face, eyes, hands and so on. These tools, are specially demanded by this kind of investigations, as it could considerably save time.

Therefore, some of the papers include external detectors. Particularly, Glowinski [1] used a *skin color tracking algorithm* to extract the blobs of the head and the two hands. Gunes and Piccardi [2] utilised a *CamShift tracker* to extract each region of interest (location of body parts represented as boundary boxes). Kapur [3] did not use any external detector itself, but they fit spheres covered with reflective tape, known as *markers* (14 markers) to record the trajectory of specific parts of the body.

Castellano, Villalba and Camurri [5] included the *EyesWeb platform* to extract the whole silhouette and the hands of the subjects.

Commonly, Saha [6], Psaltis [8] and Vaufreydaz, Johal and Combe [9] made use of an external hardware tool called *Kinect sensor*. It generates a human skeleton represented by 3-dimensional coordinates corresponding to twenty body joints (which are head, shoulder center, spine, hand left, wrist left, elbow left, shoulder left, hand right, wrist right, elbow right and shoulder right) and allows real time pose and gesture recognition. Kosti [7] also included a *dense-ASM tracking framework* which consists in a two-three layer neural network with a hidden layer to detect the action units of the face.

E. Machine Learning Algorithm

In order to test the performance of the architectures proposed by the several researchers introduced in this document, different machine learning algorithms were considered.

Glowinski [1] used an unsupervised learning (clustering) approach. In other words, every new entry (video) was classified into one of the possible clusters, regarding its closeness with the centroid of the cluster computing the Euclidean distance. To define the number of clusters, they applied bootstrapping techniques. The final structure of the model imitated a 4-quadrants graph with 2 axis, one indicating the level of arousal and the other the level of valence of the emotion perceived.

On the contrary, Gunes and Piccardi [2] employed frame-based and sequence-based classifiers. The frame-based classifiers were tested in Weka, obtaining the following results: BayesNet (28.97% face-monomodal, 73.2% body-monomodal and 72.73% bimodal); Support Vector Machine with Sequential Minimal Optimization (SVM-SMO) (32.49% face-monomodal, 64.51% body-monomodal); Random Forest (RF)(33.56% face-monomodal, **76.87% body-monomodal** and 80.72% bimodal); Adaboost using C4.5 (**35,22% face-monomodal**, 73.14% body-monomodal); Adaboost with Random Forest (**82.65% bimodal**); and Neural Networks (NN) (80.27%). As sequence-based classifiers they utilized a Hidden Markov Model (HMM) obtaining 11% and 12.6% accuracy for the face and body in the monomodal classifier, and **17.3% in the bimodal classifier**.

In addition, Kapur [3] also tested the next frame-based classifiers: Logistic regression (85.6 %); Naive bayes (NB) with a single multidimensional Gaussian distribution modeling each class (66.2 %); Decision tree classifier based on the C4.5 algorithm (86.4 %); Multilayer perceptron backpropagation artificial neural network (91.2 %); And support vector machine using the Sequential Minimal Optimization (SVM-SMO) (**91.8%**).

Similarly, Castellano, Villalba and Camurri [5] used Weka to test the performance of their model against these classifiers: Simple 1--Nearest-Neighbor (1NN); Decision Tree - J48; And Bayesian Network/Hidden Naive Bayes (HNB).

The experiments headed by Saha [6] obtained the following percentages: Binary decision tree (76.63%); Ensemble decision tree **(90.83%)**; k-nearest neighbour (KNN) (86.77%); Support vector machine with radial basis function kernel (87.74%); and neural network based on backpropagation learning (89.26%).

Vaufreydaz, Johal and Combe [9] tested their database using Neural Network and Support Vector Machine algorithms. The percentages of correctly predicting the label “will interact” for NN were 0,90 precision and 0,87 recall, while for SVM were 0,92 precision and 0,71 recall.

On the contrary, the Kosti [7] approach was based on deep learning. Specifically, it consisted in a Convolutional Neural Network (CNN) model which gathered three independent modules. The first module centered its attention to the zone of the subject with most expressiveness, trying that way to obtain the most relevant features. The other module extracted features from the global image in order to take into account the context. The last module merged the first two, associating each image to its label and its level of VAD (valance, arousal, dominance). Different configurations of the CNN parameter were considered for the experiments. Presenting the best results when fusing the whole-image features along body features using the loss function L comb (obtaining 28% average precision).

