Effects of Emotion Grouping for Recognition in Human-Robot Interactions

Daniel C. Tozadore, Caetano M. Ranieri,
Guilherme V. Nardari, Roseli A. F. Romero
Institute of Mathematics and Computer Sciences
University of São Paulo
São Carlos, SP, Brazil
Email: {tozadore, cmranieri, guinardari}@usp.br,
rafrance@icmc.usp.br

Vitor C. Guizilini
School of Information Technologies
University of Sydney
Sydney, NSW, Australia
Email: vitor.guizilini@sydney.edu.au

Abstract—**Understanding people's emotions may be important to achieve success in behavior adaptability and, consequently, to sustain long-term human-robot interactions. Most emotion recognition systems consist in classifying a given input into one out of seven basic emotions, following Ekman's model. However, it is sometimes enough for the customization of a robot's behavior to recognize whether an emotion is positive or negative, in order to approach more often subjects which display more positive emotional reactions. In this article, two approaches to that effect are proposed and compared. The first one, named pre-grouping, refers to combining the four negative emotions into one single class and use it to train a classifier. The second one, named post-grouping, refers to applying classifiers to classify the seven basic emotions and interpret their negative outputs as related to a single class. Furthermore, a novel dataset entitled QIDER, based on queries in a search engine and well defined facial cues, is introduced and made available for public use. Both approaches led to more balanced precision scores among all classes, which may make them a suitable choice for applications in human-robot interaction. Several experiments have been performed and post-grouping is shown to produce better overall accuracy.**

I. Introduction

Human-robot interaction (HRI) research analyzes interactions in which robots play collaborative roles with human users. One drawback found in such studies is that, after a given user gets familiarized with the robot, a decrease in his motivation is verified, due to long-term factors such as lack of novelty in successive interactions [1]. Dynamic customization on the robot's behavior is as a potential solution to minimize this problem. Thus, adaptation may be essential for successful autonomous personalized interactions over a longer period of time. For HRI, one of the most relevant forms of non-verbal cues, which plays a fundamental role in interpersonal communication, are facial expressions. Fast analysis and adaptation to these indicators may be crucial to hold users' attention and achieve more engagement in several domains of interactive tasks. For instance, it is possible to interpret a user's smile as an emotive reaction to an enjoyable robot's behavior, and therefore reinforce this behavior, increasing its probability to be displayed in the future.

Although recent advances on computer vision have significantly improved results in emotion recognition based on facial expressions [2], the available techniques have led to poor precision when recognizing certain emotions, partially because they are underrepresented in the available datasets. It may be noticed that these emotions correspond to negative emotions, as may be observed in Russel [3]. As distinguishing between negative emotions is not central to most HRI applications, which aim to apply emotions mostly in rules for reinforcement learning, grouping them into a single class may provide improvements in precision without undesirable drawbacks.

In this paper, we have considered two consolidated network architectures, Inception-v3 [4] and MobileNet [5], to classify static images from human faces into a subset of the Ekman's model of affect [6], which consists of six basic emotions: *happiness*, *anger*, *sadness*, *fear*, *disgust* and *surprise*, plus the *neutral* emotion. The key idea is to group the negative emotions, defined as *anger*, *sadness*, *fear* and *disgust*, and use them as a single entity for classification purposes. Three datasets were considered for training and evaluating the models: the FER+ [7] and RAF [8] plus a novel dataset, the Query-Based Image Dataset for Emotion Recognition (QIDER), based on images downloaded from the internet using predefined queries on a search engine and filtered based on well-defined features [9], analyzed by human observation. Besides training and evaluation on each individual dataset, additional experiments were conducted in which the training subsets of these databases were merged and used as input for the networks' training stage. This merged dataset was named *Global*.

The remainder of this paper is organized as follows: In Section II, a review of related methods and applications in robotics is presented. Details about the datasets, network topologies and architectures adopted in our study are presented in Section III. In Section IV, experiments are performed with various datasets, with results being described and then discussed in Section V. Finally, Section VI concludes the paper and provides direction for future work.

