Towards Multimodal Affective Expression: Merging Facial Expressions and Body Motion into Emotion

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

***Abstract*— Affect recognition plays an important role in human everyday life and it is a substantial way of communication through expressions. Humans can rely on different channels of information to understand the affective messages communicated with others. Similarly, it is expected that an automatic affect recognition system should be able to analyse different types of emotion expressions. In this respect, an important issue to be addressed is the fusion of different channels of expression, taking into account the relationship and correlation across different modalities. In this work, affective facial and bodily motion expressions are addressed as channels for the communication of affect, designed as an emotion recognition system. A probabilistic approach is used to combine features from two modalities by incorporating geometric facial expression features and body motion skeleton-based features. Preliminary results show that the presented approach has potential for automatic emotion recognition and it can be used for human-robot interaction.**

***Index Terms*— Emotion recognition, probabilistic approach, human-robot interaction**

I. Introduction

Emotion perception is pursued by different fields of research such as psychology, computer science and engineering, and advanced robotics. In robotics, which is the focus of this paper, a robot (seen as an AI agent) can be endowed with the ability of analysing verbal and non-verbal behavioural cues displayed by the user to infer the underlying communicated affect. In the latter, an affectively competent AI agent exploits affective states to successfully interact with humans; the perception and interpretation of emotional states through different body expressions allows an artificial agent to act more socially and engage with humans more naturally. Humans can rely on different channels of information to understand the affective messages communicated by others. Similarly, it is expected that an automatic affect recognition system should be able to analyse different types of affective expressions. When it comes to emotional expression in human-to-human communication, the face is one the main area of attention [1], since it transmits relevant information about emotions. However, an increasing attention is being also paid to the possibility of using body motion as expressions in order to build affectively aware technologies. The relevance of body motion expressions and the benefits of developing applications is evident in many areas, such as security, games and entertainment, education, and health care. An important issue to be addressed is the combination of different channels of expression, which must be designed by taking into account the relationship and correlation across different modalities. According to the Facial Action Coding System (FACS) [2], humans share seven emotional expressions regardless of ethnic group, culture, and country, and they are: Happiness; Sadness; Anger; Fear; Surprise; Disgust; and Contempt. Affect expression can also occurs through combinations of verbal and nonverbal communication channels including bodily expressions [3], however, it is evident that the study of perception of whole-body expressions lags so far behind facial expressions. In this work, facial and bodily expressions are addressed as channels for the communication of affect. This research shows a probabilistic framework to recognise human emotional expression by merging multimodal features. Preliminary results evidence the potential of this framework that can surely be used for human-robot interaction.

The remainder of this paper is organised as follows. Sec. II describes the recognition system. Sec. III describes the multimodal features and experimental setup. Sec. IV reports preliminary results for multimodal affect recognition and Sec. V presents conclusions and future work.

II. Probabilistic Recognition Framework

Our approach is based on the ensemble Dynamic Bayesian Mixture Model (DBMM) [4], [5], which is inspired on dynamic Bayesian networks for time-dependent problems and mixture models as weighting fusion. It was successfully used in different applications [6], [7], [8], [9]. We are extending this model to deal with different channels of affect, and to do so we have designed a two-layered fusion step as shown in Fig. 1. The first layer is a fusion at classification level for each modality separately, and the second one is used to merge the different channels of affect for emotion recognition. The two-layered fusion model is given by:

$$P(C^t|A^t) = \tfrac{1}{\beta} \times$$

$$\times \prod_{k=t}^{t-T} \left\{ \overbrace{P(C^k|C^{k-1})}^{\text{dynamic prior}} \overbrace{\sum_{y=1}^{M} \left[w_{2_y}^k \times \left(\underbrace{\sum_{i=1}^{N} w_{1_{y,i}}^k \, P_{y,i}(A^k|C^k)}_{\text{fusion of base classifiers}} \right) \right]}^{\text{fusion of multiple mixtures}} \right\}, \tag{1}$$

D. R. Faria is with [1]School of Engineering and Applied Science, Aston University, Birmingham, UK. 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, {fernanda, cpremebida}@isr.uc.pt).

