









# Towards a Better Understanding of Human Emotions: Challenges of Dataset Labeling

Hajer Guerdeli<sup>1,2</sup> , Claudio Ferrari<sup>3</sup> , Joao Baptista Cardia Neto<sup>4</sup> , Stefano Berretti<sup>1</sup> , Walid Barhoumi<sup>2,5</sup> , and Alberto Del Bimbo<sup>1</sup> 

<sup>1</sup> University of Florence, Florence, Italy

{hajer.guerdeli, stefano.berretti, alberto.delbimbo}@unifi.it

<sup>2</sup> University of Tunis El Manar, Tunis, Tunisia

<sup>3</sup> University of Parma, Parma, Italy

claudio.ferrari2@unipr.it

<sup>4</sup> FATEC Catanduva, Catanduva, Brazil

joao.cardia@fatec.sp.gov.br

<sup>5</sup> Université de Carthage, Ecole Nationale d'Ingénieurs de Carthage, Tunis, Tunisia

walid.barhoumi@enicarthage.rnu.tn

**Abstract.** A major challenge in automatic human emotion recognition is that of categorizing the very broad and complex spectrum of human emotions. In this regard, a critical bottleneck is represented by the difficulty in obtaining annotated data to build such models. Indeed, all the publicly available datasets collected to this aim are either annotated with (i) the six prototypical emotions, or (ii) continuous valence/arousal (VA) values. On the one hand, the six basic emotions represent a coarse approximation of the vast spectrum of human emotions, and are of limited utility to understand a person's emotional state. Oppositely, performing dimensional emotion recognition using VA can cover the full range of human emotions, yet it lacks a clear interpretation. Moreover, data annotation with VA is challenging as it requires expert annotators, and there is no guarantee that annotations are consistent with the six prototypical emotions. In this paper, we present an investigation aiming to bridge the gap between the two modalities. We propose to leverage VA values to obtain a fine-grained taxonomy of emotions, interpreting emotional states as probability distributions over the VA space. This has the potential for enabling automatic annotation of existing datasets with this new taxonomy, avoiding the need for expensive data collection and labeling. However, our preliminary results disclose two major problems: first, continuous VA values and the six standard emotion labels are often inconsistent, raising concerns about the validity of existing datasets; second, datasets claimed to be balanced in terms of emotion labels become instead severely unbalanced if provided with a fine-grained emotion annotation. We conclude that efforts are needed in terms of data collection to further push forward the research in this field.

**Keywords:** Emotion recognition · Valence-Arousal · Annotated dataset

## 1 Introduction

Human-to-human interaction occurs between individuals and may include hearing, tact, and vision data that describe the interaction [25]. During an interaction, it is important that all the involved actors understand the meaning of what is happening, especially because part of the communication occurs in non-verbal ways. In this context, facial expressions play a very important role, because they bring a significant non-verbal meaning. So, facial expressions have a central role in manifesting the emotional state of people, and showing an expression can modify the mood of the interaction [7]. Given its importance in social interactions, understanding facial expressions and the related emotional states, using automatic systems based on computer vision is a task that has been studied since decades, with a vast literature [1, 10, 15, 28, 29, 31]. Since it is present in most human interaction scenarios, facial expressions are relevant to understand non-verbal features. It is possible to apply a Facial Expression Recognition (FER) system in several scenarios, such as to measure tourist satisfaction [8], engagement evaluation [5], and human-robot affective interaction [24].

It is possible to divide FER systems into two main categories: *categorical* and *dimensional*. A system that operates in the first category, typically aims to classify a given facial expression into one of the six universal emotions plus the neutral one [6]. The main problem with this classification is that human emotions are more complex and subtle than those categories, and this rigid classification does not cover the real range of emotions. In dimensional emotion recognition [19, 21, 22], instead, a FER system usually extracts two values, *valence* and *arousal*, from a face image. Valence refers to a pleasure/displeasure degree, while arousal refers to the intensity of the displayed expression. This deals with the shortcoming of emotion recognition but the valence/arousal values are hard to interpret since they are not intuitive: while it is easy to understand what the label “happy” means, it is not so straightforward to understand what a value of 0.4 of valence and -0.2 of arousal does signify. Furthermore, a clear and unique mapping between the two representations has not been accurately defined.

