Are Facial Expression Recognition Algorithms Reliable in the Context of Interactive Media? A New Metric to Analyse Their Performance

Emmanuel V.B. Sampaio, Lucie Lévêque, Matthieu Perreira da Silva, Patrick Le Callet

Are Facial Expression Recognition Algorithms Reliable in the Context of Interactive Media? A New Metric to Analyse Their Performance

EMMANUEL V. B. SAMPAIO, LUCIE LÉVÊQUE, MATTHIEU PERREIRA DA SILVA, and PATRICK LE CALLET, Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Emotions, and consequently facial expressions, play an essential role in communication - and thus in everyday life. With the increase of human-machine interactions, and more especially of multimedia applications, automatic recognition of facial expressions has emerged as a challenging task, particularly under naturalistic conditions. In the present work, a benchmark is firstly conducted using four open source deep learning solutions on four labeled image datasets. Thanks to an exhaustive analysis based on two distinct, yet complementary approaches, we show how the four models performed depending on the studied emotions. Furthermore, we present a novel metric based on the Euclidean distance between two given emotions (i.e., ground truth and predicted) to better measure the performance of said models in the context of interactive media, where human sensibility needs to be taken into consideration.

Additional Key Words and Phrases: Emotions, facial expression recognition, benchmark, distance metric, interactive media.

1 INTRODUCTION

Coppin & Sander stated a few years ago that the "topic of emotion rarely leaves individuals unemotional." [1]. In fact, emotions play an essential role in decision making, perception, and learning, and are thus a crucial part of people's everyday life. They have been under investigation for the last decades in a wide range of fields, including, but not limited to: psychology, sociology, neuroscience, cognitive science, and more recently, computer science, with the advent of a specific domain called *affective computing* [2]. Affective computing is the study and development of systems and devices able to recognise, interpret, process, and simulate human affects - or, the so-called emotions - which appear to be complex inherent characteristics of human beings.

Facial expressions represent an important part of implicit (i.e., nonverbal) communication. Mehrabian *et al.* indeed stated that facial expressions reflect 55% of human communication, directly followed by voice [3]. As facial expressions allow communicating feelings, analysing them can consequently contribute to the investigation of human emotions. Darwin himself advocated that facial expressions were "residual actions of more complete behavioural responses to environmental challenges" [4]. Measuring facial expressions indeed allows to get insights on emotions in real time, and via a non-invasive approach (compared to the use of physiological signals for instance).

In the domain of computer vision, three major perspectives to distinguish facial expressions appear. The most popular one is a categorical model, widely known as the *7 basic emotions*, derived from Ekman's theory [5]. According to his theory, every emotion can be classified in one of the following categories: anger, disgust, fear, happiness, surprise, sadness, or neutrality. Furthermore, the model states that each of these emotions can be characterised by a unique facial expression [6], and the latter can be considered universal [7]. The other two main descriptor models aim to characterise a larger scope of emotions. The *Facial Action Unit System (FACS)* indeed classifies facial muscle movements, and provides 44 distinct action units (AUs) [8]. This taxonomy has been widely embraced for complex facial expressions, as it enables the description of specific facial muscles. At last, the continuous *2D valence and arousal space*, further developed to the *3D valence, arousal, and dominance space*, allows a representation of an emotion according to two - or three - independent and complementary axis [9]. More specifically, the valence, or pleasure, ranges from unhappiness to happiness; the arousal, or affective

IMX '22 - Emotion, 2022, Portugal

activation, ranges between sleep and excitement; and the dominance, or level of control of the emotion state, from submissive to dominant.

The ability to automatically recognise human emotions is an interesting - yet challenging - issue spreading accross several fields, including, but not limited to: human-computer interaction (HCI), healthcare, education, and multimedia experience [10] [11]. As a matter of fact, there has been a growing interest in automatic emotion recognition over the last few years as emotions not only play an important role in human relationships, but also in how they interact with computers and applications. Automatic facial expression recognition (FER) has been widely investigated as it allows a non-invasive analysis of emotions, compared to the measurement of physiological signals for instance, and it has been made possible with the advent of image processing technologies. Indeed, conventional FER approaches are based on different steps, including image pre-processing, face and landmark detection, feature extraction and selection, and classification [12]. Recent FER algorithms have been using deep learning, or more precisely convolutional neural networks (CNN) [13]. Existing models, based on the seven basic emotions, usually propose a binary classification (i.e., correct *vs.* incorrect) of expressions as output.

In the particular context of interactive media experiences, emotions can be taken into consideration in a wide variety of ways and at different levels. Among several examples, the *RIOT* prototype can be cited, an immersive and emotionally responsive live-action film using FER to put users in a riot situation [14]. More subtleness in the emotion classification can be needed as even humans make mistakes in recognising expressions from their peers. Face expression recognition is indeed one of the most challenging tasks in social communication [15]. Some errors can be considered more important than others in terms of emotion classification - and that both for humans and machines. For instance, the gap between happiness and surprise is manifestly smaller than the gap between happiness and sadness.

In this paper, we firstly present a benchmark of four open source facial expression recognition (FER) libraries, i.e., DeepFace, EmoPy, Py-FEAT, and RMN, on four publicly available annotated image datasets, i.e., JAFFE, FER-2013, RAF-DB, and AffectNet, presented in the next sections. Using, in the first instance, classical classification metrics, we compare the performance of these algorithms. Secondly, we propose a new metric which aim is to improve the current binary classification of the results for the particular context of interactive and immersive media experience. Finally, we open a discussion on the obtained results, as well as on the ground truth of existing datasets.

Table 1. Summary of the datasets used in our benchmark.

	JAFFE [16] [17]	**FER-2013** [18]	**RAF-DB** [19] [20]	**AffectNet** [21]
Dataset size	213 images	35.887 images	29.672 images	1 million images
Benchmark sample size	213 images	7178 images	6134 images	3500 images
• Angry	30 images	958 images	162 images	500 images
• Disgust	29 images	111 images	160 images	500 images
• Fear	32 images	1024 images	74 images	500 images
• Happy	31 images	1774 images	1185 images	500 images
• Surprise	30 images	831 images	329 images	500 images
• Sad	31 images	1247 images	478 images	500 images
• Neutral	30 images	1233 images	679 images	500 images
Image type	Greyscale	Greyscale	Colour	Colour
Condition	Posed	Wild	Wild	Wild

2 PRESENTATION OF THE DATASETS

Table 1 summarises the four annotated facial expression databases we used for our benchmark, i.e., JAFFE, FER-2013, RAF-DB, and AffectNet, which are further described in the following subsections.

2.1 JAFFE

The Japanese Female Facial Expression (JAFFE) dataset consists of 213 256x256 greyscale images of different facial expressions posed by ten Japanese female subjects [16] [17]. Each subject was asked to freely pose seven facial expressions (i.e., the six basic emotions, and a neutral or expressionless face). Images were annotated using the average semantic ratings given by sixty human observers on seven distinct 5-level Likert scales (i.e., one for each emotion). Figure 1 is a sample of seven images (i.e., one for each emotion) which can be found in the JAFFE dataset. Human accuracy was estimated around 81% over all poses of the JAFFE dataset [16] [17]. However, this dataset exhibits cultural specificity, as proven in a cross-cultural study conducted with Japanese and American annotators (the average agreement of the American group was overall significantly smaller than of the Japanese group on some "Anger" and "Disgust" poses) [22].