II. Related Work

Improving emotion recognition is important not only for HRI, but also for any computational application that demands affective feedback from the user. For this reason, a lot of recent research focuses on designing systems for this task. Burket *et al.* [10] proposed a Convolutional Neural Network (CNN) for motion classification composed by a classic CNN architecture followed by parallel independent layers for further feature extraction. The combination of multiple machine learning models to achieve better accuracy is a well established technique. However, in the case of deep learning architectures, high model complexity can make ensembles impracticable, especially in robotics where resources are limited and fast response time is necessary. Surace *et al.* [11] proposed a CNN solution solution based on a novel combination of deep neural networks and Bayesian classifiers to the Group Emotion Recognition in the Wild challenge. Their method achieved an accuracy of 64.68% on the test set, significantly outperforming the 53.62% competition baseline, which consist in images of humans took in their daily tasks (including crowd). In Yan *et al.* [12], a convolutional neural network does a coarse prediction, which is then used to select a specialized network that is responsible for the fine prediction, using low-level features from the first model and specific learned features for the corresponding group. With a balanced random subsample of the training set, a confusion matrix is calculated and used to define the class groups. Deng *et al.* [13] proposed a graph structure representation of classes that considers both hierarchical and exclusive relationships. Guo *et al.* [14] presents a framework where CNNs are used as features extractors and its last layers are substituted by recurrent neutral networks (RNNs) to model the sequential relationship of hierarchical classes. Hamester *et al.* [15] used a Multi-channel Convolutional Neural Network (MCCNN) combined with Convolutional Autoencoder (CAE) for training booster. Later, they tested the proposed model in HRI experiments [16]. The authors claimed to achieve significantly better results, even though posed databases were used for training.

HRI applications, studies comparing adaptive systems have shown significant improvements in user experience when applying such techniques [17]. Moreover, perceiving emotion from images is considered the most contribution factor in multimodal systems, and deep learning is a method that shows appropriate support in real time applications and seems to set a promising path to follow [18]. Ruiz-Garcia *et al.* [19] proposed a CNN for emotion recognition in a real-time empathic robot system. Although no tests with robot in real world were performed so far, the proposed model achieved near state-of-art results in the Karolinska Directed Emotional Faces (KDEF) dataset. With a study about the effect of pre-training a Deep CNN as a Stacked Convolutional Auto-Encoder (SCAE) in a greedy layer-wise unsupervised fashion for emotion recognition using facial expression images, a performance of 92.52% was obtained, against 91.15% from models with randomly trained weights.

III. Methodology

As already stated, this work consisted in evaluating the Inception-v3 and MobileNet architectures in three datasets of static facial expressions, related to the six Ekman emotions plus neutral. Two approaches for grouping the negative emotions into one single class were considered, in order to provide well-suited models for HRI scenarios. The next subsections will give further details on each of these steps.