Motivated by the above considerations, in this work we propose a way to transfer the continuous emotion description provided by VA values to a fine-grained human-readable taxonomy of expressions. The goal of this procedure is that of finding a way to avoid the burdensome process of collecting and labeling new data, while simultaneously analyzing the quality and reliability of existing annotation. The proposed procedure relies on two established studies in the emotion literature: the categorical/dimensional mapping of Russell *et al.* [22], and the tree-like emotion categorization proposed by Parrott [18].

Among the works of psychological constructionists, Russell *et al.* [22], identified 151 fine-grained emotional terms and also provided a map from each of them to a distribution in the VA plane. For example, the term “Excited” is associated with the values of 0.62 valence and 0.75 arousal, with a dispersion range of 0.25 and 0.2, respectively. Doing this mapping for all the terms resulted in a complete coverage of the valence/arousal plane, though the dispersion regions associated to each term largely overlap. Our idea here is to start from this fine

mapping and define a way to decide the emotion labels at coarser levels, while keeping a quantitative mapping between textual labels and the valence/arousal values. For the aggregation of the terms, we followed the term classification proposed by Parrott [18], a Basic Emotion Theorist, where emotions were categorized into primary, secondary and tertiary layers. The first layer includes the six basic emotions, while the secondary and tertiary layers are derived from the first one, and include, respectively, 25 and 115 emotions (a total of 140 emotion labels were provided with this categorization). Overall, with this terms aggregation, we obtained an intermediate ground that can better describe the range of human emotions. In summary, the main contributions of this work are:

- We proposed an emotion classification based on merging Russell’s 151 terms and the 140 terms of Parrott’s classification; the intersection between the two resulted in a first-level 6 emotion terms (Love, Joy, Surprise, Anger, Sadness, Fear) taken from the primary emotions, and a second-level 32 terms from the secondary and tertiary emotions;
- We experimentally showed that classifying expressions according to the proposed fine terms still represents a difficult task for expression classification based on deep learning methods as originally developed for the classification task into the six prototypical emotions;
- From the experimental outcomes, we highlighted two problems: (i) existing datasets suffer from a severe class unbalance if provided with fine-grained annotations, and (ii) there exists an inconsistency between emotion labels and corresponding VA values.

## 2 Related Work

Works in the literature mainly focused on emotion recognition, spontaneous expression recognition, micro-expressions detection, action units detection, and valence/arousal estimation [12]. Given the focus of our work, we will discuss some related works, focusing mainly on two tasks: emotion recognition and valence/arousal estimation. We just mention that some works combined Action Units (AUs) to emotional states. An example is the work in [4], where authors simulated emotional facial expressions according to pleasure, arousal, and dominance (PAD) values, and animated virtual human’s facial muscle AUs.

*Emotion Recognition.* In [13], a Multi-Scale Convolutional Neural Network was proposed that combined dilated convolutional kernels and automatic face correction aiming to improve the learned features from a CNN. Each expression was classified in one of the six universal expressions (*i.e.*, happiness, anger, sadness, disgust, surprise, fear) [6] or neutral. One shortcoming of this approach is the difficulty in correctly detecting the fear expression. To deal with pose and occlusion, the authors in [26] proposed a new approach called Region Attention Network (RAN), which captures the importance of facial regions in an adaptive manner. This aimed to increase robustness with respect to the aforementioned problems. To do so, RAN generated a compact representation aggregating and embedding

a varied number of region features that come from a backbone CNN. A critical aspect of this approach is a region-biased loss, which increases the attention weights for essential regions. Spite being overall better than the baseline, using RAN the accuracy decreased on specific emotions. This decrease in performance was mainly observed for *disgust* on the Occlusion-AffectNet data, for *disgust* and *neutral* on Pose-AffectNet, and for *fear* on Pose-FERPlus. In [30], two methods were proposed for facial expression recognition: the Double-channel Weighted Mixture Deep Convolution Neural Network, and the Deep Convolution Neural Network Long Short-Term Memory network of double-channel weighted mixture. The former focused on static images being able to quickly recognize expressions and provide static image features. The latter utilized static features to extract temporal features and precisely recognize facial expressions. The work in [14] proposed an approach with an end-to-end network utilizing attention mechanisms. It comprised four modules designed, respectively, for feature extraction, attention, reconstruction, and classification. The method combined Local Binary Pattern images with the attention modules to enhance the attention model, focusing on useful features for facial expressions recognition.