Fig. 1. Examples of images from JAFFE (from left to right: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutrality).

2.2 FER-2013

The Facial Expression Recognition 2013 (FER-2013) dataset was first introduced at the *International Conference on Machine Learning* [18]. It consists of 48x48 pixel greyscale images of faces, which were automatically registered so that the face is more or less centred and occupies about the same amount of space in each image. Images were collected thanks to a Google search. The final dataset is composed of 35887 images, with 4953 "Anger" images, 547 "Disgust" images, 5121 "Fear" images, 8989 "Happiness" images, 6077 "Sadness" images, 4002 "Surprise" images, and 6198 "Neutral" images. The test set contains 7,178 examples. Figure 2 is a sample of seven images which can be found in the FER-2013 dataset. Human accuracy was estimated around 65.5% on this dataset [18].

Fig. 2. Examples of images from FER-2013 (from left to right: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutrality).

2.3 RAF-DB

The Real-world Affective Faces Database (RAF-DB) contains approximately 30.000 facial images with uncontrolled poses and illumination [19] [20]. All the images were obtained from the Internet. It is a very diverse dataset in terms of age, gender, and ethnicity, but also of lighting, occlusions (e.g., glasses), etc. Each image was manually

annotated by about 40 individuals using crowdsourcing. Annotators were asked to classify each image in one of the seven basic emotion classes. Figure 3 is a sample of seven images which can be found in RAF-DB.

Fig. 3. Examples of images from RAF dataset (from left to right: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutrality).

2.4 AffectNet

AffectNet contains about one million of facial images collected from the Internet by querying three major search engines in six different languages [21]. It is, so far, the largest available in-the-wild facial expression database. About half of the retrieved images were manually annotated using a categorical model (i.e., seven discrete facial expressions). The rest of the images were automatically annotated using a ResNeXt neural network trained on all manually annotated training set samples with average accuracy of 65%. However, as the full AffectNet dataset is huge, only a small version was released, containing 291,651 images manually annotated with eight labels (i.e., 0: Neutral, 1: Happiness, 2: Sadness, 3: Surprise, 4: Fear, 5: Disgust, 6: Anger, 7: Contempt). The test set is not released at this moment; we therefore used the validation set, which contains 3,999 images. Figure 4 is a sample of eight images which can be found in AffectNet.

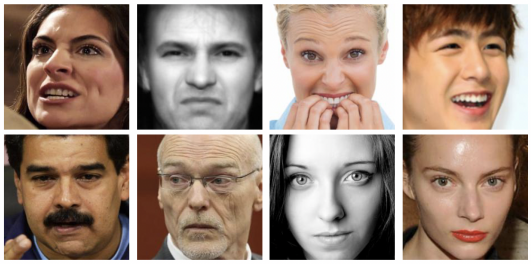


Fig. 4. Examples of images from AffectNet (from left to right, and from top to bottom: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutrality, and Contempt).

3 PRESENTATION OF THE OPEN SOURCE LIBRARIES

In the context of interactive media projects, open source libraries represent a low cost tool that simplifies access to a computational solution. For projects requiring facial expression analysis from images or videos, the major tool is the emotion interpreter, that is the system able to translate the emotion from the analysed frame. In a computational domain, facial expression recognition has improved with the emergence of neural network models trained on large databases [23]. Several open source solutions offer access to such models as a plug and play solution. The four libraries used for our benchmark, i.e., DeepFace, EmoPy, Py-FEAT, and RMN, are further described in the following subsections.

3.1 DeepFace

DeepFace is a lightweight Python library that offers a range of models for facial recognition and attribute analysis [24]. This library offers as a convolutional neural network (CNN) pre-trained on the FER-2013 training set, which obtained an accuracy of 57%. It is able to classify images between seven possible emotions.

3.2 EmoPy

EmoPy is an open-source Python library that contains different pre-trained neural network architectures to be applied to FER projects [25]. Those pre-trained models are CNNs that can be applied to specific cases with a reduced set of emotions, as well as in cases where the classical seven emotions are considered.

3.3 Py-FEAT

Py-FEAT is an open-source package for facial expressions research written in Python [26]. It includes tools to detect faces, extract emotional facial expressions, facial muscle movements, and facial landmarks, from videos and images of faces, as well as methods to pre-process, analyse, and visualise facial expression data.

3.4 RMN

RMN is a Python library that offers access to a pre-trained residual mask network (RMN) that can be used to predict emotions on videos and static images [27]. This model can also be accessed *via* the Py-FEAT library.

4 PRESENTATION OF THE BENCHMARK

To analyse the performance of the pre-trained models presented in Section 3, a sample of images containing expressions associated to one of the seven basic emotions was selected from each dataset, as detailed in Table 1. A set of classical statistical metrics - calculated by comparing the ground truth and predicted labels for a given image - were used, with two different (yet complementary) approaches. For the first approach, taking into account the existence of seven classes, i.e., one for each basic emotion, we considered the macro precision, recall, and F1-score, as well as Cohen's Kappa and accuracy. To further compare the classification performance on each class, and to investigate misclassification, normalised confusion matrices were generated. On the other hand, the second approach only considers the existence of two classes in the calculus of the metrics, i.e., one representing the analysed emotion and the other representing the absence of the analysed emotion. Using this approach, precision and recall were calculated for each emotion.

4.1 Benchmark on JAFFE images

Table 2. Performance of common classification metrics on JAFFE.

Metric	DeepFace	RMN	Py-FEAT	EmoPy
Accuracy	0.474	**0.479**	0.469	0.207
Precision	0.430	0.575	**0.601**	0.149
Recall	0.471	**0.476**	0.467	0.206
κ	0.386	**0.392**	0.381	0.074
F1-Score	0.429	**0.429**	0.428	0.141

Table 2 presents the results of all libraries on JAFFE images. RMN obtained the highest accuracy, recall, and Cohen's Kappa; while Py-FEAT obtained the highest precision. Lowest values for all metrics are associated with EmoPy. Normalised confusion matrices, illustrated in Figure 6, show that the four models had difficulties

on correctly predicting anger, disgust, and fear expressions; whereas they performed better with happiness, neutrality, sadness, and surprise. DeepFace accomplished recall values greater than 70% on the prediction of images containing happy, neutral, and surprised expressions. EmoPy only had a recall above 70% for happiness (recall was below 50% for the other emotions).

Figure 7 shows a comparison between the models that achieved the best performance, i.e., DeepFace and Py-FEAT. In terms of precision, Py-FEAT and DeepFace obtained the highest precision for different emotions, with a highlight for angry expressions on Py-FEAT results. Py-FEAT tended to only classify as angry images containing anger expressions, while DeepFace also classified in this category images containing disgust expressions. In terms of recall, Py-FEAT correctly classified images containing sad expressions. On the contrary, images containing anger, disgust, and fear expressions tended to be misclassified. DeepFace correctly classified images containing happy, neutral, and surprised expressions. However, it had difficulties classifying images with sadness, anger, fear, and disgust.

4.2 Benchmark on FER-2013 images

Table 3. Performance of common classification metrics on FER-2013.