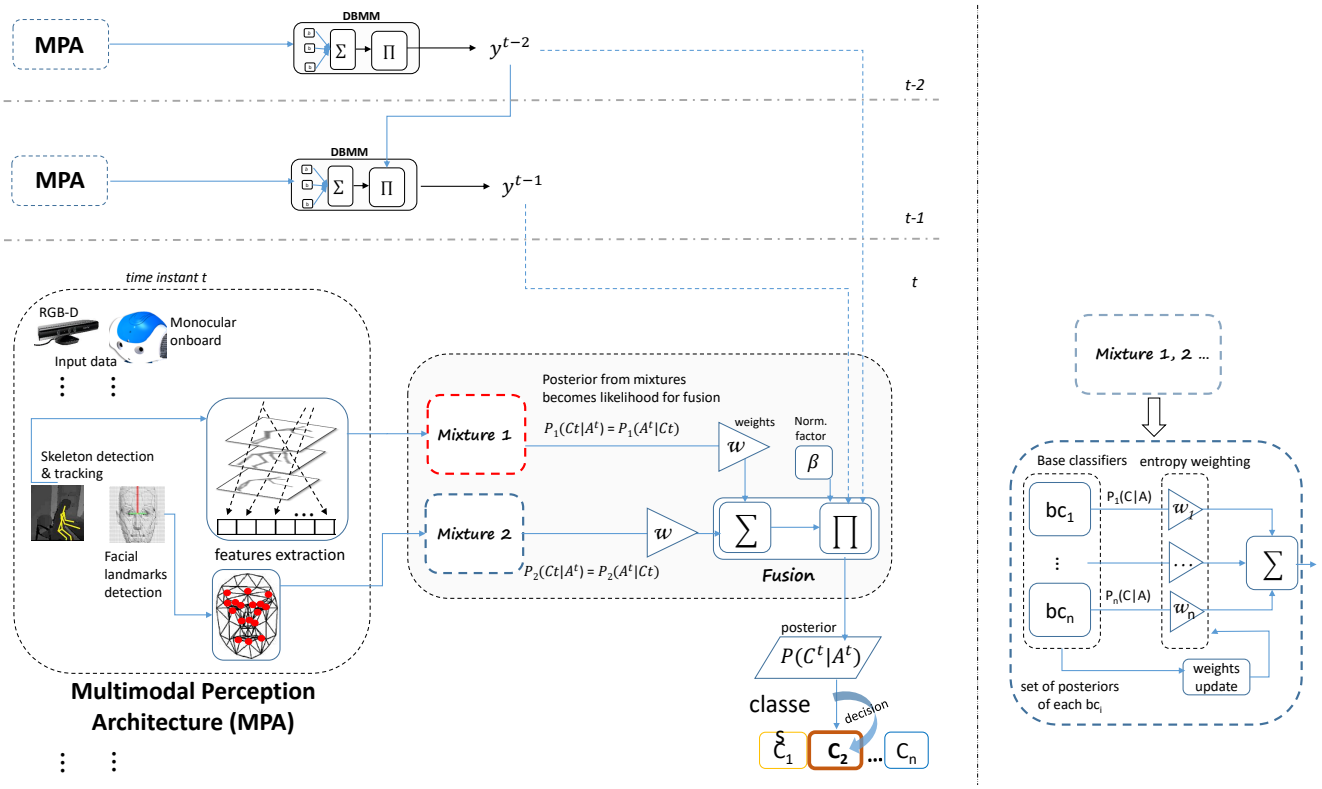


Fig. 1: Two-layered DBMM fusion for multimodal affect recognition using facial expressions and bodily motion features.

where $P(C^t|A^t)$ is the posterior with C representing classes of emotion and A feature models; $\frac{1}{\beta}$ is a normalization factor; $P(C^k|C^{k-1})$ is the dynamic prior given by posteriors from previous time slices; $y = \{1,...,M\}$ is an index for mixture models; $i = 1...,N$ is an index for the base classifiers; t is the current time instant, k is an index for time instants and T is the number of time slices; $P(A|C)$ denotes the output conditional probability from a learning model (base classifier); $w1$ is the weight for the first layer fusion (base classifiers) and $w2$ is the weight for the second layer that works as a fusion layer i.e., it combines the different channels of affect. The weights can be computed using entropy-based weighting and probability residual energy as show in [10], [11], [5].

The model is dynamic not only by the temporal transitions as function of the time slices, but also by the update of weights as well. During the classification stage, the performance of base classifiers in the ensemble model can change over time. Thus, a local update of weights during the classification can benefit from the fact that weights adjusted using information from previous observed frames will produce a higher belief for the next time instants. Basically, the update process occurs given a temporal slide window of precedent posteriors to the current frame.