The main problem with emotion recognition is that it summarizes human emotion into seven classes, not being able to represent the diversity in which humans portray their emotions. In this sense, a more fine-grained representation of human emotion would be required.

*Dimensional Emotion Recognition.* In [12], a CNN-RNN based approach was proposed to dimensional emotion recognition, defined as the prediction of facial affection. In this case, VA varied in a continuous space [12]. A set of RNN subnets exploited the low-, mid- and high-level features from the trained CNN. The work in [2] utilized a Convolutional Autoencoder (CAE) that learned a representation from facial images, while keeping a low dimensional size for the features. A CNN is initially trained on a facial expression recognition dataset and the weights of the pre-trained network were used to initialize the convolutional layers in the proposed CAE. CAE was then used as an encoder-decoder to learn the latent representation of the data, which in turn was used to train a support vector regressor to infer the VA values. In [23], the EmoFAN network was proposed that combined networks performing facial alignment with expression recognition. In this way, as emotion classification, it was possible to obtain VA as well as fiducial points on the face. Instead of utilizing an attention mechanism, the keypoints of the face were used to focus on relevant regions of the surface.

Different from emotion recognition, dimensional emotion recognition is less straightforward to interpret. Annotating data in the dimensional space is also harder and more prone to error given the subjective nature of the human emotional state. One way to deal with such a problem is to provide a middle ground that motivated us to exploit the terms taxonomy defined by Russell [22], and the hierarchical label categories proposed by Parrott [18]. This is relevant because it generates a more realistic way of representing human emotions: it is simpler to interpret the results than dimensional emotion recognition, while increasing the diversity of representations, differing from the basic 6 classes used in

emotion recognition. Though we did not delve into this, the proposed fine grained terms can be further extended following our approach, by finding additional correspondences between Russell’s and Parrott’s terms.

### 3 Proposed Emotion Taxonomy

In this section, we discuss the selection processes and the taxonomy we chose for our work.

**Russell Vs. Parrott.** Describing an emotional state of a person with the six prototypical emotions, *i.e.*, joy, sadness, disgust, fear, anger and surprise, provides us with her/his expression but does not give us the person’s emotions. In the work of Russell [22], a set of 151 terms were provided with the corresponding distributions of *valence*, *arousal* and *dominance* values (this is given as a mean value plus a standard deviation term). The 151 terms not only describe the emotions in a precise way but also include terms that represent emotional interaction between two individuals. Russell defined the emotional state by three independent and bipolar dimensions, pleasure-displeasure, degree of arousal, and dominance-submissiveness. He defined it as a result of a study on 200 subjects and 42 verbal-report emotion scales. In our work, we use the *dimensional* representation of emotions based on VA, without using the dominance dimension. This is motivated by a subsequent work of Russell [20] that introduced the *circumplex* model of emotion, where the dimensions of excitation and valence were distributed in a two-dimensional circular space. Most of facial expression datasets that are labeled with the dimensional representation use these two dimensions.

In the same context, Parrott [18] defined a taxonomy for emotion-related terms in a tree-like structure starting with six primary emotions (*i.e.*, love, joy, surprise, anger, pain and fear), then secondary emotions and tertiary emotions, for a total of 140 terms. Although the Parrott’s hierarchy accurately describes the emotion of a person, it does not provide the measure of VA. Hence, one founding idea of our work is that of taking the common terms between Russell’s work and Parrott’s classification to derive the VA of each emotion.

**Terms Selection and Mapping.** To define a larger set of emotional terms with respect to the six basic emotions, we first merged the terms of Russell and Parrott; then, we selected the subset of most similar terms to get a mapping from Parrott’s structure to Russell’s values for VA. For example, the terms *joy* and *joyful* appear in both taxonomies, and the related (*valence,arousal*) pair given by Russell for joyful is (0.76,0.48). So, we can both exploit the values mapping and the Parrott’s hierarchical structure. This process is depicted in Fig. 1.

