

# RAF-AU Database: In-the-Wild Facial Expressions with Subjective Emotion Judgement and Objective AU Annotations

Wen-Jing Yan<sup>1\*</sup>, Shan Li<sup>2\*</sup>, Chengtao Que<sup>1</sup>, Jiquan Pei<sup>1</sup>, and Weihong Deng<sup>2</sup>

<sup>1</sup> JD Digits, Beijing, China

<sup>2</sup> Beijing University of Posts and Telecommunications

eagan-ywj@foxmail.com, {ls1995, whdeng}@bupt.edu.cn,  
{quechengtao, peijiquan}@jd.com,

**Abstract.** Much of the work on automatic facial expression recognition relies on databases containing a certain number of emotion classes and their exaggerated facial configurations (generally six prototypical facial expressions), based on Ekman’s Basic Emotion Theory. However, recent studies have revealed that facial expressions in our human life can be blended with multiple basic emotions. And the emotion labels for these in-the-wild facial expressions cannot easily be annotated solely on pre-defined AU patterns. How to analyze the action units for such complex expressions is still an open question. To address this issue, we develop a RAF-AU database that employs a sign-based (i.e., AUs) and judgement-based (i.e., perceived emotion) approach to annotating blended facial expressions in the wild. We first reviewed the annotation methods in existing databases and identified crowdsourcing as a promising strategy for labeling in-the-wild facial expressions. Then, RAF-AU was finely annotated by experienced coders, on which we also conducted a preliminary investigation of which key AUs contribute most to a perceived emotion, and the relationship between AUs and facial expressions. Finally, we provided a baseline for AU recognition in RAF-AU using popular features and multi-label learning methods.

## 1 Introduction

Of all nonverbal behaviors—body movement, posture, gaze, proxemics and voice—the face is probably the most commanding and complicated, and perhaps the most confusing. The face, and especially facial movement (or expression), is commanding because it is always visible and therefore always providing information such as emotion and intent [1]. Due to their utility for understanding a human being’s mental state, the recognition of automatic facial expressions is becoming a popular field in computer vision.

Automatic facial expression recognition relies heavily on training datasets. When training a system to recognize facial expressions, the investigator must assume that the training and test data have been accurately labeled. This assumption may or may not be accurate. Traditionally, researchers have categorized facial expressions as expressing

---

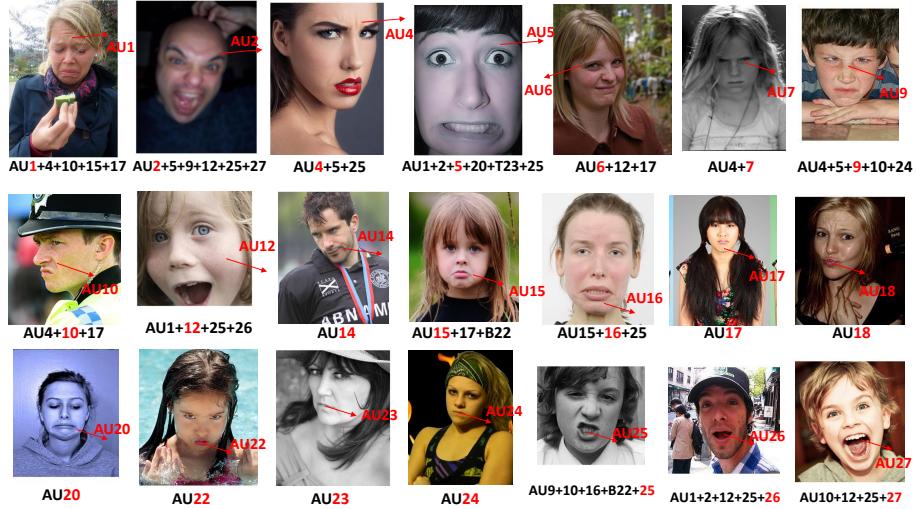
\* These authors contributed equally to this work.

happiness, sadness, surprise, fear, anger and disgust. Each of these prototypical facial expressions can be described via a pattern of action units (AUs), based on the Facial Action Coding System (FACS) developed by Ekman and colleagues [2] and supported by Basic Emotion Theory (BET) [3].

