Affective Facial Expressions Recognition for Human-Robot Interaction

Conference Paper · August 2017

CITATIONS
0

READS
38

4 authors, including:

Diego R. Faria
Aston University
37 PUBLICATIONS **171** CITATIONS
SEE PROFILE

Mário Vieira
University of Coimbra
5 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Cristiano Premebida
University of Coimbra
33 PUBLICATIONS **528** CITATIONS
SEE PROFILE

Affective Facial Expressions Recognition for Human-Robot Interaction

Diego R. Faria[1], Mario Vieira[2], Fernanda C.C. Faria[2] and Cristiano Premebida[2]

Abstract— **Affective facial expression is a key feature of non-verbal behaviour and is considered as a symptom of an internal emotional state. Emotion recognition plays an important role in social communication: human-to-human and also for human-to-robot. Taking this as inspiration, this work aims at the development of a framework able to recognise human emotions through facial expression for human-robot interaction. Features based on facial landmarks distances and angles are extracted to feed a dynamic probabilistic classification framework. The public online dataset Karolinska Directed Emotional Faces (KDEF) [1] is used to learn seven different emotions (e.g. angry, fearful, disgusted, happy, sad, surprised, and neutral) performed by seventy subjects. A new dataset was created in order to record stimulated affect while participants watched video sessions to awaken their emotions, different of the KDEF dataset where participants are actors (i.e. performing expressions when asked to). Offline and on-the-fly tests were carried out: leave-one-out cross validation tests on datasets and on-the-fly tests with human-robot interactions. Results show that the proposed framework can correctly recognise human facial expressions with potential to be used in human-robot interaction scenarios.**

Index Terms— **Affective facial expressions, emotion recognition, human-robot interaction**

I. Introduction

In recent years, there has been a growing interest in recognising people's affective state which launched the interest in developing new technologies. One good example is towards Human-Robot Interaction (HRI), where automatic recognition of human emotions and generation of expressive behaviour for virtual avatars and robots are key challenges. When it comes to emotional expression in person-to-person communication, the face is one of the main focus of attention [2] since it transmits different information about emotions. There are evidences that facial expression displays human characteristics regarding expressiveness [3], [4], which is a meaningful way for social interaction between humans and robots. On the other hand, affect expression can occurs through combinations of verbal and nonverbal communication channels including bodily expressions [5], however, it is evident that the study of perception of whole-body expressions lags so far behind facial expressions [6].

Research on human emotion is mainly focused on facial expressiveness, usually defined by the Facial Action Coding System (FACS) [7]. It was developed to measure facial activity called "facial actions", i.e. component motion, to provide an objective description of facial signals. Studies adopting FACS revealed that humans share seven emotional expressions regardless of ethnic group, culture, and country; they are: happiness, sadness, anger, fear, surprise, disgust, and contempt. Although FACS is the leading method for measuring facial expressions, there are studies that considers facial muscles to express twenty one categories of emotions [8] derived from the seven ones aforementioned. For many years facial expression recognition has been an active topic of research, and it is still challenging, especially for real-time applications. Feature extraction for face detection, as in [9] and [10]) is the initial and important step in the recognition process, and usually methods can be divided into geometrical computed from landmarks positions of essential parts of the face [11], and appearance-based methods that work directly on image and not on single extracted points [12]. Usually, the latter uses a larger amount of data, making them less suitable for a real-time application due to the high computational cost.

In this work, we present an emotion recognition approach for HRI that uses the Dynamic Bayesian Mixture Model (DBMM [13]]) and feature-based machine learning classifiers to detect and recognize affective facial expressions. The main contributions of this paper are: (i) effective geometrical and log-covariance features; (ii) a novel uncertainty measure useful to compute global weights and runtime weights update for the DBMM; (iii) assessment of the proposed features using different state-of-the-art classifiers to validate their effectiveness over two affective facial expression datasets. In addition, tests on-the-fly using a humanoid robot during human-robot interaction were carried out, where appropriate robot reactions are executed given the emotional state of a person. Figure 1 depicts an illustration of the general idea of this work when its come to the human-robot interaction.

The remainder of this paper is organized as follows. Section II presents the proposed features extraction approach, the setup and datasets. Section III presents the classification model and the contribution for weights computation. Section IV presents the on-the-fly performance of the proposed approach. Finally, Section V presents the conclusion and future work.