Following that path, Psaltis [8] also created a multimodal fusion Neural Network, training independently the facial and corporal modalities,

merging them afterwards (they obtained 98.3% of accuracy).

F. Database

The database is a crucial factor affecting the performance of systems based on machine learning, as the system will learn what you have previously taught it. That could be the main reason why almost every researcher decided on implementing themselves their own corpus.

Nonetheless, Glowinski [1] adapted the Geneva multimodal emotion portrayals (GEMP) selecting a subset of 120 scenes representing 12 emotions (mentioned in the Output Emotion section) by 10 actors. The data has been obtained through a 25 frames-per-second video camera and validated by extensive ratings that ensured high believability (assessed capacity of the actor to communicate a natural emotion impression), reliability (interrater agreement) and recognizability (accuracy scores) of the encoded emotion.

Considering the time Gunes and Piccardi [2] started their experiments, there was not readily available a database which combined affective face and body information in a bimodal manner. Hence, they created the FABO database by collecting information from two cameras. The left camera focused on facial gestures, recording only the face of the subjects. And, the right camera fixated on corporal gestures, recording both the upper body and the internal region of the face. It is significant to mention that the subjects did not know which emotions they had to perform. Instead, they were presented specific situations and they had to act out accordingly.

Kapur's team [3] also built their own corpus, constituted by 500 raw data files storing the x, y, and z coordinates extracted from the markers. The data was collected from 5 subjects, where each subject performed 25 times each emotion for a length of 10 seconds.

On the other hand, Castellano, Villalba and Camurri [5] used a corpus created by 240 gestures collected during the Third Summer School of the HUMAINE (Human-Machine Interaction Network on Emotion) in Genoa 2006. The data was acquired from 10 participants (six male and four female) that were asked to act eight emotional states (anger, despair, interest, pleasure, sadness, irritation, joy and pride) equally distributed in the valence-arousal space. Even though, just 4 emotions (the ones mentioned in the Output Emotion section) were used. In order to perform the emotions, the subjects were ask to stand in the rest position while rising and lowering the arms in the coronal plane. The gesture was performed 3 times per emotion. The full body was recorded using a camera capable of recording 25 frames per second. The background was dark and all the participants wore a long-sleeve shirt to make feasible the hand tracking.

Saha [6], Psaltis [8] and Vaufreydaz, Johal and Combe [9] created their corpus using the Kinect sensor, a device which records videos at a rate of 30 frames per second.
In the scenario of Saha [6], 10 subjects around 25+-5 years old were asked to express the emotion/s they were feeling with their body. To motivate the subjects to perform the 5 emotions mentioned before, the conditions of the room were altered. Every sample lasts 60 seconds.

On the contrary, the samples obtained by Psaltis [8] were significantly shorter, during 3 seconds per video. The whole database gathered 450 videos, obtained from 15 subjects starting from the neutral expression. To collect the displays, a video was shown to every subject illustrating each of the emotions and then, the subject had to perform the same emotion 5 times with its own style in front of the Kinect sensor. The majority of the videos recorded the full body of the subject. Nevertheless, a bit of noise was added by including some videos showing the facial or the corporal region independently.

The corpus used by Vaufreydaz, Johal and Combe [9] included 29 videos produced by 15 different participants (50% were male and 50% were female) from 20 to 25 years old. The recording setup was meant to extract the most realistic data. For that reason, the participants were left in a "living lab" environment similar to a flat, with the purpose of creating emergent interactions. Even though, some actions were randomly given to each one of the participants to stimulate the HRI, like walking, sitting, playing cards or pouring water from the sink. Considering all the data, not all the classes were equally distributed ("someone around" 22.3%, "will interact" 1.47%, "interact" 24.94%, "leave interact" 1.29%).

The last dataset was created by Kosti [7], called "Emotions in Context Database" (EMOTIC). It was composed of images from MSCOCO and Ade20k datasets along with images downloaded using the Google search engine. In order to label these images with its respective category or continuous dimension, workers were previously trained and examined.

Every image was annotated with two complementary systems: the affective category representing the image (the label indicating the emotion), and three numbers (in the scale 1 to 10) representing its level of valence (that measures how positive or pleasant an emotion is, ranging from negative to positive), arousal (that measures the agitation level of the person, ranging from non-active / in calm to agitated / ready to act), and dominance (the control level of the situation by the person, ranging from submissive/non-control to dominant/ in-control).
The total number of images of the corpus was 18,316. It is important to highlight that an elevate percentage of images did not present the face in a clearly visible angle.