The Kullback-Leibler (KL) divergence [12] is an asymmetric measure of the difference between two probability distributions. However, a symmetric measure is obtained by averaging the KL divergence, also known as Jensen-Shannon divergence (a.k.a. total divergence to the average [13]). Based on that, the divergence between prior and posterior distributions is computed, where the prior is the global weight learnt given the training set, and the posteriors are given by the classified test set frames precedent to the current frame. Given that, the weights for fusion are updated (runtime) employing the following steps:

$$D_{KL_i}(P(o_i^{\{1:t-1\}}) \parallel P(w_i^g)) = \sum_{l=1}^{t-1} P(o_i^l) \frac{P(o_i^l)}{P(w_i^g)}, \quad (2)$$

$$D_{KL_i}(P(w_i^g) \parallel P(o_i^{\{1:t-1\}})) = \sum_{l=1}^{t-1} P(w_i^g) \frac{P(w_i^g)}{P(o_i^l)}, \quad (3)$$

$$w_i^t = \frac{0.5}{\alpha} \times \left[D_{KL_i}\left(P(o_i) \parallel P(w_i^g)\right) + D_{KL_i}\left(P(w_i^g) \parallel P(o_i)\right)\right], \quad (4)$$

where $P(w_i^g) = w_i^g$ is the prior model, representing the global weight learnt from the training set using the one of the strategies described in [10], [11], [4]; $P(o_i) = P_i(C|A)$ is an observation that composes a set of posteriors from previous time instants $\{1,...,t-1\}$, i.e., previous posteriors from an i^{th} base classifier; $\frac{0.5}{\alpha}$ is a normalization factor to keep a symmetric measure, with $\alpha = \sum_i^N 0.5 \times \left[D_{KL_i}\left(P(o_i) \parallel P(w_i^g)\right) + D_{KL_i}\left(P(w_i^g) \parallel P(o_i)\right)\right]$.

In this work, the base classifiers used to compose the two-layered DBMM ensemble were: Support Vector Machines (SVM); Random Forest Classifier (RFC); and a linear regression based on Stochastic Average Gradient (SAG).

III. Experimental Setup and Multimodal Features Extraction

An environment to stimulate the participants' emotions was set up. In order to awaken emotions such as {happy/joy, angry, disgusting, afraid/scared, surprised, sad, neutral}, we asked participants to watch a sequence of videos. This sequence consists of emotional adverts, jokes and pranks that are expected to awaken different feelings. We carefully selected successful videos with thousands of views on youtube channels, also following suggestions of a psychologist. In this experimental setup, we have used a 50" tv screen to display the videos, a monocular camera to record the facial expressions, and an RGB-D sensor to track the body motion. Six individuals, three males and three females participated voluntarily. A dataset of emotions was built, consisting of RGB images with facial expressions and 3D skeleton data (i.e body joint motions) from the RGB-D sensor along 17 minutes of sequence of different videos as stimuli, resulting a data set with 30600 frames (1020 seconds $\times$ 30 frames per second) for both, images and skeleton data. We have manually annotated the sequence of videos and prepared a ground truth (i.e. expected reactions given the videos segments). More details about this experimental setup can be found in [10]. In addition, and to complement the data for learning purposes, we created a new dataset of body motion, an acting dataset, where we have asked the same 6 participants to show some body expressions that they usually do when they are in some emotional state. Each participant repeated an expression along 60 seconds, so that for each expression we acquired $6 \times 60 = 360$sec $\times 30 =$ 10800 frames. We have combined both datasets, i.e. the skeleton from the body expressions acquired along the video stimuli and from this acting sessions into one single dataset of body expressions. An additional public online available dataset for facial expressions, Karolinska Directed Emotional Faces (KDEF) [14], was also used in this work in order to improve the emotion learning. 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 individuals, 35 females and 35 males. This dataset comprises of actors performing specific facial expressions.

A. Extraction of Facial Features from Images

Given a single image with a human face, several sets of geometric features are extracted. First, 68 facial landmarks (e.g. contour of the face, lips, eyes and nose) are detected using the Dlib library [15]. Given that, we have computed several subsets of features as follows:

- Euclidean distances among all landmarks, obtaining a 68×68 symmetric matrix with a null diagonal. A normalization is performed to make the features scale-invariant. Finally, we compute the *log-covariance* over this matrix;
- Given the detected landmarks, triangles between them are computed. All three angles of a total of 91 triangles are computed.

In this work we have followed our previous work for facial expression features extraction. Full details about the features extraction is explained in [10].

B. Body Motion Features from Skeleton Data

In order to map body posture and movements into affect, we tried to exploit skeleton spatio-temporal features to characterise them. Since in previous works [4], [5] we have been successful in human daily activity recognition using spatio-temporal features from skeleton data, herein, for features extraction we are based on these previous works using features such as: Euclidean distances between joints; angles formed between joints; torso inclination; energy and log-energy entropy of joints velocities; and skeleton poses. More details can be found in [5].