II. Setup, Dataset and Features Extraction

We had set up an environment to stimulate the emotions of the participants for recording the sessions. In order to awaken emotions such as happy/joy, angry, disgusting, afraid/scared, surprised, sad and neutral, we asked participants to watch

Diego R. Faria is with [1]System Analytics Research Institute, School of Engineering and Applied Science, Aston University, Birmingham, UK. M. Vieira, F. Faria and C. Premebida are with [2]Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Portugal. (emails: d.faria@aston.ac.uk, {mvieira, cpremebida, fernanda}@isr.uc.pt).

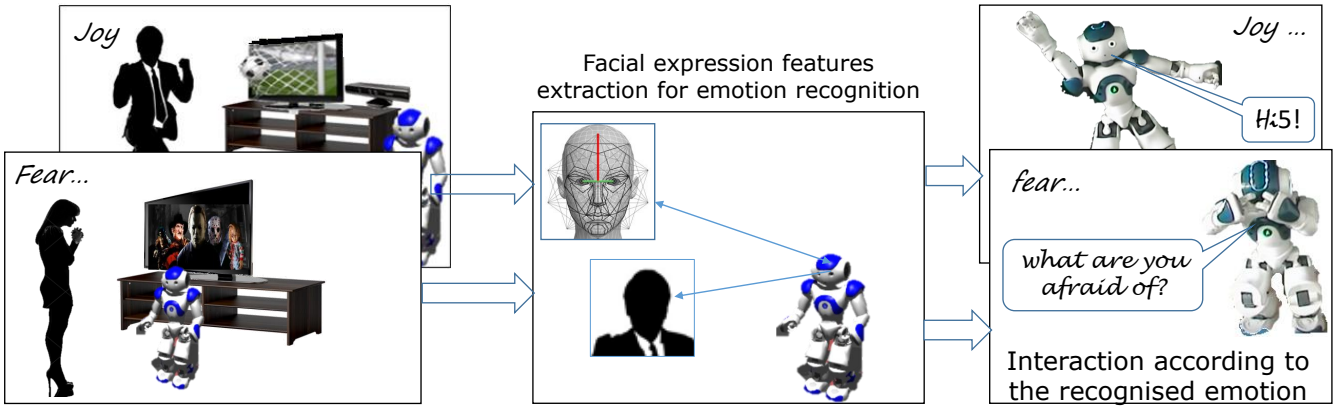

Fig. 1: Illustration of a robot using its monocular camera to detect and recognize facial expressions in order to infer human emotions and then to react and interact according to the person's emotional state.

a video sequence. This sequence is online[1] (i.e. youtube channels) and consists of emotional advert, jokes and pranks that are expected to make people smile, laugh or even make people surprised or scared. We carefully selected successful videos with thousands of views on youtube channels, also following suggestions of a psychologist, in order to be aware of what kind of videos could potentially stimulate the participants' emotions while they watch the videos.

The experimental setup comprises a 50" tv screen to display the selected videos, and a monocular camera (Sony video camera) to record the facial expressions, and an RGB-D sensor to track the body motion. Nevertheless, in this work we are not dealing with bodily expressions. Figure 2 shows the experimental setup for sessions recording of participants' facial expressions.

Six individuals, three males and three females between 22 and 38 years old, with average of 29.6 ± 6.23 participated voluntarily. A dataset of emotions was built, consisting of RGB images with facial expressions and 3D skeleton data (i.e body joints) from the RGB-D sensor along the 17 minutes of sequence of different videos as stimuli, which gives a total of about 30600 frames (1020 seconds $\times$ 30 frames per second) for both, images and skeleton data. We have asked participants to stand at a distance of 2m far from the tv screen, and we have told them to be comfortable and watch the video sequences in the most natural way as possible. We have manually annotated the sequence of videos with 17 minutes of expressions for each participant and also a ground truth (i.e. expected reactions given the videos segments). Through this step we have noticed that the video sequence we have chosen correctly stimulated the expected emotions, with women being more emotional. From these sessions, looking at the annotated data, this dataset was sufficient to detect and properly discriminate most of facial expressions as shown in Figure 3.

An additional publicly available dataset, Karolinska Directed Emotional Faces (KDEF) [1], was also used in this work in order to improve the emotion learning step. The KDEF dataset comprises 4900 pictures of 7 different facial expressions (e.g. Angry, Fearful, Disgusted, Happy, Sad, Surprised, and Neutral) that were performed by 70 subjects, 35 females and 35 males between 20 and 30 years old. The KDEF dataset images have 562×762 pixels and were taken from 5 different angles $(-90, -45, 0, +45, +90$ degrees), however we did not use images with expression performed at 90 and -90 degrees relative to the camera since these images are not good enough for detecting facial landmarks and geometrical features extraction.