III. PROPOSED METHODOLOGY

Subsequently, I have tried to choose the best combinations of choices in order to build an outstanding system, considering the papers introduced before.

A. Body Regions

This first model proposes a system which will be focused on the upper-body kinematics to extract useful information regarding the emotional state of the person.

The idea is to represent the upper-body of the person through blobs situated in the next areas: *head, shoulders, upper-trunk, elbows, hands and hip*.

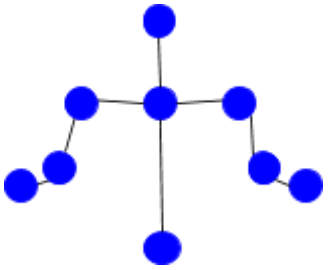

Fig.1. Graphical representation of the blobs which will be analysed by the proposed model.

Additionally, if the quality of the camera allows to extract information of the internal region of the face (situating the participant at a medium distance from the robot), a multimodal system fusing facial and upper-body information will be implemented.
In this case, also features from *the forehead, eyes, nose, upper right lip, lower right lip, upper left lip, lower left lip, and chin* will be included.

Fig.2. Graphical representation of the blobs of the internal region of the face that would be analysed by the proposed model if the quality of the camera allows it.

B. Input Features

Once the body regions have been decided, its external shape will be ignored keeping only information related to its dynamics.

Precisely, the features representing the descriptor of every individual video will be grouped into the following categories:

- Energy (*velocity and acceleration*).
- Spatial extent (*bounding, density and index of contraction*).
- Smoothness (*curvature and fluidity*).
- Symmetry (*symmetry between hands, elbows and shoulders*).
- Leaning angle (*considering leaning of the head, of the trunk and both together*).

C. Output Emotion

Taking into account that this document proposes the first approach of a P.h.D thesis, it has been decided to explore deeply every aspect of the possible emotional states that a human could feel.

Although there is not a finite number of emotions, several psychologist and scientists have tried to defined some standards. As mentioned before, Paul Ekman [15] discovered that every person could transmit and perceived six basic emotions (sadness, anger, happiness, fear, surprise and disgust) offering a discrete categorization of them. Additionally, Plutchik

[17] introduced a hierarchical view of emotions, called "the Wheel of Emotions", conceiving them as infinite processes of feedback loops. In contrast, some researchers opted for representing the emotions in the multidimensional Euclidean space, being one of the most relevant the Valence-Arousal-Dominance space (VAD) [18]. It represents every emotional state through three numbers. The first represents the positive-negative value, the second the level of intensity and the third the level of control which the person has over the emitted emotion.

In order to decide the method to imitate, the HRI field has been considered, asking which emotion/s would be the most useful for a robot. Trying to obtain the maximum veracity in real situations, the idea of utilizing a discrete categorization of several emotions has been discarded. Instead, one single emotion will be analysed, showing its continuous level of arousal.
That is to say, the emotional state to study will be the level of conformity of a person in an interaction. It will go from zero (meaning that the person is feeling rejection/disgust) to 100 (meaning complete satisfaction/love) percent of conformity. Internally, 25% will mean partial disagreement, while 50% will indicate a neutral state, where the person won't show any sign of agreement or disagreement and 75% will mean partial agreement.

D. External Detector

In order to identify the blobs explained in the Body Regions section, the "Biological Motion detector" developed by Alessia Vignolo [19] would be detailed studied and considered.

E. Machine Learning Algorithm

Following the example of the papers studied, several state-of-the-art algorithms will be tested in Weka, and afterwards, the one with highest accuracy will be implemented into the ICub.

The candidate algorithms are:

- Support Vector Machine (SVM)
- Simple Neural Network (XNN)
- Decision tree - J48
- Naive Bayes
- Logistic Regression

F. Database

Given it does not present special difficulties and the output emotion to consider is rather exceptional, the corpus will be exclusively built for this project.
To do so, 300 videos will be acquired from 20 participants. Every display will contain the reaction to a story of one participant, in which they will have to simulate being the protagonist. As explained in the Output Emotion section before, the continuous level of conformity would be split into five levels. At the same time, every level will contain three associated stories. That way, every participant will be asked to react to 15 stories, during from 5 to 15 seconds, obtaining a wide range sample of conformity states.
The camera which will be used to record the videos will be the one integrated into the ICub robot, as the intention is to train the system with the most realistic data. Doing so, the final system should be capable of predicting unseen information in real HRI-scenarios.