Metric	DeepFace	RMN	Py-FEAT	EmoPy
Accuracy	**0.554**	0.506	0.503	0.207
Precision	**0.555**	0.457	0.446	0.149
Recall	**0.522**	0.475	0.465	0.206
κ	**0.460**	0.410	0.404	0.074
F1-Score	**0.535**	0.445	0.444	0.141

Table 3 summarises the results obtained on the FER-2013 sample. Figure 8 illustrates the comparison of ground truth and predicted labels on each image of the FER-2013 sample. It can be quickly observed that happy expressions tended to be more accurately recognised than any other emotion. Contrarily, disgust and fear expressions were poorly recognised by the models used in our benchmark. In terms of performance metrics, presented in Table 3, DeepFace performed better than the other three libraries on this sample. This result is also expressed by the main diagonal of its confusion matrix, showing sensitivity values greater than 40% for each emotion. DeepFace easily recognised happiness and surprise.

Figure 9 represents a comparative view of the precision and recall per emotion, for DeepFace and Py-FEAT. DeepFace performed with a better precision and recall for almost all studied emotions, while Py-FEAT had a higher recall for angry expressions. In terms of precision, as it can be observed on the confusion matrices, Py-FEAT had more than 40% of accuracy for six emotions out of seven, with a precision greater than 70% for happiness and surprise. Yet, both models poorly performed on expressions of fear, with a precision lower than 30% for DeepFace, and than 20% for Py-FEAT. Regarding recall, an interesting point is the behaviour observed on the residual mask model implemented in Py-FEAT and RMN: these models obtained the highest tax on correctly classifying anger, generating the highest recall.

4.3 Benchmark on RAF-DB images

As presented in Table 4, the residual mask model implemented in Py-FEAT and RMN obtained the highest performance on RAF-DB. In this context, the residual masking model was able to have the highest rates of correct classifications. Figure 10 illustrates the confusion matrices generated. It can be seen that images containing happy, angry, and sad expressions were better identified compared to the other emotions. For EmoPy and Py-FEAT, the

Table 4. Performance of common classification metrics on RAF-DB.

Metric	DeepFace	RMN	Py-FEAT	EmoPy
Accuracy	0.507	**0.662**	0.630	0.347
Precision	0.422	**0.528**	0.521	0.269
Recall	0.345	**0.553**	0.530	0.279
κ	0.347	**0.567**	0.531	0.184
F1-Score	0.336	**0.519**	0.495	0.222

highest rate of correct predictions are associated to anger. For DeepFace and RMN, the highest rate is associated to happy expressions.

Figure 11 presents a comparative view of the precision and recall calculated for DeepFace and RMN. It can be seen from Figure 11a that RMN had a higher precision for images containing sad, neutral, and happy expressions, while DeepFace obtained the highest precision for surprise. Both models performed poorly on the classification of fear expressions. In terms of recall, illustrated on 11b, RMN obtained the highest values, with a highlight for sadness, happiness, anger, and neutrality. Both models performed poorly on disgust expressions.

4.4 Benchmark on AffectNet images

Table 5. Performance of common classification metrics on AffectNet.

Metric	DeepFace	RMN	Py-FEAT	EmoPy
Accuracy	0.351	**0.575**	0.552	0.256
Precision	0.415	**0.599**	0.589	0.251
Recall	0.351	**0.575**	0.552	0.256
κ	0.243	**0.504**	0.477	0.132
F1-Score	0.314	**0.573**	0.550	0.223

In terms of statistical metrics, presented in Table 5 for AffectNet, the residual mask model obtained higher Cohen's Kappa and accuracy values compared to EmoPy and DeepFace. From Figure 12, one can see that images with truly happy expressions were easily identified compared to the other emotions. Images containing disgust expressions were indeed badly identified by each library.

Using the one *vs.* all approach, libraries with the best performance were evaluated, as presented in Figure 13. Precision values found for Py-FEAT and DeepFace were similar for surprised and disgust images. However, for the other emotions, highest precision values were found by Py-FEAT. It is interesting to observe a high precision value for disgust and fear expressions, despite a low number of correct classifications for images labelled with these emotions. Such precision results are due to the low number of other expressions classified as these emotions. The lowest precision was found for neutral expressions, despite a high rate of correct classifications (many expressions were misclassified as neutral). As far as recall is concern, the higher performance was found on happy images for DeepFace and Py-FEAT. Py-FEAT obtained higher recall values for the other emotions. Lowest recall is linked to disgust and fear for both libraries.

5 A NEW METRIC TO ANALYSE FACIAL EXPRESSION RECOGNITION ALGORITHMS' PERFORMANCE

Although several metrics have been proposed to evaluate the performance of facial expression recognition (FER) algorithms, the latter are mainly based on a binary (i.e., true *vs.* false) classification of the results. To address this limitation, we propose a new metric which evaluates the error between ground truth and predicted emotion.

5.1 Proposed metric

The valence-arousal-dominance (VAD) model [28] is built on three independent dimensions, i.e., valence (pleasure), ranging from unhappiness to happiness; arousal (or affective activation), ranging between sleep and excitement; and dominance (level of control of the emotion state), from submissive to dominant. Table 6 represents the values of valence, arousal, and dominance to represent each of the six basic emotions, i.e., Anger, Disgust, Fear, Happiness, Surprise, and Sadness, as defined by Russell & Mehrabian [28]. We further added Neutrality as origin of the coordinate system.

Table 6. Valence, arousal, and dominance values of the seven basic emotions.

	Valence	Arousal	Dominance
Anger	−0.43	0.67	0.34
Disgust	−0.60	0.35	0.11
Fear	−0.64	0.60	−0.43
Happiness	0.76	0.48	0.35
Surprise	0.40	0.67	−0.13
Sadness	−0.63	0.27	−0.33
Neutrality	0.00	0.00	0.00

Let $\mathcal{E}$ be the set of emotions e, i.e., the six basic emotions plus the neutral state. Let $P \in \mathbb{R}^7$ be the FER model output, such that $p_e \geq 0 \ \forall e \in \mathcal{E}$, and $\sum_{\forall e \in \mathcal{E}} p_e = 1$. In this context, let (v_e, a_e, d_e) be the value of valence, arousal, and dominance for each emotion e in the emotion set $\mathcal{E}$. Using the probabilities and ground truth values presented in Table 6, the predicted valence, arousal, and dominance for a given emotion can easily be calculated as follows.

$$\hat{v} = \sum_{\forall e \in \mathcal{E}} p_e \cdot v_e \qquad \hat{a} = \sum_{\forall e \in \mathcal{E}} p_e \cdot a_e \qquad \hat{d} = \sum_{\forall e \in \mathcal{E}} p_e \cdot d_e$$

To measure the prediction error for a given image, the Euclidean distance between $(\hat{v}, \hat{a}, \hat{d})$, representing the prediction in the VAD space, and (v_{gt}, a_{gt}, d_{gt}), corresponding to the ground truth label of the image, was calculated as in (1).

$$z = \sqrt{(\hat{v} - v_{gt})^2 + (\hat{a} - a_{gt})^2 + (\hat{d} - d_{gt})^2} \tag{1}$$

Euclidean distances between the ground truth emotions' representation on the VAD model were calculated to introduce a comparative perspective on our analysis. Such values, presented in Figure 5, show the existence of similarities and dissimilarities between the emotions on this model. For example, sadness and fear are the most similar emotions, with the smallest distance (i.e., 0.34) between their respective points in the VAD model compared to the distance between other emotions. On the contrary, happiness and fear are the less similar, as the distance between their respective points in the VAD model is the largest (i.e., 1.51).