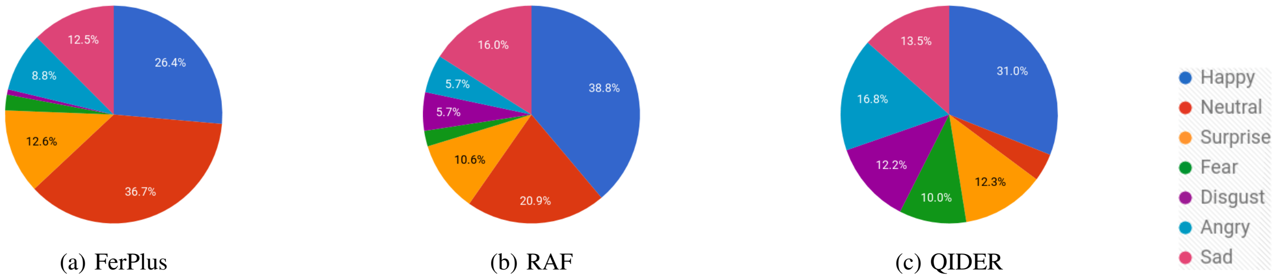
A. Datasets

There are two kinds of emotion datasets based on facial expressions: *posed* or *real-world*. On one hand, posed datasets are generally small collections of images, obtained by taking pictures of actors performing predefined facial expressions under controlled conditions. Examples are the Radbound Faces Database (RAFD) [20] and the Extended Cohn-Kanade (CK+) [21]. Although these datasets may be good platforms for performing experiments that require homogeneous data, as they are comparatively small, classifiers trained on them may provide poor generalization when dealing with different lighting conditions, ethnic characteristics and poses. Real-world datasets, on the other hand, are often obtained by downloading images from the internet and labeling them. Examples are the Facial Expression Recognition dataset (FER) [7] and the Real-world Affective Faces (RAF) [8] datasets. These models are considerably bigger, noisier and more heterogeneous, which makes them suitable for deep neural networks, and may potentially lead to better generalization.

In this work, we have chosen to perform the experiments on real-world datasets, since we are focused on applications for HRI. All presented models and results are based on the three datasets described in the next subsections. These datasets were FER+ [7], RAF [8] and a new dataset that we have gathered, named QIDER. The class distribution for each dataset is shown in Figure 1. Images were normalized to gray-scale and resized to 224×224 by bilinear interpolation. To increase variance in the training data and make the model less sensitive to the position of the user, the following data augmentation techniques were applied: random rotations by up to 40^o, random width and height shifts with 20% of the corresponding dimension, random shear with an angle of 120^o, random zooming with range $[0.8, 1.2]$ and horizontal flipping.

1) FER+: The FER dataset was first made available as a Kaggle competition. It was composed by a total of 35,713 images. This dataset was largely adopted as a standard benchmark for emotion recognition from static facial expressions [22], which makes it suitable for comparison with existing works on the literature. However, by observing the data, one may conclude that it presented a significant number of inconsistent labels, probably due to wrong interpretations during the annotation process. To provide better annotations, the FER+ labels were proposed [7], in that a set of 10 people have labeled each image, from which the total score for each emotion was provided. An additional label, called *contempt*, was also defined and considered in the annotations. To assign a single emotion for an image, three approaches were proposed: majority voting, multi-label learning and probabilistic label drawing. In this paper, the majority voting approach was adopted in all scenarios in which the FER+ dataset was applied, and the *contempt* class was disconsidered, since this class, represented by only 221 instances, was absent of the other datasets analyzed and do not figure in Ekman's model.

2) RAF: The RAF dataset [8] was elaborated to provide both single and compound emotions for images collected on the internet. It provides 29,672 images, from which 15,339 refer to the Ekman's emotions plus neutral. Each image was labeled by 40 different people, and the class was calculated via a reliability estimation based on the annotator's background. This dataset was included in this work to provide additional data and to serve as another platform for comparisons between our models and others on the literature.



(a) FerPlus (b) RAF (c) QIDER

Fig. 1: Class distribution for the datasets described

3) QIDER: The Query-Based Image Dataset for Emotion Recognition (QIDER) dataset was proposed based on the visual criteria from facial expressions that define Ekman's emotions. It was composed by images searched on the internet, with queries built according to the following scheme: for each emotional state there is an *emotion*, an *emotion feature* and a *human profile*, as shown in Table I. All the *emotion features* were combined with all the *human profiles* to compose the emotions classes and then be filtered for labeling. For instance, "happy man", "smiling black woman" and "smiling woman" are queries to the *happiness* class.

TABLE I: Emotion features (left) and human profiles (right) used in the queries scheme.

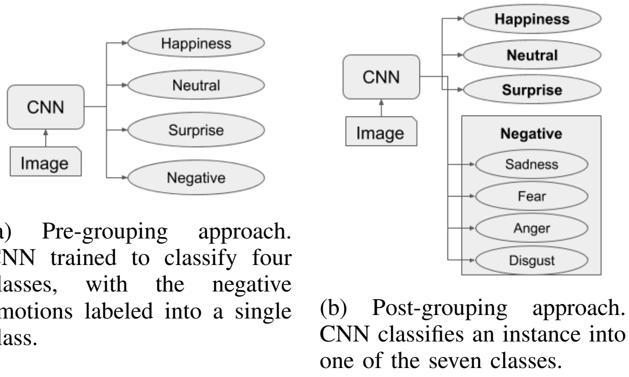
Emotion Class	Emotion Features
Happy	happy, smiling
Anger	angry, furious
Disgust	disgusted, aversion
Fear	fear, scared
Sad	sad, disappointed
Surprise	surprised, shocked
Neutral	neutral