This process, resulted into 32 terms that we organized as reported in Table 1. In addition, we also included the *neutral* emotion to the selected set of terms. In Parrott’s classification of emotions, 3 positive and 3 negative terms were proposed, respectively, for primary emotion, 11 positive and 14 negative terms

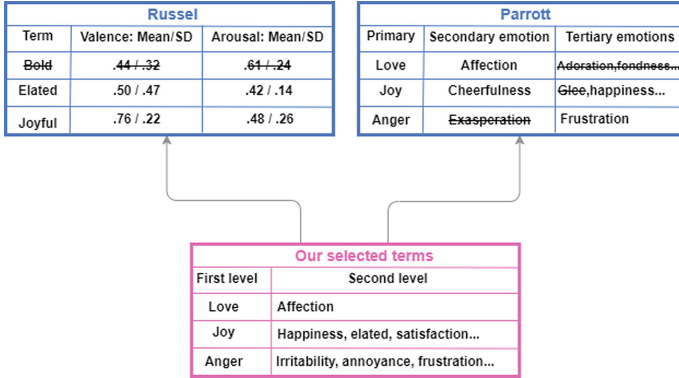


Fig. 1. Mapping between emotion definitions.

Table 1. Our selected emotional terms (6 and 32, respectively. at the first and second level of the categorization) plus neutral.

First level emotions	Second level emotions
Love	Affection
Joy	Happiness, Elation, Satisfaction, Excitement, Triumph
Surprise	Astonishment
Anger	Irritability, Annoyance, Frustration, Rage, Hostility, Hatred, Scorn, Disgust, Contempt
Sadness	Depression, Despair, Displeasure, Shame, Guilt, Regret, Refeatism, Embarrassment,
	Humiliation, Insecurity, Isolation, Loneliness, Rejection
Fear	Terror, Anxiety, Distress

for secondary emotion, and 42 positive and 73 negative terms for tertiary emotion, respectively. With this classification of emotions, we can see an imbalance between positive and negative terms. Since our classification is derived from Parrott’s, we can observe the dominance of negative emotions over positive ones in the proposed 32 terms: 3 positive and 3 negative terms, and 7 positive and 25 negative terms for the first and second level emotions, respectively.

**Terms Taxonomy.** After merging Russell’s and Parrott’s terms, the intersection between the two classifications yields 38 terms (see Table 1) from the primary, secondary, and tertiary layers of the Parrott term representation. Taking Parrott’s classification as a guideline for our classification, we divided the 38 terms into two levels: the first level includes 6 terms (*love, joy, surprise, anger, sadness, fear*); the second level comprises 32 terms plus the neutral state (*neutral* is positioned in the origin of the VA plane). The choice of having only two levels in the hierarchy, unlike Parrott’s classification, is due to the fact that most

of the selected terms come from Parrott’s tertiary layer. In the experiments, we used only the terms in the second level to re-label the dataset because the terms in the first level are independent emotions, intense, long-lasting and irreversible: even in case of reversing them, they need longer period. Contrarily, emotions in the second class are less intense and dependent upon primary emotions.

## 4 Experiments

The objective of our evaluation is to demonstrate the applicability of the proposed relabeling to a benchmark dataset. In the following, we considered the AffectNet dataset [16]. Due to the limited number of datasets annotated with both emotions and VA [9], the AffectNet dataset was chosen.

### 4.1 Relabeling the AffectNet Dataset

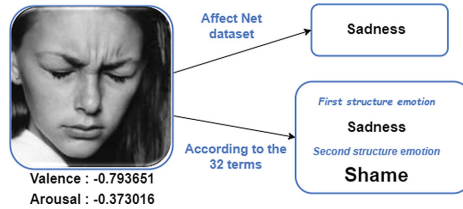
AffectNet [16] is a large facial expression dataset acquired in the wild, with around 0.4 million images manually annotated for the presence of six categorical facial emotions (*i.e.*, *happy*, *sad*, *surprise*, *fear*, *disgust*, *angry*), plus contempt and neutral. Values of VA are also provided. In addition, the three non-emotional labels of *none*, *uncertain*, and *no-face* are used. In particular, the *none* (“none of the eight emotions”) category is an expression/emotion (such as sleepy, bored, tired, seducing, confuse, shame, focused, *etc.*), that could not be assigned by annotators to any of the six basic emotions, contempt or neutral. However, VA could be assigned to these images. The relabeling process was performed according to the VA annotations in AffectNet. These values were used to link each image with one of the proposed 32 emotional terms plus neutral. The relabeling is as follows:

**Validation Set** - 19 terms were used in the relabeling (Insolent, Anxious, Disgusted, Insecure, Self-satisfied, Frustrated, Astonished, Depressed, lonely, Shamed, Excited, Affectionate, Elated, Hate, Defeated, Hostile, Irritated, Enraged, Happy). For a total of 248 images (6,20% of the relabeled images) no matching terms were found;

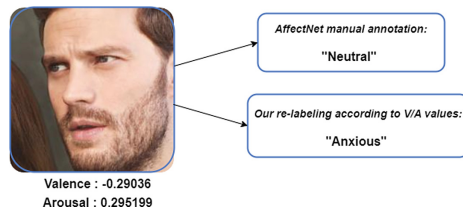
**Training Set** - 21 terms were used in the relabeling (Insolent, Affectionate, Self-satisfied, Anxious, Disgusted, Insecure, Shamed, Elated, Depressed, Hate, Excited, Astonished, Irritated, lonely, Happy, Frustrated, Defeated, Enraged, Regretful, Hostile, Despairing). In this case, 44,871 images (15,60% of the relabeled images) were not relabeled.

Figure 2 shows the multi-term relabeling process for an example image from AffectNet. With the proposed terms, emotions can be described with more than the basic terms: for example “sad” has a variety of terms, such as depression, despair, displeasure, shame, guilt, regret, embarrassment that describe the emotion more accurately. More in detail, the image on the left was originally annotated in AffectNet with VA values of, respectively,  $-0.793651$  and  $-0.373016$ ,

and “Sadness” as expression label. With the proposed labeling, we can associate the point in the VA plane with first structure emotion “Sadness” and second structure emotion “Shame”. Applying the proposed terms hierarchy, we can derive a coarser expression characterization with 6 first class emotions (love, joy, surprise, anger, sadness, fear) and 32s class emotions, that are more descriptive of the main emotion.



**Fig. 2.** Example of relabeling an image from the AffectNet dataset according to the proposed approach.



**Fig. 3.** Difference between AffectNet and our re-labeling.

It turns out that AffectNet is manually annotated and labeled with a balanced number of classifications for the emotions (500 images for each emotion). But when we re-labeled the AffectNet images based on VA values, we found a large imbalance in the re-labeling results. This is due to the large difference between annotating a dataset according to the VA values and according to manual labeling. An example showing the efficacy of our proposed terms is illustrated in Fig. 3. In this case, an image with valence  $-0.29036$  and arousal  $0.295199$  was originally labeled in AffectNet as “Neutral”, though the “Neutral” label is the non-emotional state associated with the origin of the VA plane ( $0.0$  value for both dimensions). The relabeling process applied to this image resulted in a relabeling with the “Anxious” term. According to Russell’s work, the distributions for VA associated to this term are as follow: mean valence of  $0.01 \pm 0.45$ , and arousal of  $0.59 \pm 0.31$ .

Our observation here, is that the VA values provide a viable way to expand the categorical emotion representation with 7 (6 basic plus neutral) or 8 (with contempt) terms to a finer grained set of emotional states. Proposing the use of Russell’s mapping and the relation between Russell’s and Parrott’s terms, we



contribute a motivated and flexible way to include more terms in the classification of expression datasets fostering the design of more effective and realistic classification solutions.

## 4.2 Results

In this section, we aim at evaluating the proposed classification emotion annotations. We are interested in estimating the consistency of our mapping with the manual annotations for the basic 8 emotion classes. With this aim, we used the Distract your Attention Network (DAN) [27]. The DAN model was proposed for facial expression recognition, while we used it as classifier using our proposed annotations as classes to perform emotion recognition. We used 287,651 images for the training set and 3,999 images for the validation set, according to the annotation of VA values provided by AffectNet. DAN includes three key components: a Feature Clustering Network (FCN) extracts robust features by adopting a large-margin learning objective to maximize class separability; a Multi-head cross Attention Network (MAN) instantiates a number of attention heads to simultaneously attend to multiple facial areas and build attention maps on these regions; finally, an Attention Fusion Network (AFN) distracts these attentions to multiple locations before fusing the attention maps to a comprehensive one.

In AffectNet, a total of 291,650 images manually annotated with eight labels were released for public research. We used the validation set including 3,999 images (the test set is not released). Results of our experimentation (Table 2) shows an accuracy of **27.94%** compared to the original accuracy of **61%** obtained by running DAN as classifier on the original 8 expression labels of AffectNet. Observing the distribution of the 32 emotions over the whole dataset, and the lack of homogeneity between the number of images for each term, an increased difficulty of the classification task is revealed. We believe this result can open the way to new challenges in the task of emotion recognition. Especially in the availability of datasets that are annotated not only with more than the 6 basic emotions but also with the values of VA.