IV. GOAL & FUTURE PLANS

The main intention with this project is to implement a light system which identifies the level of conformity of every subject interacting with ICub.

A very simple view, as the one shown below, could be added to yarp dataplayer so everyone could quickly understand the perception of ICub against this affective state in every moment.

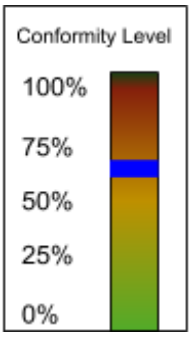

Fig.3. Measurement bar regarding the level of conformity of possible participants in future experiments. The blue line will indicate the exact percentage of conformity.

Once this functionality is working, a personality profile will be given to ICub, to make it capable of answering to the level of conformity perceived. Initially, this personality will try to maximize the level of conformity of the human. Thus, ICub will respond accordingly with a physical manifestation of its decision. In other words, several options will be previously detailed for example, smiling plus moving its arms trying to catch

the attention of the human, producing the surprise emotion by moving its facial leds and arms or emitting an annoying gesture trying to express indignation for being ignored. This prefixed reactions will be studied when the time comes. The idea is to create a social interaction between ICub and the participants considering non verbal information.

Subsequently, the other two modules regarding the distance (*facial-model* when the distance will be closer and *whole-body-model* when the distance will be further) between the robot and the subject will be proposed.

V. REFERENCES

1. *D. Glowinski, M. Mortillaro, K. Scherer, N. Dael, G. V. A. Ca- murri, Towards a minimal representation of affective gestures, in: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 498–504.*
2. *GUNES, Hatice; PICCARDI, Massimo. Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, vol. 39, no 1, p. 64-84.*
3. *KAPUR, Asha, et al. Gesture-based affective computing on motion capture data. En International conference on affective computing and intelligent interaction. Springer, Berlin, Heidelberg, 2005. p. 1-7IEEE International Conference on. IEEE, 2016. p. 435-439.*
4. *VAUFREYDAZ, Dominique; JOHAL, Wafa; COMBE, Claudine. Starting engagement detection towards a companion robot using multimodal features. Robotics and Autonomous Systems, 2016, vol. 75, p. 4-16.*
5. *GOLEMAN, Daniel. Inteligencia social: la nueva ciencia de las relaciones humanas. Editorial Kairós, 2006.*
6. *ATREY, Pradeep K., et al. Multimodal fusion for multimedia analysis: a survey. Multimedia systems, 2010, vol. 16, no 6, p. 345-379..*
7. *POLLICK, Frank E., et al. Perceiving affect from arm movement. Cognition, 2001, vol. 82, no 2, p. B51-B61.*
8. *CASTELLANO, Ginevra; VILLALBA, Santiago D.; CAMURRI, Antonio. Recognising human emotions from body movement and gesture dynamics. En International Conference on Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg, 2007. p. 71-82.*
9. *SAHA, Sriparna, et al. A study on emotion recognition from body gestures using Kinect sensor. En 2014 International Conference on Communication and Signal Processing. IEEE, 2014. p. 056-060.*
10. *KOSTI, Ronak, et al. Emotion recognition in context. En The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.*
11. *PSALTIS, Athanasios, et al. Multimodal affective state recognition in serious games applications. En Imaging Systems and Techniques (IST), 2016.*
12. *YANG, Zhaojun; NARAYANAN, Shrikanth S. Analysis of emotional effect on speech-body gesture interplay. En Fifteenth Annual Conference of the International Speech Communication Association. 2014.*
13. *KESSOUS, Loic; CASTELLANO, Ginevra; CARIDAKIS, George. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 2010, vol. 3, no 1-2, p. 33-48.*
14. *LOPES, André Teixeira, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognition, 2017, vol. 61, p. 610-628.*
15. *Ekman, P., 1978. Facial Expression 1.*
16. *EKMAN, Paul. Facial action coding system (FACS). A human face, 2002.*
17. *17. R. Plutchik, "Emotion: Theory, research, and experience" vol. 1. Theories of emotion 1, New York: Academic, 1980.*
18. *18. J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," J. Res. Personality, vol. 11, no. 3, pp. 273–294, 1977.*
19. *VIGNOLO, Alessia, et al. Detecting biological motion for human–robot interaction: A link between perception and action. Frontiers in Robotics and AI, 2017, vol. 4, p. 14.*