Profiles
Man
Woman
Boy
Girl
Child
Black man
Black woman
Asian
Human face

The images were filtered by human analysis following the criteria stated in Scott and Brave [9], in which each emotion is defined by a set of well-defined facial cues. Three people analyzed each image, and the annotated label was the one with more votes. After applying the described filter in the data, the built-in OpenCV implementation of the Viola and Jones method [23] was applied to detect and extract only the faces from the samples. The extracted faces were converted to grayscale images, in order to standardize with the other datasets considered in this work. The resulting dataset was composed by 3,904 images of human faces displaying emotions. The labels were distributed as follows: 656 as anger, 477 as disgust, 390 as fear, 528 as sadness, 1,211 as happiness, 481 as surprise and 161 as neutral. The QIDER dataset may be acquired on request, by contacting the authors of this paper.

B. Proposal

From an application point of view, it makes sense to group negative emotions into a single class, since the goal is usually to give a negative feedback for the behaviors of an interactive system. There is psychological understanding that positive and negative emotional states are especially distinguishable from the subject's perspective [24]. Besides, negative emotions are usually underrepresented on almost all real-world datasets, which leads to poor precision scores [25], making it almost impossible to build refined interaction strategies based on these particular emotions. Given this scenario, we have defined a reduced set of labels for classification, by combining the four negative emotions - *fear*, *disgust*, *anger* and *sadness* - into a single class. As result, four classes were considered as outputs for the classifiers: *happiness*, *neutral*, *surprise* and *negative*. These groups of labels are shown in Table II. The two approaches considered, named *pre-grouping* and *post-grouping*, are represented in Figure 2.



(a) Pre-grouping approach. CNN trained to classify four classes, with the negative emotions labeled into a single class.

(b) Post-grouping approach. CNN classifies an instance into one of the seven classes.

Fig. 2: Proposed approaches.

TABLE II: Groups of emotions considered in the network models.

Group 1	*Happy, Surprised, Neutral, Negative*
Group 2	*Sadness, Fear, Anger, Disgusted*

C. Hierarchical Models

Hierarchical approaches consist of training several neural networks, each related to a different level of abstraction [13]. In order to investigate whether taking a model with the group 1 classes as the first level of an hierarchical network and fine-tune the negative emotions in a second stage could improve classification accuracy, we have analyzed three different approaches. The first approach, named *single-step*, consists of providing a model with no hierarchy at all, with the network simply taking an image as input and trying to classify it into one out of the seven Ekman's emotions. Both other approaches, respectively called *hierarchical 4-4* and *hierarchical 7-4*, consist of two neural networks, each one comprising a *level*. In the hierarchical 4-4 approach, the first CNN followed the pre-grouping approach, in which the image is classified into one of the group 1 labels. In the hierarchical 7-4 approach, the first CNN followed the post-grouping approach, classifying the input image into one of the seven emotions and grouping the negative emotions *a posteriori*. In both cases the second CNN, activated only when the *negative* class was produced, would further classify the input image into one of the group 2 labels. The second CNN shared some of the weights with the first, in order to take advantage of the facial features that may have been learned by the first convolutional layers, trained with the whole dataset. In other words, the model performed transfer learning on the bottom layers, responsible for the generation of lower-level features [26], and reused their weights without further training, while the rest of the network is randomly initialized. For the models based on Inception-v3, the parameters were reutilized on all layers until the 6th concatenation (56th convolutional layer), re-training the remaining layers on the second level network. For the models based on MobileNet, parameters were reutilized until the penultimate set of operations, and only the last group of operations was re-trained, consisting of the last convolution, batch normalization, ReLU, global average pooling and softmax layers.