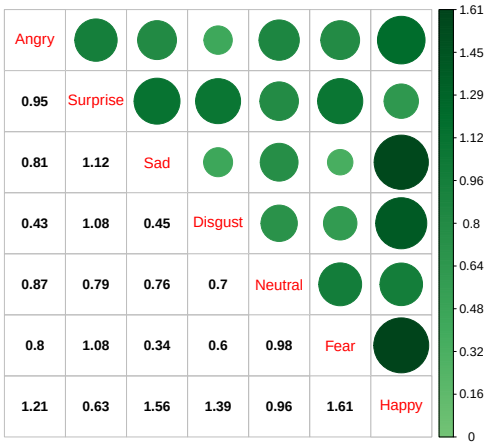


Fig. 5. Euclidean distances between the valence, arousal, and dominance points for each pair of basic emotions.

The following subsections present the univariate analyses of the distance distributions calculated for each model on each dataset. A discussion on the central tendency and variability metrics calculated for the unconditional distribution of the distances calculated for the predictions of each model is firstly presented. Secondly, a discussion on the same metrics is raised, although, in this case, the distance distribution conditioned by the ground truth emotions is considered.

5.2 Distance analysis on JAFFE images

For the JAFFE images, the central tendency and variability metrics, presented in Table 7, indicate that RMN predictions are closer to the ground truth values, as the distribution obtained for this model has the smallest mean and median and the highest positive skewness. Such metrics also indicate that EmoPy produced predictions far from the ground truth emotions, as the distribution for this model has the smallest skewness and the largest mean and median values. Based on this result, RMN is the most accurate model while EmoPy was the less accurate.

Based on the distribution of the distances for each emotion, represented on Figure 14, distances associated with images containing happy and neutral expressions are mostly concentrated near to the first quartile calculated for the distances found using DeepFace, Py-FEAT and RMN. On Py-FEAT and RMN, images containing sad expressions are also associated with distances in the mentioned region. The model with the lowest performance, i.e, EmoPy had the majority of distances situated after the median calculated for this model.

Table 7. Measures of central tendency and variability of the distance distribution on JAFFE.

	DeepFace	RMN	Py-FEAT	EmoPy
Mean	0.384	**0.374**	0.387	0.759
Standard deviation	0.322	**0.299**	0.314	0.357
First quartile	**0.048**	0.079	0.094	0.547
Median	0.359	**0.333**	0.334	0.783
Third quartile	0.654	**0.605**	0.650	1.035
Variance	0.104	**0.089**	0.099	0.128
Skewness	0.478	**0.448**	0.520	-0.329

Reciprocally, the distributions show the existence emotions in which the model predictions are far from their respective ground truth values. For example, the images labeled as containing angry or disgusted expressions had their predictions distances mostly concentrated after the median value calculated for DeepFace, Py-FEAT and RMN.

5.3 Distance analysis on FER-2013 images

For FER-2013 images, the central tendency and variance metrics, displayed in Table 8, indicate that DeepFace predictions are closer to the ground truth values, as the distribution obtained for this model has the smallest mean and median and the highest positive skewness. Such metrics also indicate that EmoPy produced predictions far from the ground truth, as the distribution for this model has the lowest skewness, with the highest mean and median values. Such analysis matches the results found via classical metrics, as it showed DeepFace as best model, and EmoPy as least accurate.

Based on the distribution of the distances for each emotion, represented on Figure 15, distances associated with images containing happy expressions are mostly concentrated near the first quartile calculated for the distances found using DeepFace, Py-FEAT and RMN. The model with the lowest performance, i.e, EmoPy has most of its distributions per emotion concentrated before the distance median value.

Reciprocally, the distributions show the existence of emotions for which model predictions are far from their respective ground truth values. For example, images labeled as containing disgust or fear expressions had their prediction distances mostly concentrated after the median value calculated for DeppFace, Py-FEAT, and RMN. For DeepFace the distance associated with images labeled as containing angry expressions are also situated before the median value.

Table 8. Measures of central tendency and variability of the distance distribution on FER-2013.

	DeepFace	RMN	Py-FEAT	EmoPy
Mean	**0.392**	0.420	0.425	0.572
Standard deviation	**0.416**	0.339	0.339	0.364
First quartile	**0.011**	0.084	0.095	0.284
Median	**0.253**	0.385	0.392	0.594
Third quartile	**0.724**	0.697	0.695	0.802
Variance	**0.173**	0.115	0.115	0.133
Skewness	**0.900**	0.476	0.498	0.129

5.4 Distance analysis on RAF-DB images

On RAF-DB images, the central tendency and variability metrics presented in Table 9 indicate that RMN predictions are closer to their respective ground truth values, as the mean and median values are the lowest in comparison to the ones calculated for distances obtained with the other models. Such metric also indicates that EmoPy produced predictions that are far from the ground truth, as distances generate a distribution with the highest mean and median values, and lowest skewness. Results found via the distance metric match the ones found via classical metrics.

On Figure 16, distances associated with images containing happy and neutral expressions are mostly situated before the median for both DeepFace and RMN models. Distances calculated for fear and surprise images are mostly concentrated after the third quartile of the distances calculated for DeepFace predictions. For the RMN model, disgust and fear expressions are largely situated after the median value. With these results, and based on Figure 16, happiness and neutrality are easily recognised by both models. Sadness and anger are also easily recognised by RMN.

Distributions show the existence of emotions for which the model predictions are far from their respective ground truth values. In this context, surprise and disgust are mostly concentrated after the median value of the distance calculated for Py-FEAT and RMN.

Table 9. Measures of central tendency and variability of the distance distribution on RAF-DB.

	DeepFace	RMN	Py-FEAT	EmoPy
Mean	0.472	**0.332**	0.342	0.618
Standard deviation	0.473	**0.342**	0.342	0.365
First quartile	0.021	**0.039**	0.042	0.332
Median	0.348	**0.215**	0.247	0.649
Third quartile	0.786	**0.550**	0.576	0.879
Variance	0.224	**0.117**	0.117	0.133
Skewness	0.768	**1.058**	0.948	-0.017

5.5 Distance analysis on AffectNet images

On AffectNet images, the centrality and variability metrics, presented in Table 10, indicate that RMN obtained the predictions that are closer to their respective ground truth values, as the mean and median values are the lowest in comparison to the ones calculated for the distances obtained with the other models. Such metric also indicates that EmoPy produced predictions are far from their respective ground truth, as the distances calculated generate a distribution with the highest mean and median values, as well the lowest skeweness. Results found via the distance metric on AffectNet shows that the best performance on classification metric (Table 5) is associated with models that predicted points near their respective ground truth positions on the VAD space.