[1]Video sequence available at: `https://youtu.be/2-kFYHHJLZQ`

A. Proposed Image-based Facial Features

In this work, given a single image with a human face, several sets of geometrical features are extracted. First, 68 facial landmarks (e.g. contour of the face, lips, eyes and nose) are detected using the Dlib library [14]. Given that, we have computed several subsets of geometrical features as follows:

- Subset $S1$: Euclidean distances among the face landmarks, obtaining a 68×68 symmetric matrix with a null diagonal. Let $\{L_p, L_q\}$ be two facial landmarks with 2D coordinates $\{x, y\}$, then, Euclidean distances $\delta(L_p, L_q) = \sqrt{(L_p^x - L_q^x)^2 + (L_p^y - L_q^y)^2}$ were computed $\forall \{L_p, L_q\}$. Subsequently, we removed the null diagonal, obtaining a 67×68 matrix $\mathbf{M}$ and a normalization was performed: $\mathbf{M} = \frac{\mathbf{M}}{\max(\mathbf{M})}$. This step is important, since it makes the features scale-invariant (i.e. the size of the image or distance of person and camera). Finally, we compute the *log-covariance* of the matrix $\mathbf{M}$ as follows:

$$\mathbf{M}_{lc} = \mathbf{U}(\log(\mathrm{cov}(\mathbf{M}))), \tag{1}$$

  where the covariance for each element in $\mathbf{M}$ is given by $cov_{ij} = \mathrm{cov}(\mathbf{M}) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{M}^{ik} - \mu_i)(\mathbf{M}^{kj} - \mu_j)$; $\log(\cdot)$ is the matrix logarithm function (logm) and $\mathrm{U}(\cdot)$ returns the upper triangular matrix elements composed of 2346 features.
- Subset $S2$: Composed of the same steps of $S1$, but with

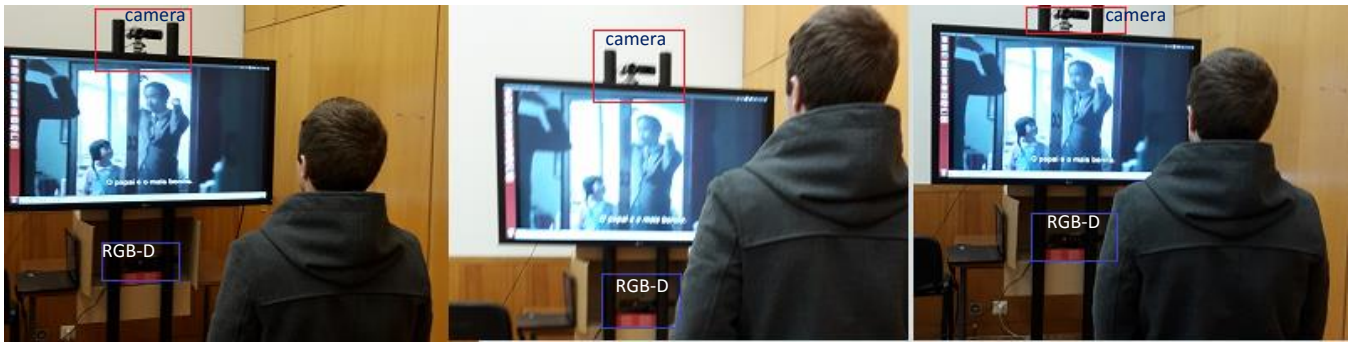

Fig. 2: Experimental setup of the recording sessions for affective facial expressions.

Euclidean distances for each dimension *(x,y)* individually, obtaining a set with 4692 features.

- Subset $S3$: Subset $S1$ without the *log-covariance* step, obtaining 2346 features.
- Subset $S4$: Subset $S2$ without the *log-covariance* step, obtaining a subset of 4692 features.
- Subset $S5$: Given the detected landmarks L, triangles between them are computed. All three angles of a total of 91 triangles (e.g. as shown on the faces in Figure 8) were computed. The angles θ_1, θ_2 and θ_3 are given by:

$$\theta_a = \arccos\left(\frac{(\delta_1)^2+(\delta_2)^2-(\delta_3)^2}{2\times\delta_4\times\delta_5}\right), \qquad (2)$$