IV. Experiments and Results

The experimental setup was elaborated to provide comparisons between the proposed architectures within themselves and with other research available in the literature.

A. Experimental Setup

For all experiments, the Keras[1] framework was used, with Tensorflow[2] back-end. This framework enables fast prototyping and validation of ideas with ready-to-use implementations of important architectures, including Inception-v3 and Mobilenet. Training was performed on a GeForce Titan Xp GPU. The code was made available online at Github[3].

The already mentioned combination of the three train sets, named Global train set, was used to train the models. The network architectures presented in Section III were trained from scratch for 100 epochs with batches of size 32, where one epoch is a sequence of updates in which all inputs were processed. The models based on Inception-v3 and MobileNet were optimized with the stochastic gradient descent (SGD) algorithm, with learning rate 10^{-3} and momentum 0.9. The cross-entropy was adopted as the cost function. For evaluation, each of the test sets was considered individually.

B. Results

The results shown in Tables III and IV were obtained from network models trained on the Global train set and tested individually in each test set. In Table III, to perform emotion classification into the four Group 1 classes, two approaches were considered: *pre-grouping* the four negative (i.e., group 2) classes into a single label and train the model with only four classes or *post-grouping* the outputs of a model trained to classify the seven original classes. Table IV contains results shown by the models trained for classifying the seven emotions (i.e., seven labels as possible outputs). We have chosen to train the models with Global dataset, rather than training them in each dataset individually, because of our focus on the deployment of applications for human-robot interaction.

TABLE III: Accuracies for the Group 1 classes.

Dataset Model	FERPlus	RAF	QIDER
InceptionV3 pre	83.1%	82.7%	82.6%
InceptionV3 post	84.7%	81.1%	82.2%
MobileNet pre	83.2%	80.8%	79.8%
MobileNet post	84.3%	83.7%	82.6%

TABLE IV: Accuracies for the seven classes.

Dataset Model	FERPlus	RAF	QIDER
InceptionV3	82.8%	78.5%	72.2%
InceptionV3 4-4	80.7%	77.1%	75.5%
InceptionV3 7-4	82.9%	79.5%	75.7%
MobileNet	81.9%	80.3%	76.2%
MobileNet 4-4	80.9%	76.5%	72.3%
MobileNet 7-4	82.4%	80.2%	77.0%

To present the performances of both MobileNet and Inception, the F1-scores provided by them, when evaluated on the Global test set, are shown in Fig. 3. Corroborating to our prior discussion, the lowest scores were found for the negative labels, probably due to their under-representation on all datasets.

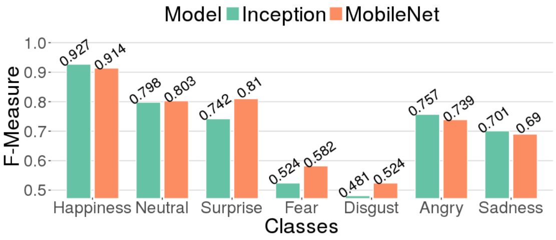


Fig. 3: Results from Inception-v3 and MobileNet trained and tested on the *Global* dataset.

To allow comparisons with other results provided in the literatura, an additional experiment was made, in which the models were trained and tested individually on the FER+ and RAF datasets. Results are shown in Table V. The single-step and hierarchical models were elaborated based on both Inception-v3 and MobileNet, presented separately in the table. Results reported in the literature, obtained from VGG architectures in both reference papers, were included for comparison. We have also provided results for the QIDER dataset, which may be used for further reference.

In order to evaluate the performance of each model regarding computer resources, the number of parameters and the relative time elapsed for feed-forward, on a standardized set of 100 samples, were measured and are presented in Table VI. For the hierarchical models, the worst case (i.e., negative emotions) was considered.