IV. Preliminary Results

A. Facial Expressions Recognition

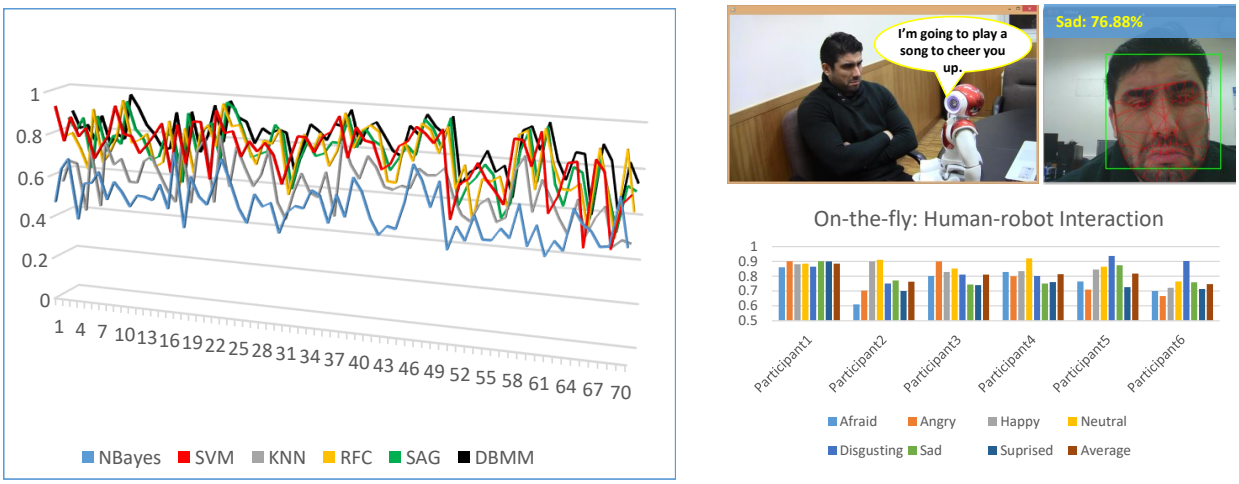
We have tested our approach on the KDEF dataset [14] adopting the leave-one-out cross-validation tests for the 70 persons in this dataset perfroming 7 emotions. Just to check the efficiency of the features shown in [10], we have compared some base classifiers and a single layered DBMM. Figure 2(a) presents all tests done on the KDEF in terms of F-Measure to show that the DBMM ensemble attains best classification performance compared to individual classifiers. Figure 2(b) presents the results attained during on-the-fly tests, where participants interacted 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 represent the testing set (unseen persons). For the on-the-fly tests, we have used a humanoid robot: Aldebaran NAO robot, taking advantage of its monocular camera to detect and recognise human facial expressions. We have used the python API from Aldebaran to access the NAO cameras and to provide some spoken and physical feedback as a way of interaction, once the facial expression is recognised.

B. Body Expression Recognition from Dataset

Given the created dataset, we have used the strategy of leave-one-out, obtaining the following overall results: precision 62.48%; recall 61.26%. The overall confusion matrix is presented in Fig. 2 (a). In this specific case, classifying the body expressions was not so easy, because the participants controlled their body expressions while watching the video sequence, showing more facial expressions than bodily. When analysing the features, the ones that evidences better the emotion are the head, shoulders and arms.

C. Multimodal Emotional Expression Recognition

Figure 2 (b) shows the overall results for multimodal affect recognition after the performance of 6 participants. In this case, the training sets were all the aforementioned datasets for both, facial and bodily expressions (naturally stimulated). The overall result obtained were precision: 82% and recall: 81.67% after merging the final classification using



(a) (b)

Fig. 2: (a) Results on the KDEF dataset in terms of F-Measure for Naive Bayes (56%), SVM (80.0%), k-Nearest Neighbor (65.0%), Random Forest (78.0%), linear regression by SAG (78.0%) and a single layer DBMM (82.8%). (b) On-the-fly tests using a robot with 80.6% of overall accuracy (average). Top: a sample from the acquired dataset. Bottom: overall results (average after each individual performing 3 times each emotion).