The distance distribution calculated for each emotion, illustrated in Figure 17, shows that images containing happy expressions are associated with the highest concentration of distances in a region below the first quartile found for DeepFace, Py-FEAT, and RMN. Other facial expressions have distribution concentrated in this region. As an example, the neutral expressions and, only for Py-FEAT and RMN model, anger expressions. The model with the lowest performance on AffectNet, i.e., EmoPy, has most of its distance distributions per emotion situated before the median calculated for the distances found using this model predictions.

Table 10. Measures of central tendency and variability of the distance distribution on AffectNet.

	DeepFace	RMN	Py-FEAT	EmoPy
Mean	0.530	**0.351**	0.362	0.591
Standard deviation	0.398	**0.322**	0.328	0.320
First quartile	0.162	**0.053**	0.056	0.363
Median	0.531	**0.280**	0.296	0.610
Third quartile	0.791	**0.581**	0.613	0.805
Variance	0.159	**0.104**	0.107	0.102
Skewness	0.427	**0.803**	0.776	0.037

Distributions show the existence of emotions for which model predictions are far from their ground truth values, as their distances are mostly concentrated on the region after the median. Using DeepFace, images containing disgust, anger, surprise, and fear expressions had their predicted distances mostly concentrated after the median. For RMN and Py-Feat, fear and disgust expressions had their predicted distances mostly concentrated after the median.

6 DISCUSSION

In spite of the rapid growth of artificial intelligence technology, reading people's emotions using FER algorithms still appears extremely complicated for some emotions, as demonstrated in Sections 4 and 5. To be used, and then trusted, these algorithms firstly need to be reliable. Furthermore, while certain (e.g., commercial) systems perform well under controlled conditions (i.e., with high resolution images and frontal faces), they usually face difficulties with real world applications, like realistic distortions [29] [30].

According to [31], classical classification models fail on capturing fine-grained differences in dynamic expressions. A possible solution, suggested by the authors, is therefore to move the FER problem from a classification problem to a regression one, where the model prediction corresponds to a point in the valence-arousal-dominance (VAD) space. In this context, our approach creates a link between the classification result and the VAD space, as it directly maps the discrete classification into this space. We also defined a useful metric to evaluate the regression model on this domain, i.e., the distance to ground truth values.

Current publicly available datasets, such as JAFFE, FER-2013, RAF-DB, and AffectNet, only consider a single value (i.e., discrete emotion) as ground truth. However, in some cases, even human annotators do not agree when labeling images; ambiguities on the ground truth remain [32]. Therefore, to further explore the relationship presented above, it would be interesting to have new datasets where ground truth labels are represented in the VAD space, or as percentages of the seven basic emotions instead of only one emotion.

7 CONCLUSION

In this paper, we carried out a benchmark involving four facial expression recognition (FER) algorithms on four datasets composed of head shots. All the datasets used are publicly available, and labeled with the seven basic emotions. The use of common performance metrics, e.g., F1-score and confusion matrices, yielded two distinct observations. Firstly, existing open source libraries for automatic FER do not seem to be good enough (as their overall accuracy remains below 50%) to be reliable. Secondly, current metrics to analyse the performance of such algorithms remain limited as they only consider a binary classification of the results.

We therefore proposed a new metric based on the Euclidean distance between ground truth and predicted emotions in the valence-arousal-dominance space. With this metric, we aimed to provide new information on the output of different FER models for interactive media, where human sensibility needs to be considered.

ACKNOWLEDGMENTS

This work is financially supported by West Creative Industries[1], a cluster in the French Pays de la Loire region.

Special thanks go to Thomas Choquard and Allan Ghazhali, Master students at Nantes University, who helped starting this work during their final year project.

REFERENCES

[1] Géraldine Coppin and David Sander. Theoretical approaches to emotion and its measurement. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 3–30. Woodhead Publishing, 2016. ISBN 978-0-08-100508-8. doi: https://doi.org/10.1016/B978-0-08-100508-8.00001-1. URL https://www.sciencedirect.com/science/article/pii/B9780081005088000011.

[2] Rosalind Picard. *Affective Computing*. The MIT Press, 1997.

[3] Albert Mehrabian. Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation*, 1(6):203–207, 1968.

[4] Charles Darwin. *The expression of the emotions in man and animals*. New York ;D. Appleton and Co.„ 1916. URL https://www.biodiversitylibrary.org/item/24064. https://www.biodiversitylibrary.org/bibliography/4820.

[5] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068. URL https://doi.org/10.1080/02699939208411068.

[1] https://www.westcreativeindustries.org/

[6] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[7] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27:46, 1997.

[8] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[9] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[10] C Vinola and K Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 14(2):24–44, 2015.

[11] Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michal R Wrobel. Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 51–62. Springer, 2014.

[12] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2), 2018. ISSN 1424-8220. doi: 10.3390/s18020401. URL https://www.mdpi.com/1424-8220/18/2/401.

[13] Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions, 2017.

[14] Karen Palmer. Riot, 2018. URL https://thoughtworksarts.io/projects/riot/.

[15] I Michael Revina and WR Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, 33(6):619–628, 2021.

[16] Michael Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998. doi: 10.1109/AFGR.1998.670949.

[17] Michael Lyons. "excavating AI" re-excavated: Debunking a fallacious account of the JAFFE dataset, 2021.

[18] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2014.09.005. URL https://www.sciencedirect.com/science/article/pii/S0893608014002159. Special Issue on "Deep Learning of Representations".

[19] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. doi: 10.1109/CVPR.2017.277.

[20] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. doi: 10.1109/TIP.2018.2868382.

[21] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. doi: 10.1109/TAFFC.2017.2740923.

[22] Matthew N Dailey, Carrie Joyce, Michael J Lyons, Miyuki Kamachi, Hanae Ishi, Jiro Gyoba, and Garrison W Cottrell. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, 10(6):874, 2010.

[23] Wafa Mellouk and Wahida Handouzi. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.07.101. URL https://www.sciencedirect.com/science/article/pii/S1877050920318019. The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC),The 15th International Conference on Future Networks and Communications (FNC),The 10th International Conference on Sustainable Energy Information Technology.

[24] Sefik Serengil. Deepface: The most popular open-source facial recognition library, 2021. URL https://viso.ai/computer-vision/deepface/.

[25] Angelica Perez. Emopy: A machine learning toolkit for emotional expression, 2018. URL https://www.thoughtworks.com/insights/blog/emopy-machine-learning-toolkit-emotional-expression.

[26] Jin Hyun Cheong, Tiankang Xie, Sophie Byrne, and Luke J. Chang. Py-feat: Python facial expression analysis toolbox, 2021.

[27] Pham Luan, Vu Huynh, and Tran Tuan Anh. Facial expression recognition using residual masking network. In *IEEE 25th International Conference on Pattern Recognition*, pages 4513–4519, 2020.

[28] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[29] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, 19(8):1863, 2019.