[1] https://keras.io/

[2] https://www.tensorflow.org/

[3] https://github.com/LAR-USP/emotion-recognition

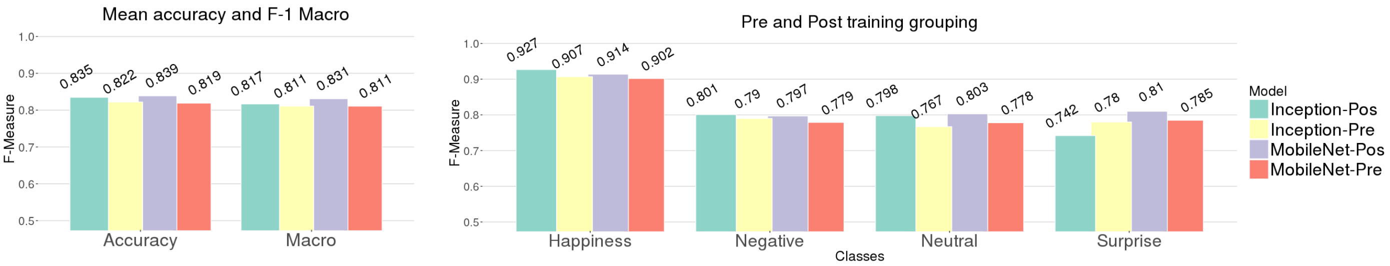


Fig. 4: Mean accuracy and F-measures for classification of four emotions. F1-macro was included to give a global understanding of how balanced were the classifiers.

TABLE V: Accuracy results for the models trained and evaluated on each dataset.

Dataset	CNN Model	Single-step	Hierarchical 4-4	Hierarchical 7-4
FER+	Reported [7]	84.5%	-	-
	Inception-v3	81.7%	76.4%	76.5%
	MobileNet	80.6%	79.5%	79.0%
RAF	Reported [8]	74.3%	-	-
	Inception-v3	76.6%	73.6%	71.7%
	MobileNet	75.4%	68.6%	73.2%
QIDER	Inception-v3	49.4%	59.6%	48.8%
	MobileNet	58.5%	55.9%	58.4%

TABLE VI: Complexity and average time elapsed for the tests.

Model	**Parameters**	**Average time (ms)**
Inception	$23,851,784$	44 ± 1
Inception H	$38,013,358$	56 ± 2
MobileNet	$4,253,864$	33 ± 1
MobileNet H	$5,282,960$	37 ± 2

V. Discussion

Regarding classification of group 1 classes, post-grouping performed better than pre-grouping in all the MobileNet scenarios, although it performed slightly better when the Inception model was tested on RAF and QIDER datasets. Hierarchical approaches showed no improvement in classification performance. For most image processing tasks, concept-based hierarchies may result on relatively similar visual features between instances on each level. For example, one may define a class *feline* by grouping cats, lions and tigers. The task of classifying emotions, however, may be especially challenging, due to its subjectivity and possibility of compound emotions. It means that emotions which are similar from a biological point of view may share few geometrical or texture features, which possibly explains the lack of significant improvements obtained by grouping the negative emotions as a single label for the CNN, even though these emotions share relevant psychological factors [24]. This fact may explain why, in this work, pre-grouping did not led to significant improvements, even making the results worse in most scenarios. This result differs from Deng *et al.*, which could improve their results by composing hierarchies with the HEX graphs.

Concerning the complexity and the time elapsed to perform the tests, shown in Table VI, the Inception-v3 single-step model requires about 5.6 times the number of parameter of the MobileNet single-step, and takes about 1.3 times to perform the tests. The hierarchical approaches led to more distinct performance, even on the 4-4 approach, which has shown to be slower than the 7-4. In this case, the Inception-v3 required 7.2 times the number of parameters of the MobileNet and took 1.5 times longer to process the samples. Although, regarding the accuracies, the results were not considerably different, and the F1-Macro related to the MobileNet-post performed even slightly better than the Inception-v3-post. These results indicate that, between the proposed models, MobileNet hierarchical 7-4 is the most suitable for HRI applications.