where $a=\{1,2,3\}$, i.e., (2) is computed for θ_1, θ_2, and θ_3, while δ_1 to δ_5 are Euclidean distances between two landmarks of a triangle for each θ_a. Each triangle has three landmarks, L_1, L_2 and L_3. For θ_1, $\delta_1=\delta_{L_{12}}$, $\delta_2=\delta_{L_{23}}$, $\delta_3=\delta_{L_{13}}$, $\delta_4=\delta_{L_{12}}$, $\delta_5=\delta_{L_{23}}$. For θ_2, $\delta_1=\delta_{L_{13}}$, $\delta_2=\delta_{L_{23}}$, $\delta_3=\delta_{L_{12}}$, $\delta_4=\delta_{L_{13}}$, $\delta_5=\delta_{L_{23}}$. For θ_3, $\delta_1=\delta_{L_{12}}$, $\delta_2=\delta_{L_{13}}$, $\delta_3=\delta_{L_{23}}$, $\delta_4=\delta_{L_{12}}$, $\delta_5=\delta_{L_{13}}$. In total, we obtained $3\times 91=273$ angle features.

Even though we have a large set of features, it does not demand complex computational cost, being feasible for on-the-fly applications (e.g. around 10-15 frames/second) with considerable performance as shown in results section.

III. Recognition Model

A. Background

A probabilistic ensemble of classifiers called Dynamic Bayesian Mixture Model (DBMM) [13] [15], which was used in different classification applications (e.g. human daily activity recognition [16]; semantic place categorization [17], [18]; and social behaviour classification [19]) is employed in this work for facial expression recognition. The DBMM takes inspiration on Dynamic Bayesian Networks (DBN), using the concept of mixture model to fuse different classifier posteriors, incorporating temporal information through time slices. A random variable A (e.g. feature model for a specific classifier) is considered to be independent on previous A-nodes: $P(A^t|A^{t-1},C^t,C^{t-1})=P(A^t|C^t,C^{t-1})$, where C represents a set of possible classes (e.g. emotions). The nodes are not conditionally dependent of future nodes $P(A^{t-2}|C^t,C^{t-1},C^{t-2})=P(A^{t-2}|C^{t-2})$. With that, the transitions between classes reduces to the probability of the current-time class $P(C^t)=P(C^t|C^{t-1})$. Modeling $P(A^t|C^t)$ by a mixture of probabilities, then the explicit expression for the DBMM with finite $T=\{1,2,...,x\}$ time slices assumes the form:

$$P(C^t|C^{t-1:t-T},A^{t:t-T})=\frac{\prod_{k=t}^{t-T}\left(\sum_{i=1}^{n}w_i^k\times P_i(A^k|C^k)\right)\times P(C^k)}{\sum_{j=1}^{nc}\left[\prod_{k=t}^{t-T}\left(\sum_{i=1}^{n}w_i^k\times P_{i,j}(A^k|C^k)\right)\times P_j(C^k)\right]}, \qquad (3)$$

where n is number of classifiers; nc is the number of classes; w is the weight for each base classifier learned from the training set. In this work, (3) can be simplified for one time slice $T=1$ as follows:

$$P(C^t|A^t)=\frac{P(C^t)\times\left(\sum_{i=1}^{n}w_i^t\times P_i(A^t|C^t)\right)}{\sum_{j=1}^{nc}P_j(C^t)\times\left(\sum_{i=1}^{n}w_i^t\times P_{i,j}(A^t|C^t)\right)}, \qquad (4)$$

where $P(C^t|A^t)$ is the posterior probability; the prior assumes the form $\forall t>1, P(C^t)=P(C^t|C^{t-1})$, otherwise, $t=1$, $P(C^t)=1/nc$ (uniform); $P_i(A^t|C^t)$ is the likelihood model in the DBMM as the posterior probability of a base classifier; and the mixture model is obtained by $mix=w_i^t\times P_i(A^t|C^t)$. Each weight w_i, $i=\{1,2,...,N\}$ can be learnt using different weighting strategies [13].

B. Proposed Weighting Strategy for Merging Classifiers

There are various techniques one can use to estimate a finite set of weights to combine classifiers in an ensemble model. In this work a probabilistic weighting strategy is defined for both, global weights that is acquired from the training set and also for runtime weights update during the on-the-fly tests. The latter is to take advantage of classifiers' behaviours on the test set in order to guarantee better performance.