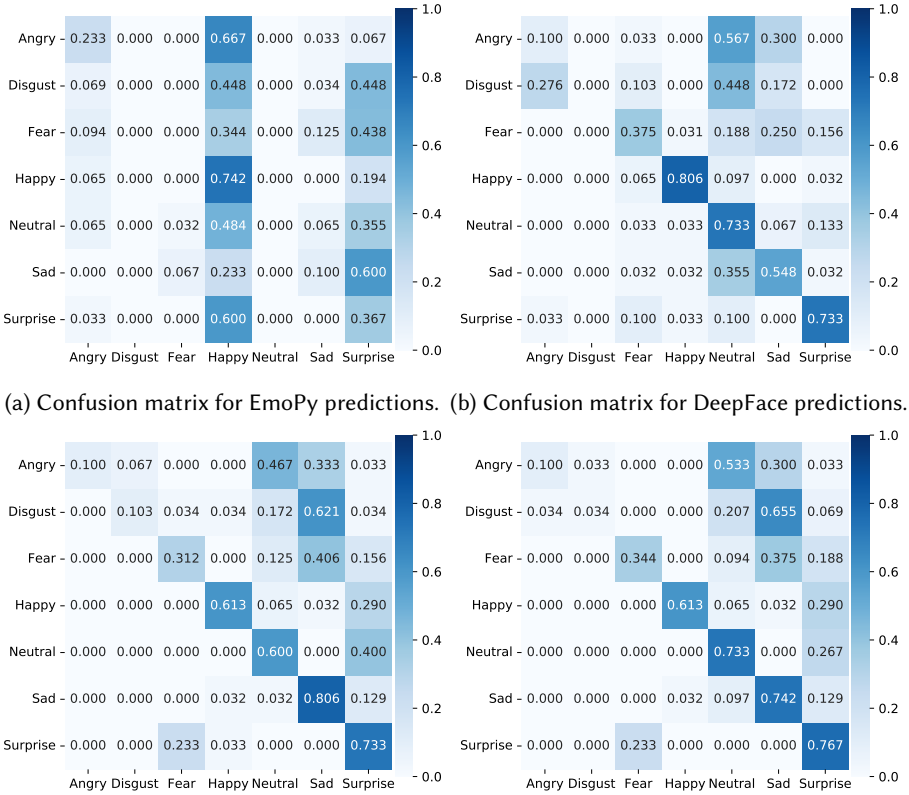
[30] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer*, 37(6):1447–1466, 2021.

[31] Feng Zhou, Shu Kong, Charless C. Fowlkes, Tao Chen, and Baiying Lei. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*, 392:38–49, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.01.067. URL https://www.sciencedirect.com/science/article/pii/S0925231220301193.

[32] William Saakyan, Olya Hakobyan, and Hanna Drimalla. Representational bias in expression and annotation of emotions in audiovisual databases. In *CAIP 2021: Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI, CAIP 2021, 20-24 November 2021, Bologna, Italy*, page 120. European Alliance for Innovation, 2021.

A APPENDIX: RESULTS OF THE BENCHMARK

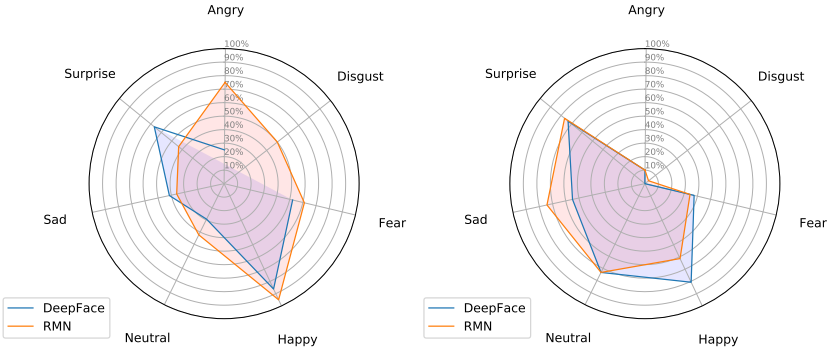
A.1 On JAFFE images



(a) Confusion matrix for EmoPy predictions. (b) Confusion matrix for DeepFace predictions.

(c) Confusion matrix for Py-FEAT predictions. (d) Confusion matrix for RMN predictions.

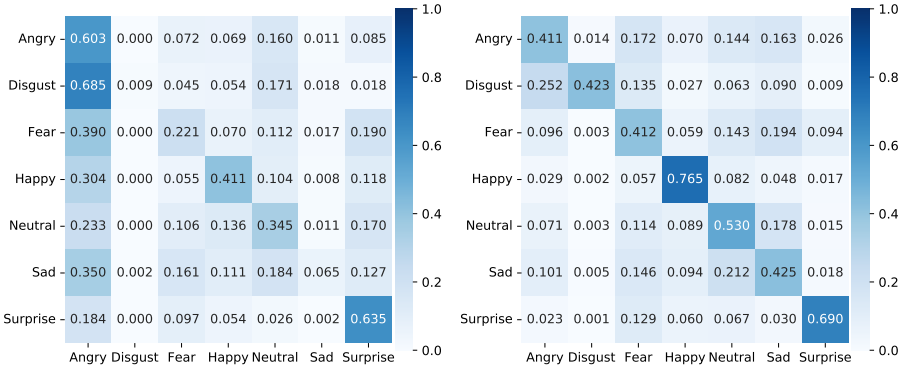
Fig. 6. Illustration of the confusion matrices calculated for the predictions of each library on the JAFFE sample.



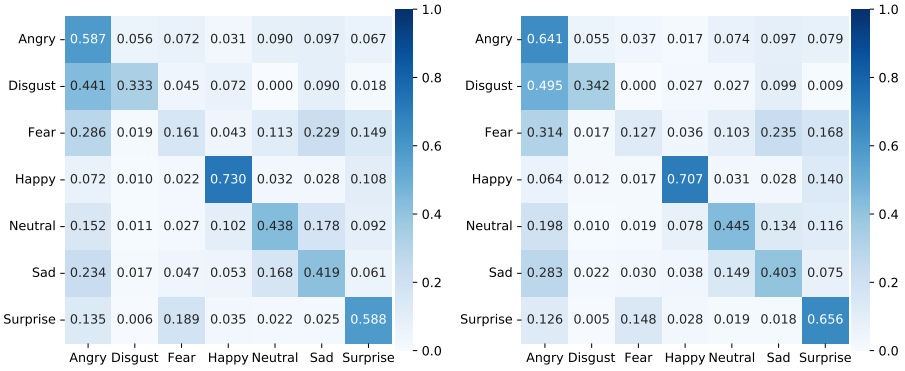
(a) Precision value for each emotion. (b) Recall value for each emotion.

Fig. 7. Illustration of the precision and recall calculated for each emotion on JAFFE images.

A.2 On FER-2013 images

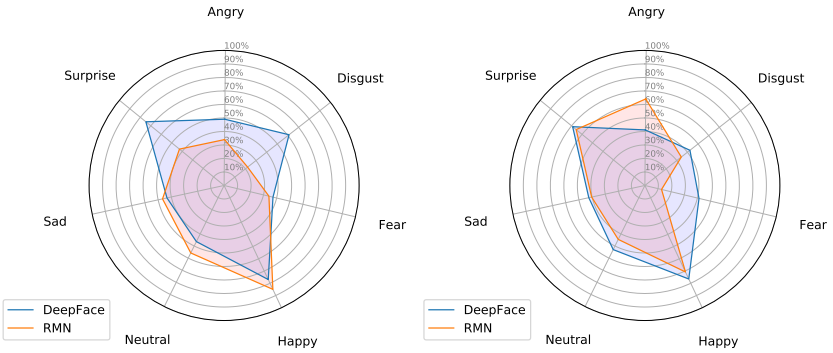


(a) Confusion matrix obtained with EmoPy. (b) Confusion matrix obtained with DeepFace.