When trained and evaluated with FER+ and RAF separately, both hierarchical architectures hit lower accuracies than the single-step CNNs. For the FER+ dataset, results between the hierarchical 4-4 and hierarchical 7-4 varied less than 0.5%, which suggests the hierarchical approach is not significantly influenced by the type of grouping. However, for the RAF and QIDER datasets, results from the hierarchical 4-4 model were perceptibly more accurate than the hierarchical 7-4 for the Inception-v3 approach, while the opposite was observed for the MobileNet, which performed almost 5% more accurate. This may be explained due to the different complexities of the Inception and Mobilenet models, as shown in Table VI. Since the models were trained from scratch, the complexity of the Inception architecture and the scarcity of data in both the QIDER and RAF dataset make it more susceptible to overfitting and favoring majority classes, thus yielding better accuracies in the single-step model while predicting the most popular negative class in the second phase.

Our classification results for the FER+ dataset hit lower accuracies than reported in related literature, although, for the RAF dataset, our single-step approach hit better accuracy. This is probably because Barsoum *et al.* [7] performed more successful fine-tuning on their network, which was not performed due to our focus on efficiency and accuracy comparison between architectures and label groups.

Recalling the fact that some robotic systems have limited resources and taking into account the evaluations of accuracy, F-measure, complexity and time, the MobileNet showed better suitability for the applications considered here.

VI. Conclusion

The main contribution of this study was the evaluation of the effects of class grouping, performance of hierarchical models for underrepresented classes and a new dataset for real-world scenarios of emotion on static images. Different approaches for emotion recognition were evaluated and compared. The Inception-v3 and MobileNet newtork architectures were trained with distinct hierarchical approaches. The emotion databases FerPlus and RAF, as well as a new one named QIDER (Query-Based Image Dataset for Emotion Recognition), introduced for the first time in this paper, were used for validation.

Training and testing the proposed hierarchical networks in each database showed results slightly lower to what is reported in the corresponding papers. Joining all the databases in the Global dataset showed that the MobileNet 7-4 provides more satisfactory results for robotic projects, achieving similar or higher accuracy than Inception-v3 in all scenarios. Considering that the main goal in HRI designs is to optimize a set of aspects that can lead the users to achieve success in the proposed tasks, sometimes it is better to treat possible system mistakes along the interaction than to use heavy methods with high accuracy, but with time delays that may compromise the robot interaction experience as a whole.

The final application of this study is the incorporation of the classifiers into an adaptive system that controls the robot's behavior base on the user's emotion. Thus, in future work we will focus on the MobileNet model for several scenarios in which the users interact with a robot controlled by the adaptive system. Emotions eventually displayed by the users will be used as ground-truth to evaluate the model in real world situations.

Acknowledgment

This research was supported by the the Brazilian National Council for Scientific and Technological Development (CNPq), the São Paulo Research Foundation (FAPESP), grants 2017/02377-5 and 2017/17444-0, and the INCT (National Institute of Science and Technology).

References

[1] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[2] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[3] J. A. Russell, "Pancultural aspects of the human conceptual organization of emotions." *Journal of personality and social psychology*, vol. 45, no. 6, pp. 1281–1288, 1983.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[7] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[8] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 2584–2593.

[9] S. Brave and C. Nass, *Emotion in human-computer interaction*, 2003, pp. 53–58.

[10] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[11] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi, "Emotion recognition in the wild using deep neural networks and bayesian classifiers," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. New York, NY, USA: ACM, 2017, pp. 593–597. [Online]. Available: http://doi-acm-org.ez67.periodicos.capes.gov.br/10.1145/3136755.3143015

[12] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2740–2748.

[13] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, "Large-scale object classification using label relation graphs," in *European conference on computer vision*. Springer, 2014, pp. 48–64.

[14] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, "Cnn-rnn: a large-scale hierarchical image classification framework," *Multimedia Tools and Applications*, pp. 1–21, 2017.

[15] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[16] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, nov 2015, pp. 582–587.

[17] N. Churamani, M. Kerzel, E. Strahl, P. Barros, and S. Wermter, "Teaching emotion expressions to a human companion robot using deep neural architectures," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 627–634.

[18] A. Hernandez-Garcia, "Perceived emotion from images through deep neural networks," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 566–570.

[19] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1586–1593.

[20] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94–101.

[22] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001.

[24] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–1071, 1988.

[25] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 443–449.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.