Probability Residuals Energy (PRE-based weighting) is an approach to quantify the uncertainty of the classifiers as a confidence level given a set of posteriors ranging from different time instants $\{t_1,...,t_S\}$. Let $P(C_k^t|A)$, $k=\{1,...,nc\}$ be the probability of each class during the classification coming from each i^{th} base classifier over time $\{t_1,t_2,...,t_S\}$ that can be inserted in a matrix mix_i, where the rows represent the number of classes, and the columns are the set of posteriors over time as follows:

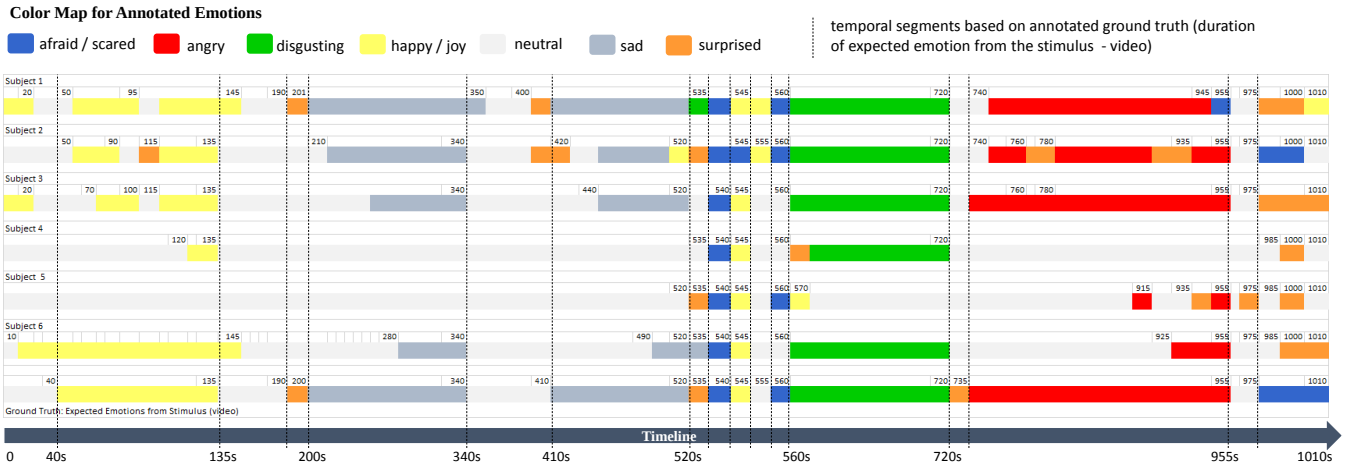

Fig. 3: Annotated data from recordings: groundtruth (GT) and participants' expressions while watching the video sequence as prepared to awaken emotions. Temporal segments are based on GT.

$$mix_i = \begin{bmatrix} P(C_1^{t_1}|A) & P(C_2^{t_1}|A) & \cdots & P(C_{nc}^{t_1}|A) \\ P(C_1^{t_2}|A) & P(C_2^{t_2}|A) & \cdots & P(C_{nc}^{t_2}|A) \\ \vdots & \vdots & \ddots & \vdots \\ P(C_1^{t_S}|A) & P(C_2^{t_S}|A) & \cdots & P(C_{nc}^{t_S}|A) \end{bmatrix}. \quad (5)$$

Given that, it is needed to know which class is the predominant one for each base classifier, i.e. the one classified with the maximum probability. In order to compute the confidence of each classifier model at time instant t, a matrix of previous posteriors as show in (5) is built. It is assumed the classified class is the column with higher energy. To do so, the following steps are recognized:

$$ev\{C_k\} = \|\vec{v}_{C_k^{\{t_1,\ldots,t_S\}}}\|_1 = \sum_{j=1}^{S} \left(v_{C_k}^{t_j}\right)^2, \quad (6)$$

$$\text{with} \quad \vec{v}_{C_k} = \{P(C_k^{t_1}|A), \ldots, P(C_k^{t_S}|A)\},$$

$$C_{MAP} = \text{argmax}\ ev\{C_k\}, \quad (7)$$

where $ev\{c_k\}$ represents the energy of each k^{th} class; $\vec{v}_{c_k}$ is a vector with a set of posteriors for each class, i.e. each column of the matrix (5); and C_{MAP} is the class that is chosen as the most probable one among all possible classes. These steps are done for each i^{th} classifier. Once the most probable class for each classifier model is known (i.e. the column of mix_i with higher energy), then the set of posteriors obtained for the chosen class over time is used to quantify the uncertainty of each model, assigning a confidence through weights to merge the different classifier posteriors into the DBMM by employing the PRE-based weighting. This step consists of computing the sum of the square differences of posteriors between the chosen class C_{MAP}^t and the other remaining less probable classes C_k^t, $k = \{1,2,\ldots,nc\}$, accumulating it over time as an energy model for each base classifier, followed by a normalization:

$$w_{i_{pre}} = \frac{\sum_{t=1}^{S}\left(\sum_{k=1}^{nc}(C_{i_{MAP}}^t - C_{i,k}^t)\right)^2}{\sum_{i=1}^{N}\left[\sum_{t=1}^{S}\left(\sum_{k=1}^{nc}(C_{k,i_{MAP},}^t - C_{k,i}^t)\right)^2\right]}, \quad (8)$$

where i is an index indicating each model up to N classifiers; t is the time instant in a set of S frames; k is an index indicating the posterior of a specific class C_k, $k = \{1,\ldots,nc\}$ from a mixture (e.g. $C_{k,i}^t = P_i(C_k^t|A)$); and $C_{i_{MAP}}^t$ is the maximum posterior $P(C|A)$ at time instant t among all C_k that was found out from (7).

Note that, a solution to find the predominant class is given through (5)-(6), which is useful for weights update in runtime when there is no labeled data, e.g. during the test set classification. When dealing with the training set and labeled data, i.e., knowing the correct class, it is employed only (8) to compute the PRE-based weighting, where C_{MAP} is the posterior acquired for the correct class given by a classifier according to the labeled data.

IV. Experimental Results

Tests offline on two datasets, and on-the-fly experiments using a humanoid robot during human-robot interaction were carried out. The strategy adopted to assess the proposed approach is Leave-One-Out Cross-Validation (LOOCV). The purpose is to verify the capacity of generalization of all classifiers, so that the strategy of "new person", i.e., learning from different persons and testing with an "unseen" person is adopted. The proposed approach was compared against other well known classifiers in literature, such as Support Vector Machines (SVM), linear regression using Stochastic Average Gradient (SAG), Naive Bayes (NB), K-Nearest Neighbors (KNN) and a Random Forest Classifier (RFC).

Extensive tests were carried out using the offline strategy to find out the best classifiers and set of features. With LOOCV on KDEF dataset, 70 offline tests were carried out for each of the 6 classifiers, using 5 different set of

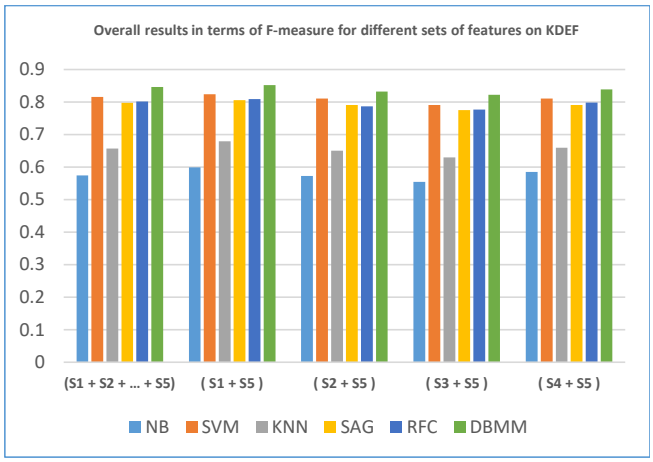

Fig. 4: Overall results in terms of F-measure for all subsets of features on KDEF dataset.

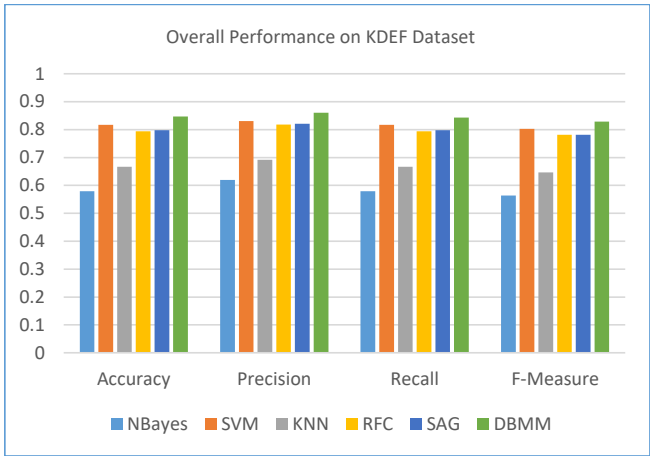

Fig. 5: Results for the KDEF dataset in terms of accuracy, precision, recall and F-Measure. Average of all 70 tests (leave-one-out cross-validation on unseen person).

features, resulting in $70 \times 6 \times 5 = 2100$ tests. For each image in the dataset, the features were computed relative to the correspondent neutral face. In this work, the DBMM consists of the following base classifiers: SVM, SAG and RFC.