(c) Confusion matrix obtained with PY-FEAT. (d) Confusion matrix obtained with RMN.

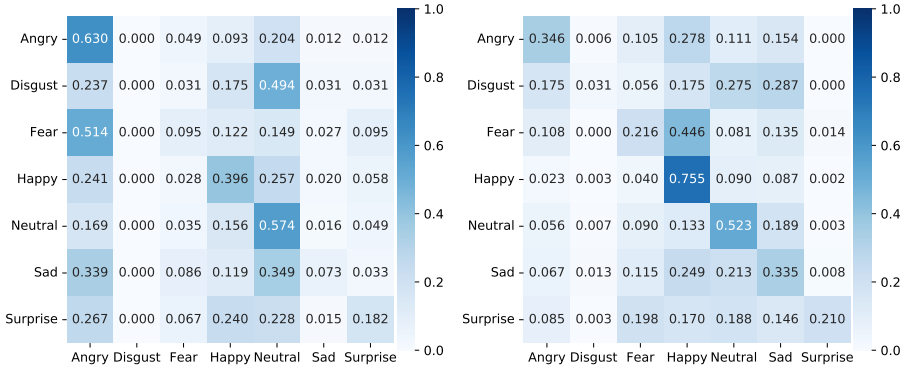
Fig. 8. Illustration of the confusion matrices calculated for the predictions of each library on the FER-2013 sample.



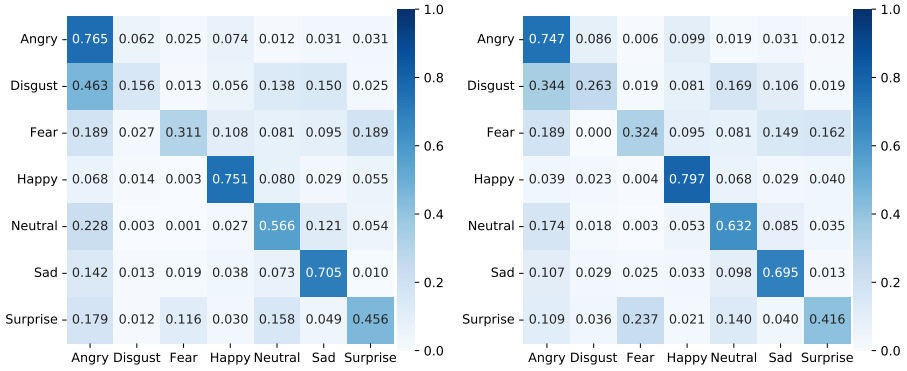
(a) Precision value for each emotion. (b) Recall value for each emotion.

Fig. 9. Illustration of the precision and recall calculated for each emotion on FER-2103 images.

A.3 On RAF-DB images

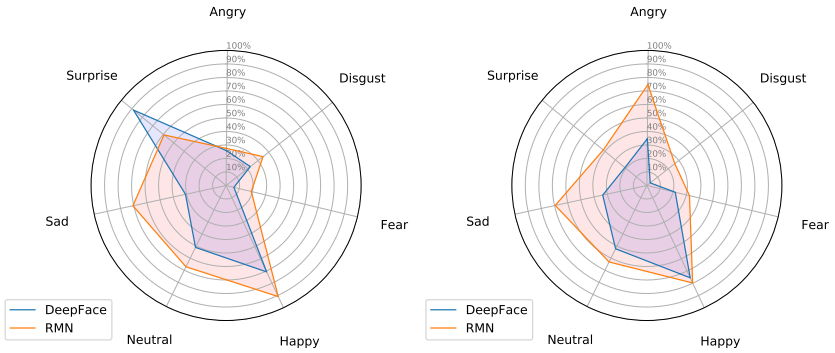


(a) Confusion matrix obtained with EmoPy. (b) Confusion matrix obtained with DeepFace.



(c) Confusion matrix obtained with PY-FEAT. (d) Confusion matrix obtained with RMN.

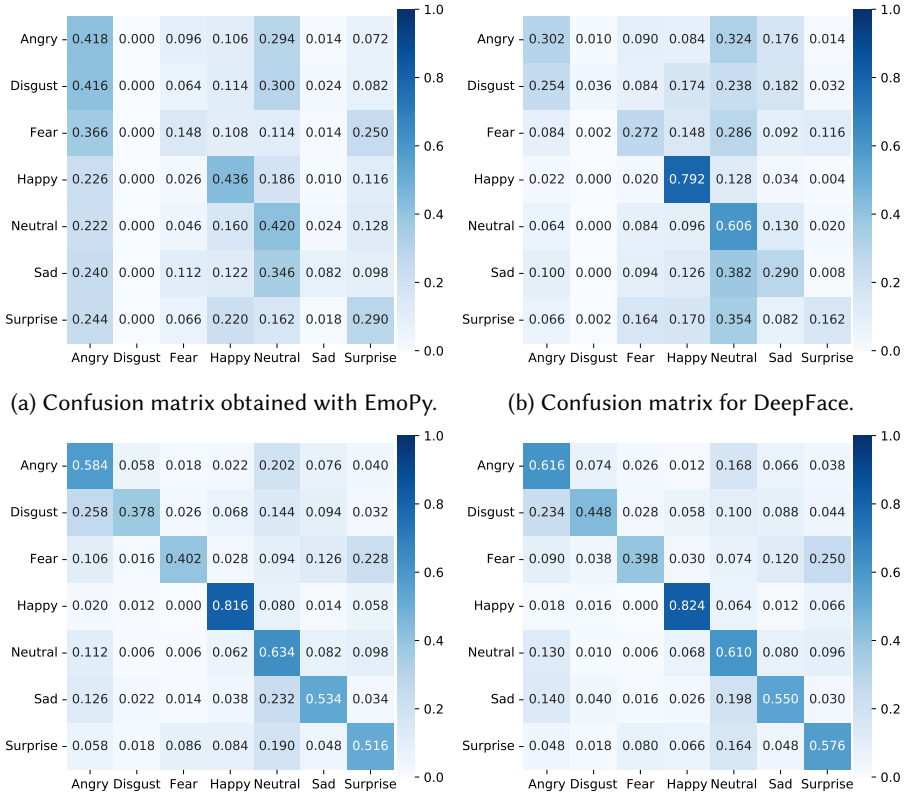
Fig. 10. Illustration of the confusion matrices calculated for the predictions of each library on the RAF-DB sample.



(a) Precision value for each emotion. (b) Recall value for each emotion.

Fig. 11. Illustration of the precision and recall calculated for each emotion on RAF-DB images.