Early in the process, samples of facial expressions are mostly viewed from the front, collected from actors required to pose preset AUs that readily match prototypical facial expressions [4, 5]. Thus, recognition of the emotion can be judged against the “correct” expression adopted by the poser. Recently, datasets of spontaneous facial expressions have multiplied [6–10]. This is remarkable progress because the goal of automatic facial expression recognition is to apply to in the real world rather than lab situations (i.e., the ideal world). Facial expressions in real life are “in the wild”, meaning that they are not necessarily frontal and direct-gazed, nor demonstrate only a very limited number of AU patterns. There may be various combinations of AUs with different gestures, head poses, gazes, and environments.

Due to the characteristics of in-the-wild facial expressions, applying an emotion label to a certain face can be difficult. AU combinations do not usually fall precisely into the six (or more) prototypical facial expressions [11]. EMFACS [4, 12], usually understood as a guide for labeling emotions expressed through AU combinations, has become inappropriate in this context. Though previous research has summarized the relationship between AUs and emotion [13, 14], researchers have only provided general or key AUs for certain emotions. Moreover, the coherence problem has long been a topic of debate in the field of facial expression [15–17]. Even for the six universally used and prototypical facial expressions, psychologists have not reached a consensus on the relationship between AU patterns and emotions, let alone facial expressions in the wild. In other words, we cannot simply categorize in-the-wild facial expressions by AU pattern, at least not for the moment.

Generally, psychologists have proposed two approaches to studying nonverbal behavior (including facial expressions), either judgement-based or sign-based [1, 13]. In judgement-based approaches, observers make inferences about the meaning of facial actions and assign corresponding labels. When classifying facial expressions into a pre-defined number of emotions or mental activity categories, the agreement of a group of annotators is taken as ground truth, usually by computing the average of the responses of either experts or non-experts. The rationale that judgement-based approaches can provide “correct” labels can be explained from an evolutionary perspective: the facial expressions broadcast from a sender (i.e., encoding the signal) should be universally understood by a perceiver (i.e., decoding the signal), or they would be useless and therefore removed. Thus, crowdsourced annotation, a practice that employs many perceivers to heuristically label a target, might be a useful way of labeling in-the-wild facial expressions. As for sign-based approaches, facial motion and deformation are coded into visual classes. Facial actions are then abstracted and described by their location and intensity, such as in FACS. Ideally, a complete description framework would contain all possible perceptible changes that might occur on a face. Most automatic facial expression analysis approaches have attempted to directly categorize facial expressions into basic emotion classes [18] by AUs, a sign-based approach.



**Fig. 1.** Example samples annotated with typical AUs in the RAF-AU dataset.

The judgement-based approach, with subjective estimation, provides the perceivers' judgement because facial expressions are born to be perceived and understood by conspecifics. From this perspective, conspecifics can provide the correct interpretation of a given facial expression. The sign-based approach, with objective description, can tell the computer where and how the facial movements will occur. Ideally, a qualified mechanism for annotating facial expressions should include both subjective and objective elements.

With such an understanding of facial expressions and their annotation, the present work provides an updated version of the existing RAF-ML database [9], called RAF-AU database<sup>3</sup>, which consists of in-the-wild facial expressions with both subjective (i.e., judgement-based) and objective (i.e., sign-based) annotations. The RAF-ML database contains facial images in different occlusions, illuminations and resolutions collected from the social network and provides multiple labels via crowdsourcing annotations for each facial image. To extend this database with objective elements, we conducted AU coding according to strict criteria by two FACS experts. Specially, two experienced coders were requested to independently code each the facial image involving 26 kinds of AUs and the inter-coder correlation was 0.6376. To our best knowledge, the RAF-AU database is the first facial expression dataset that includes both subjective emotion judgement and objective AU annotations for multi-label expression analysis in-the-wild. Figure 1 shows example images annotated with typical AUs in the RAF-AU dataset. Using both crowdsourced and AU annotations, we make the first attempt to investigate the relationship between AUs and perceived emotions in the wild. We explored which AUs contributed the most to each facial expression. We then conducted

<sup>3</sup> <http://whdeng.cn/RAF/model3.html>

AU detection experiments using popular features and multi-label learning methods. A baseline was provided for AU recognition in the RAF-AU.

## 2 Related Work

Several surveys have offered overviews of facial expression datasets [19, 20]. The present research does not repeat their work, but rather inspects how facial movements were annotated with emotion labels. Many of these databases classified facial expressions into types, based on rules such as those in EMFACS (i.e., mapping AUs to facial expressions). Usually, participants were required to pose “standard” AUs to express so-called prototypical facial expressions, and those that qualified were included in the database.