Figure 4 presents the overall result (average of all 70 tests on KDEF dataset) in terms of F-measure for all subsets of features. The subset {S1+S5} was the one with slight better performance between all subsets. Figure 5 shows the overall result (average of 70 leave-one-out cross-validation tests) obtained on the KDEF dataset in terms of accuracy, precision, recall and F-measure. Figure 6 presents all tests done on the KDEF in terms of F-Measure to show that our proposed approach attains the best classification performance compared to individual classifiers. Figure 7 presents the result attained during on-the-fly tests, where participants were interacting with a robot. In this case we have merged both datasets: KDEF and the one created through video sessions. The approach is the same, the KDEF dataset was used for training and the 6 participants are unseen persons. Figure 8 shows some examples of unseen persons running

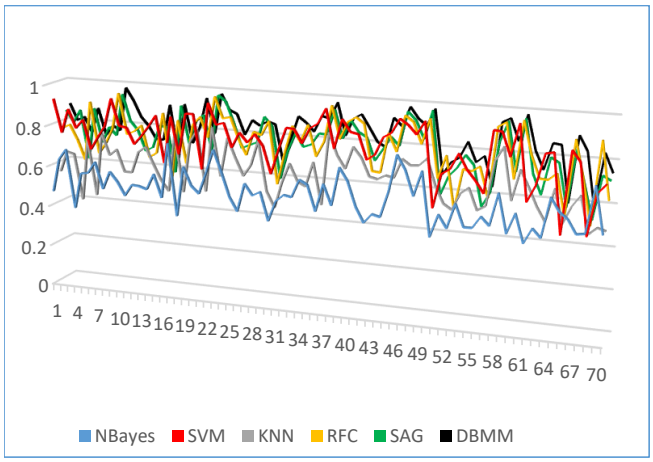

Fig. 6: Results of the 70 tests for the KDEF dataset in terms of F-Measure. We can see that the DBMM has a consistent performance when compared to its base classifiers.

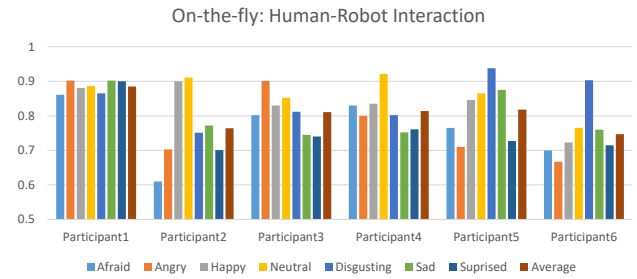

Fig. 7: Tests on-the-fly using the robot NAO. The KDEF dataset was used for training, and 6 new (unseen) subjects, 4 males and 2 females, participated in a human-robot interaction task, performing the emotional expressions. For the on-the fly tests the overall accuracy (average) was 80.6%

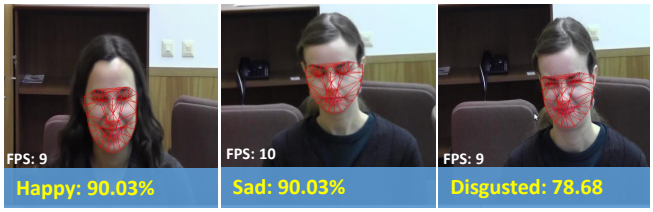

Fig. 8: Examples of facial expression recognition with unseen persons.

at 9 or 10 frames per second. These examples are from the dataset recorded during the video sessions.

A humanoid robot was used for the online tests, the well-known Aldebaran/Softbank NAO robot, taking advantage of its monocular camera to detect facial expressions. The python API from Aldebaran was used to access the NAO cameras and to provide some spoken and physical feedback as a way of interaction, once the facial expression is recognised. The libraries OpenCV [20] and Dlib [14] were used to detect and track the face landmarks. Two different sets of interactions were performed, where in the first one, the robot interacts

Fig. 9: Examples of tests on-the-fly during human-robot interaction. The robot reacts according to the person's emotion.

with a human based on s/he emotional state, i.e. in order to react according to it (e.g. Figure 9): for happy, the robot will ask for a hi5; for sad, the robot will play a song to cheer the person up; for disgusting, the robot will ask why the person is doing that disgusting face; if the person is afraid, the robot says that everything will be alright; for surprising, the robot will ask what the person saw to be that surprised.