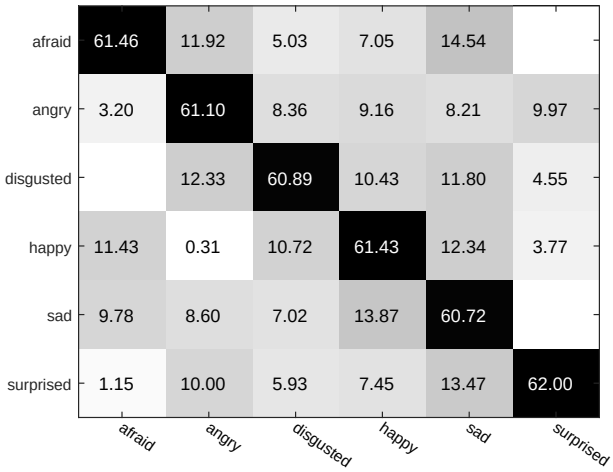


Fig. 3: Body expressions results: dataset acquired through video stimuli, Prec: 62.48%; Rec: 61.26%

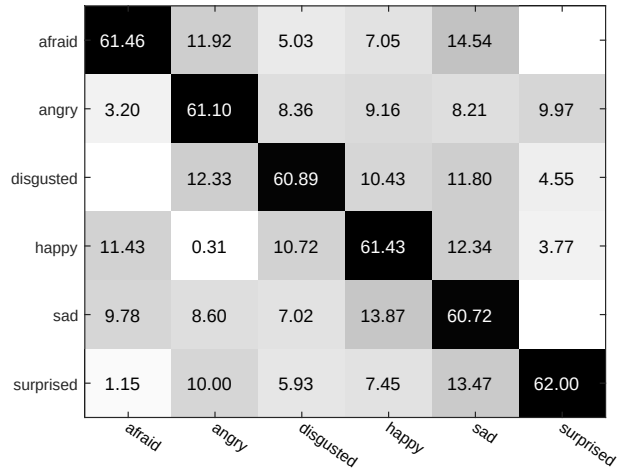


Fig. 4: Tests on-the-fly using multimodality: merging facial expressions (KDEF dataset + dataset created with video stimuli) and bodily motion. Prec: 82%; Rec: 81.67%

the two-layered DBMM. In cases where the body expressions is minimum (small body motion), the proposed approach assigns priority to the facial expressions in order to estimate the emotion, since only by body expression, depending on the person, recognising emotions is very difficult. By resorting to multimodality our approach obtained better performance when compared to the body motion modality, and slightly similar to the facial expressions. In this case, the multimodality compensates the final result during the merging step, since emotion through body motion is more difficult. Further tests will be carried out in order to investigate how is possible to improve these results.

V. Conclusion and Future Work

This paper presents a probabilistic approach for multimodal affect recognition using a two-layered dynamic ensemble. This approach can correctly classify emotional expressions with potential to be used in human-robot interaction. Based on preliminary results, when comparing both channels of affect, facial expression played the most important role for the fusion process. Future work will investigate how to improve the recognition of bodily motion to improve the overall performance. We will also explore other datasets with more data and diversity to better discriminate affective state.

REFERENCES

[1] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies," *Psy. Bull*, 1994.

[2] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consulting Psychologists Press*, 1978.

[3] R. Picard, "Toward agents that recognize emotion," *IMAGINA*, 1998.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C. Premebida, D. R. Faria, and U. Nunes, "Dynamic bayesian network for semantic place classification in mobile robotics," *AURO Springer: Autonomous Robotics*, 2017.

[9] J. Vital, D. R. Faria, G. Dias, M. Couceiro, F. Coutinho, and N. Ferreira, "Combining discriminative spatio-temporal features for daily life activity recognition using wearable motion sensing suit," *PAA Springer: Pattern Analysis and Applications*, 2016.

[10] D. R. Faria, M. Vieira, F. C. Faria, and C. Premebida, "Affective facial expressions recognition for human-robot interaction," in *IEEE RO-MAN'17: IEEE International Symposium on Robot and Human Interactive Communication, Lisbon, Portugal.*, 2017.

[11] D. R. Faria, M. Vieira, and F. C. Faria, "Towards the development of affective facial expression recognition for human-robot interaction," in *ACM PETRA'17: 10th International Conference on PErvasive Technologies Related to Assistive Environments*, 2017.

[12] S. Kullback, "Information theory and statistics," *J. Wiley & Sons*, 1959.

[13] I. Dagan, L. Lee, and F. Pereira, "Similarity-based methods for word sense disambiguation," *35 Annual Meeting of the Assoc. for Comp. Linguistics and 8 Conf. of the European Chapter of the Assoc. for Comp.Linguistics: 56-63*, 1998.

[14] D. Lundqvist, A. Flykt, and A. Ohman, "The Karolinska directed emotional faces - KDEF, Dep. of clinical neuroscience, psychology section, Karolinska Institutet, 1998."

[15] D. E. King, "Dlib-ml: A machine learning toolkit," *J. of Machine Learning Res.*, 2009.