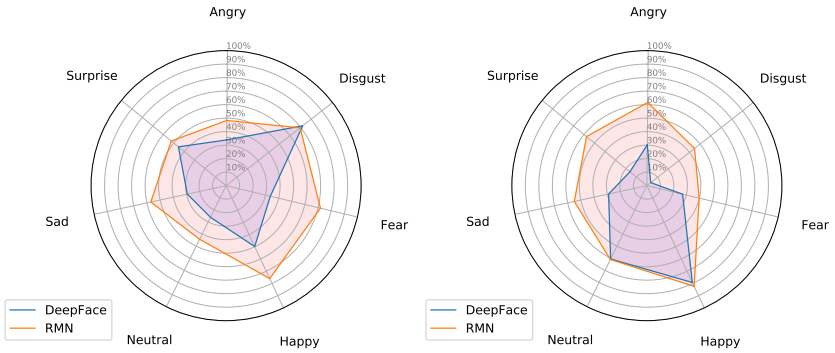
A.4 On AffectNet images



(a) Confusion matrix obtained with EmoPy. (b) Confusion matrix for DeepFace.

(c) Confusion matrix obtained with Py-FEAT. (d) Confusion matrix obtained with RMN.

Fig. 12. Illustration of the confusion matrices calculated for the predictions of each library on the AffectNet sample.

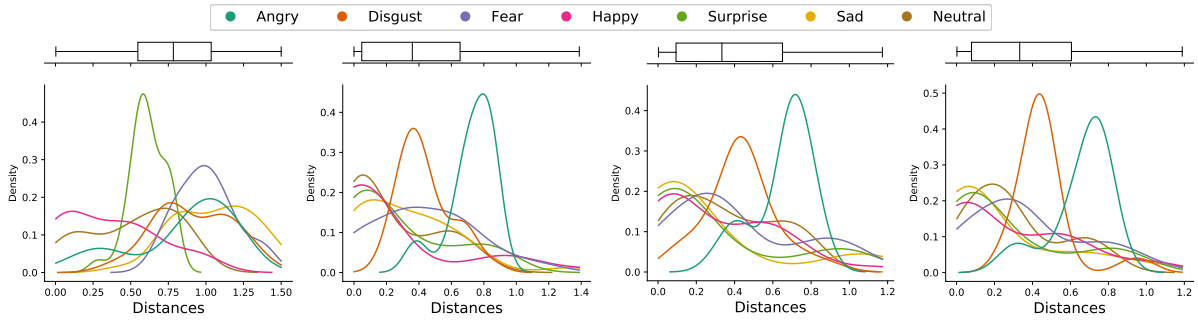


(a) Precision value for each emotion. (b) Recall value for each emotion.

Fig. 13. Illustration of the precision and recall calculated for each emotion on AffectNet images.

B APPENDIX: DISTANCE DISTRIBUTIONS

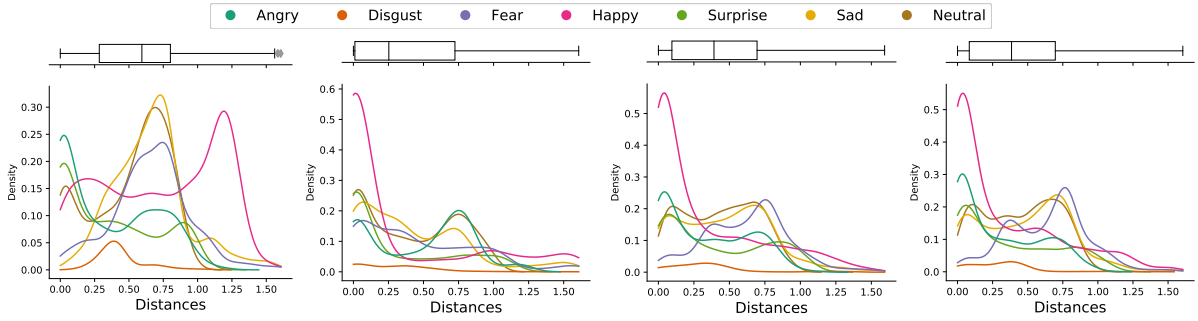
B.1 On JAFFE images



(a) Distances for EmoPy. (b) Distances for DeepFace. (c) Distances for Py-FEAT. (d) Distances for RMN.

Fig. 14. Illustration of the distance distribution for each emotion on JAFFE images.

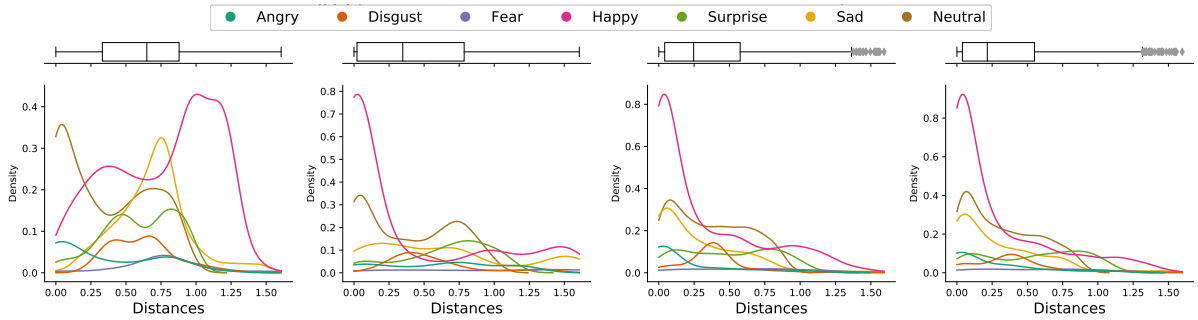
B.2 On FER images



(a) Distances for EmoPy. (b) Distances for DeepFace. (c) Distances for Py-FEAT. (d) Distances for RMN.

Fig. 15. Illustration of the distance distribution for each emotion on FER-2013 images.

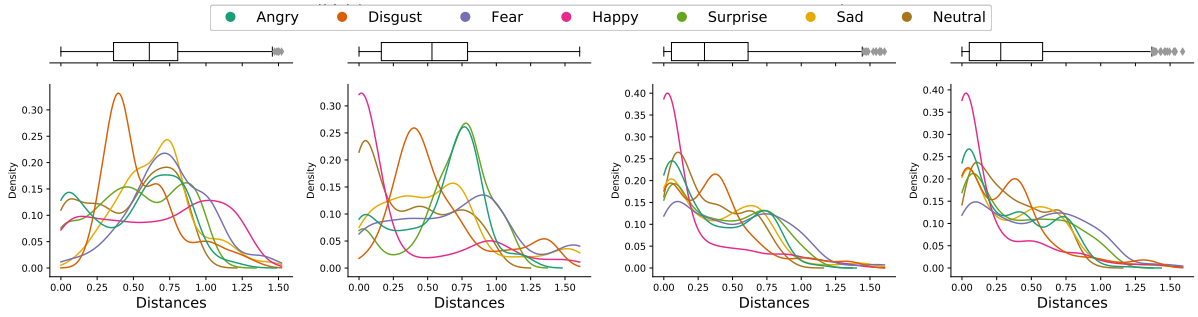
B.3 On RAF-DB images



(a) Distances for EmoPy. (b) Distances for DeepFace. (c) Distances for Py-FEAT. (d) Distances for RMN.

Fig. 16. Illustration of the distance distribution for each emotion on RAF-DB images.

B.4 On AffectNet images



(a) Distances for EmoPy. (b) Distances for DeepFace. (c) Distances for Py-FEAT. (d) Distances for RMN.

Fig. 17. Illustration of the distance distribution for each emotion on AffectNet images.