The second interaction is slightly similar, but the robot tries to express the same emotional expression instead, using its body expression, similarly the work presented by [21], but here the robot reacts resembling the human emotion after recognising it during the interaction. The intention is to verify the response of the participants (e.g. positive or negative) when they see such expressions, in order to explore in a future work whether person is positively engaged during the interaction or not. Results from on-the-fly tests presented in Figure 7 were acquired after each participant performed 3 times each emotion. The results are in terms of classification confidence (average after 3 tests). P4, P5 are females and the remaining persons are males; the emotions are: 1-Afraid; 2-Angry; 3-Happy; 4-Neutral; 5-disgusting; 6-Sad; 7-Surprised; and AV-Average. Some of the experimental results with datasets and during human-robot interaction can be watched online[2].

[2]A video presenting some classification results including human-robot interaction is available at: `https://youtu.be/xS9OMLR3Ojc`

V. Conclusion and Future Work

Emotional expression recognition is an important and challenging topic for human-robot interaction. This paper presents an approach to classify emotional expressions based on affective facial expressions, with potential to be used in human-robot interaction. The proposed approach uses supervised classification techniques, facial-based features and the DBMM method. Reported results on the KDEF dataset and on-the-fly tests using a humanoid robot show that our solution attained an overall accuracy around 85% on datasets and 80% on tests on-the-fly during human-robot interaction. Future work will address an application for child-robot interaction, where emotional expression recognition will be essential to define appropriated robot reactions.

References

[1] D. Lundqvist, A. Flykt, and A. Ohman, "The Karolinska directed emotional faces - KDEF, Dep. of clinical neuroscience, psychology section, Karolinska Institutet, 1998."
[2] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psy. Bull*, 1994.
[3] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems, vol. 42, pp. 177190*, 2003.
[4] J. Yan, Z. Wang, and Y. Yan, "Humanoid robot head design based on uncanny valley and facs," *Journal of Robotics*, 2014.
[5] R. Picard, "Toward agents that recognize emotion," *IMAGINA*, 1998.
[6] J. V. den Stock, R. Righart, and B. de Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion, vol. 7, no. 3, pp. 487-494*, 2007.
[7] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consulting Psychologists Press*, 1978.
[8] S. Du, Y. Tao, and A. Martinez, "Compound facial expressions of emotion," in *National Academy of Sciences 111 (15)*, 2014.
[9] Q. Rao, X. Qu, Q. Mao, and Y. Zhan, "Multi-pose facial expression recognition based on surf boosting," in *ACII*, 2015.
[10] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition based on facial components detection and hog features," in *Int. Workshop on Electrical and Comp. Eng.*, 2014.
[11] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors, vol. 13 (6)*, 2013.
[12] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vision Comput.*, 2009.
[13] D. R. Faria, C. Premebida, and U. Nunes, "A probalistic approach for human everyday activities recognition using body motion from RGB-D images," in *IEEE RO-MAN'14*, 2014.
[14] D. E. King, "Dlib-ml: A machine learning toolkit," *J. of Machine Learning Res.*, 2009.
[15] D. R. Faria, M. Vieira, C. Premebida, and U. Nunes, "Probabilistic human daily activity recognition towards robot-assisted living," in *IEEE RO-MAN'15, Kobe, Japan.*, 2015.
[16] M. Vieira, D. R. Faria, and U. Nunes, "Real-time application for monitoring human daily activities and risk situations in robot-assisted living," in *Robot'15: 2nd Iberian Robotics Conf., Portugal*, 2015.
[17] C. Premebida, D. R. Faria, F. A. Souza, and U. Nunes, "Applying probabilistic mixture models to semantic place classification in mobile robotics," in *IEEE IROS'15*, 2015.
[18] C. Premebida, D. R. Faria, and U. Nunes, "Dynamic bayesian network for semantic place classification in mobile robotics," *AURO Springer: Autonomous Robotics.*, 2016.
[19] C. Coppola, D. R. Faria, U. Nunes, and N. Bellotto, "Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data," in *IEEE/RSJ IROS'16: International Conference on Intelligent Robots and Systems*, 2016.
[20] "Open source computer vision library (OpenCV). Website: http://opencv.org/ (Visited January, 2016)."
[21] M. Haring, N. Bee, and E. Andre, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *IEEE RO-MAN*, 2011.