**Table 2.** Accuracy on the AffectNet dataset obtained by using DAN with our 32 terms in comparison with the 8 original terms.

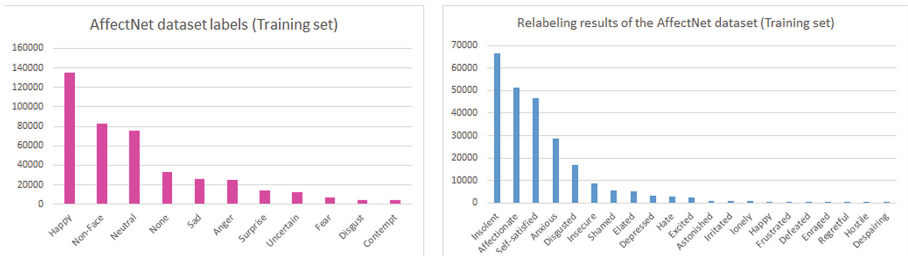
Dataset	Original terms	32 terms
AffectNet [16]	61%	27.94%

*Unbalanced Data* – In this paragraph, we address the problem of class unbalance in facial recognition datasets, and how this problem affects the final performance. Without handling unbalance during data collection results in classifications that are biased toward the majority class with poor accuracy for the minority classes [11]. The class unbalance problem represents one of eight different unbalance problems [17]; it happens when there is a noticeable disparity

in the number of examples that belong to the various classes. The unbalance in datasets is not only evident in terms of emotions, but also in the distribution of positive and negative emotions. We refer further to the unbalance in Parrott’s classification regarding positive and negative emotions.

AffectNet’s validation set, which includes 500 images for each emotion manually annotated, is unbalanced when considering the VA value of each image. However, AffectNet’s training set is out of balance in terms of VA value. In Fig. 4, we show the impact of class unbalance of the AffectNet training set on our relabeling: the relabeling results shift towards the majority of classes already distributed in AffectNet: see the over-represented class “happy”, with 13,4915 images, versus the under-represented class “contempt” with 4,250 images. This induces not only unbalance in the data but also inconsistency between the manual labeling and the assigned VA values. This leads us to one outcome of our work: data unbalance results into poor accuracy, even when data annotation is improved with the proposed 32 classes.

*Inconsistency Between Emotion and VA Labels* – As presented in Sect. 4.1 and in Fig. 3, VA values are attributed regardless of the actual value of the labeled emotion, and this is due to the manual annotation of the dataset (the manual annotation in AffectNet was done by one annotator for each image). For “neutral” the attributed VA values are not only 0.0 but also assume other values such as  $(V,A)=(-0.176846,-0.176846)$ ,  $(V,A)=(-0.135501,0.00483933)$ . This presents a challenge in relabeling an existing benchmark dataset according to its valence and arousal values, while the emotions are labelled manually.



**Fig. 4.** Impact of class unbalance on: AffectNet training set (left); our relabeling (right).

## 5 Conclusions

We believe that emotions are more extensive than the six basic categorical facial expressions, and more complicated than describing a state of a person by only “sad”, “happy”, “fear”, *etc.* In our work, we proposed a classification of emotions according to the dimensional representation of emotions based on VA proposed by Russell [22], and Parrott’s classification of emotions [18]. Taking inspiration

from both the models, and combining a subset of the Parrott's hierarchical categorization of emotions with the Russell's mapping between the categorical and the dimensional domain resulted in 38 terms, which describe more precisely the state of a person. We divided the terms into two classes: a primary class of emotions with 6 terms, and a secondary class of emotions that depend on the first class with 32 terms. We re-labeled the AffectNet dataset according to the 32 proposed terms.

Despite the fact that the results show an accuracy of 27.94% compared to the 61% obtained for the original dataset annotation with 8 emotions, they revealed the gap between manual annotation and annotating according to the VA values. The validation set of AffectNet is balanced in terms of annotating through judging if the image is happy, sad, surprise, fear, disgust, angry, contempt, presenting 500 images for each emotion, but looking at the VA value of each image, we found an unbalance in attributing the emotions in the dataset, and that is the reason behind the unbalance in the relabeling of the dataset.