In the Cohn-Kanade Expression Database [4], emotion labels refer to the expression requested rather than what may actually have been performed. AUs for each facial expression (i.e., the apex frame in a sequence) are converted into emotion-specific expressions (e.g., happiness or anger). BU-3DFE [21] and MMI [22] used a similar technique to assign emotion labels to specific facial movements. Researchers have also used this approach to annotate emotions expressed in spontaneous facial expressions. EmotioNet [6] is a large-scale database with one million facial expression images collected from the Internet. Most samples were annotated by an automatic AU detection algorithm, and the remaining 10% were manually annotated using AUs. EmotioNet contains six basic expressions and one neutral expression. The creators defined unique AU patterns that mapped AUs to specific emotions. Therefore, the emotion labels were inferred from the AUs. For example, if the face showed AU4+AU15, the emotion label attributed was “sadness”. Similar mapping rules were used to label compound facial expressions. EmotioNet defined 17 compound expressions, depending on the AU combination. For example, AU4+AU20+AU25 was defined as fearfully angry. Similarly in [14], subjects were required to practice their expressions before each acquisition. They were required to express more than one emotion based on prototypical facial expressions; these were then evaluated by experts for validity. This annotation approach was also based on prototypical facial expression protocol.

The dimension approach was used to annotate facial expressions from valence (unpleasant to pleasant) and arousal (low to high). For example, AffectNet [7] contains more than one million images from the Internet that were obtained by querying different search engines using emotion-related tags. Over 450,000 images include manually annotated labels according to eight basic expressions. Another valence-arousal dataset is Aff-Wild [23, 24] which contains 298 videos with spontaneous facial behaviors collected in arbitrary recording conditions. Most recently, the Aff-Wild dataset was extended to the Aff-wild2 [25, 26] with more videos and other attribute annotations such as basic expressions and action units. It is by far the largest database providing facial expressions with all three types of behavior states, i.e., valence and arousal (VA), basic expression and facial action unit.

The situation approach derives emotion labels from the situation (with an anticipated emotion) where facial expressions elicited are taken as a cue. The BP4D-Spontaneous database [27] used various tasks in a lab environment to elicit spontaneous facial expressions. For example, smelling an unpleasant odor should be disgusting and thus the facial

expression during the task should be one of disgust. The Aff-Wild dataset employed similar elicitation method to record videos containing naturalistic emotional states in arbitrary recording conditions.

The self-report approach is another way of annotating facial expression type. The Belfast Induced Natural Emotion Database [28] is unique. Recordings are accompanied by self-reports of emotion and intensity, serving as continuous trace-style ratings of valence and intensity. Self-reporting may be the most reliable approach to labeling emotion in facial expressions. Facial movements in this database were also annotated with valence and intensity (or arousal) according to the dimension of the emotion. In Belfast, they used many more words to describe each emotion than the popular six prototypical facial expressions. However, self-reported information is not easily collected and may not reflect the inner mind [29] and thus not commonly used.

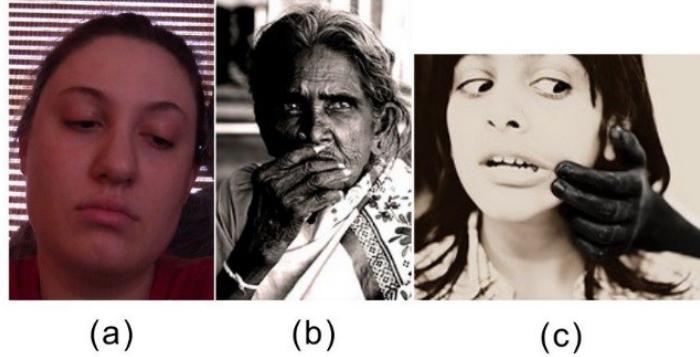
Crowdsourcing is a judgment-based approach that labels facial expressions based on human perception. The Real-world Affective Face Database (RAF-DB) [8, 30] is a real-world database containing 29,672 highly diverse facial images downloaded from the Internet. Images were randomly and equally assigned to each labeler, ensuring that there was no direct correlation among the images labeled by a single annotator. For the manually crowdsourced annotations, seven basic (including neutral) and 11 compound emotion labels were applied to the samples. Each image was ensured to be labeled by approximately 40 independent labelers. Later, RAF-ML [9] was developed. This database contains facial expressions obtained from RAF-DB that offers multiple expressions, and extends the sample collection, 4,908 in total. Another dataset, the Expression in-the-Wild Database (ExpW) [10], contains 91,793 faces downloaded using Google image searches. Each face image was manually annotated and categorized by human beings into one of seven basic expression categories, but the study didn't explain whether this coding was based on heuristic judgement or mapping rules.