Labeling a dataset manually or based on VA values are two different ways that are still difficult to compare. As future work, we will annotate other datasets that include both VA and categorical emotion labels, like the OMGEmotion Challenge dataset [3] with the overall set of terms in our proposed taxonomy.

## References

1. Abdullah, S.M.S., Abdulazeez, A.M.: Facial expression recognition based on deep learning convolution neural network: a review. *J. Soft Comput. Data Min.* **2**(1), 53–65 (2021)
2. Allognon, S.O.C., Britto, A.S., Koerich, A.L.: Continuous emotion recognition via deep convolutional autoencoder and support vector regressor. In: *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE (2020)
3. Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The omg-emotion behavior dataset. In: *IEEE Int. Joint Conference on Neural Networks (IJCNN)*, pp. 1–7 (2018)
4. Boukricha, H., Wachsmuth, I., Hofstätter, A., Grammer, K.: Pleasure-arousal-dominance driven facial expression simulation. In: *IEEE International Conference on Affective Computing and Intelligent Interaction*, pp. 1–7 (2009)
5. Carlotta Olivetti, E., Violante, M.G., Vezzetti, E., Marcolin, F., Eynard, B.: Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: a preliminary approach. *Appl. Sci.* **10**(1) (2020)
6. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **17**(2), 124 (1971)
7. Feng, W., Kannan, A., Gkioxari, G., Zitnick, C.L.: Learn2Smile: learning non-verbal interaction through observation. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4131–4138 (2017)
8. González-Rodríguez, M.R., Díaz-Fernández, M.C., Gómez, C.P.: Facial-expression recognition: an emergent approach to the measurement of tourist satisfaction through emotions. *Telematics Inform.* **51**, 101404 (2020)
9. Guerdelli, H., Ferrari, C., Barhoumi, W., Ghazouani, H., Berretti, S.: Macro-and micro-expressions facial datasets: a survey. *Sensors* **22**(4), 1524 (2022)

10. Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., Huang, T.S.: Multi-view facial expression recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–6 (2008)
11. Huang, C., Li, Y., Loy, C.C., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(11), 2781–2794 (2019)
12. Kollias, D., Zafeiriou, S.P.: Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Trans. Affect. Comput.* (2020)
13. Lai, Z., Chen, R., Jia, J., Qian, Y.: Real-time micro-expression recognition based on ResNet and atrous convolutions. *J. Ambient Intell. Hum. Comput.*, 1–12 (2020)
14. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **411**, 340–350 (2020)
15. Li, Y., Wang, S., Zhao, Y., Ji, Q.: Simultaneous facial feature tracking and facial expression recognition. *IEEE Trans. Image Process.* **22**(7), 2559–2573 (2013)
16. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**(1), 18–31 (2017)
17. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3388–3415 (2020)
18. Parrott, W.G.: Emotions in social psychology: essential readings. Psychology Press (2001)
19. Plutchik, R.: Emotion, a psychoevolutionary synthesis (1980)
20. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161 (1980)
21. Russell, J.A.: Core affect and the psychological construction of emotion. *Psychol. Rev.* **110**(1), 145 (2003)
22. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**(3), 273–294 (1977)
23. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **3**(1), 42–50 (2021)
24. Val-Calvo, M., Álvarez-Sánchez, J.R., Ferrández-Vicente, J.M., Fernández, E.: Affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access* **8**, 134051–134066 (2020)
25. Valtakari, N.V., Hooge, I.T., Viktorsson, C., Nyström, P., Falck-Ytter, T., Hessels, R.S.: Eye tracking in human interaction: possibilities and limitations. *Behav. Res. Methods*, 1–17 (2021)
26. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **29**, 4057–4069 (2020)
27. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: multi-head cross attention network for facial expression recognition. [arXiv:2109.07270](https://arxiv.org/abs/2109.07270) (2021)
28. Xiang, J., Zhu, G.: Joint face detection and facial expression recognition with MTCNN. In: International Conference on Information Science and Control Engineering (ICISCE), pp. 424–427 (2017)
29. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

30. Zhang, H., Huang, B., Tian, G.: Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recogn. Lett.* **131**, 128–134 (2020)
31. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. *IEEE Trans. Affect. Comput.* **2**(4), 219–229 (2011)