In summary, posed facial expressions are uncommon in real life, so many databases collect spontaneous samples. These spontaneous and in-the-wild facial expressions are very difficult to annotate because they do not usually display commonly accepted and pre-defined AU combinations. Mapping rules comprise the most common approach to labeling facial expression type. There are also dimension, situation, self-report, and crowdsourcing approaches. With images of facial expressions downloaded from the Internet, the subject's emotional state and mindset are unknown, and researchers can't simply infer them by the AU combinations displayed. Therefore, a crowdsourcing approach that leverages human perception to apply annotations is the most suitable of those available. In the present work, we provide a dataset of both judgement-based (i.e., subjective) and sign-based (i.e., objective) annotations. Only a combination of subjective and objective annotations will properly disclose the emotional implications of facial expressions. This is what is provided by RAF-AU.

### 3 RAF-AU

#### 3.1 AU annotations

Facial images from the Internet vary in quality; this is true not only in terms of image resolution, but also spontaneity, since many are posed and thus not emotion-motivated.



**Fig. 2.** Examples of “null” images.

There are certain individual differences in facial appearance and habitual movement. Some AUs habitually appear, making emotional perception confusing. For example, a smile with AU9 (i.e., a wrinkled nose) is more likely to appear in women, while people with gag teeth are more likely to display AU10 (a raised upper lip). AU coding of facial expressions requires a baseline (or neutral) image, or it is difficult to judge whether certain expressions were due to the face’s original appearance or facial movement. For example, eyebrow AU4 can be a permanent feature or transient movement. If we are forced to imagine a neutral face for each face image, then we have to judge whether the expression is caused by a certain AU, perpetual feature, or the influence of another AU.

In the present research, those with more than two whole faces (305 images in total) were eliminated from the RAF-ML database. There were some images that contained more than two faces. When we checked the voting, we found that participants were not always voting for the same face. For example, one image contained a smiling face and a sad face, but about 30% of the participants voted happiness while 25% voted sadness. After removing these images, 4,601 were left for further coding.

Two experienced coders independently FACS-coded these face images and arbitrated any disagreement. They usually disagreed on whether a trace of an AU was actually a baseline or emotion-motivated. They also carefully checked and discussed if AUs emerged due to other AUs. The inter-coder correlation (i.e., reliability, see Eq. (1)) was 0.6376. This relatively low reliability was due to the complexity of and lack of clarity in these in-the-wild and blended facial expressions.

$$R = \frac{2 * \#AU(C_1C_2)}{\#All\_AU}, \quad (1)$$

where  $\#AU(C_1C_2)$  is the number of AUs upon which Coders 1 and 2 agreed and  $\#All\_AU$  is the total number of AUs in a facial expression scored by two coders.

There were 253 images annotated as null. Some of these facial images didn’t show any AUs. Though there were no AUs for these faces, they were still selected for this database because according to perceivers, they expressed emotions. This is not uncom-

AU	Name	Num	AU	Name	Num
<b>1</b>	<i>Inner Brow Raiser</i>	1028+(49)	<b>21</b>	<i>Neck Tightener</i>	3
<b>2</b>	<i>Outer Brow Raiser</i>	701+(95)	<b>22</b>	<i>Lip Funneler</i>	196+(148)
<b>4</b>	<i>Brow Lowerer</i>	1808+(9)	<b>23</b>	<i>Lip Tightener</i>	80+(51)
<b>5</b>	<i>Upper Lid Raiser</i>	975+(10)	<b>24</b>	<i>Lip Pressor</i>	165+(56)
<b>6</b>	<i>Cheek Raiser</i>	404+(46)	<b>25</b>	<i>Lips part</i>	2830
<b>7</b>	<i>Lid Tightener</i>	347+(14)	<b>26</b>	<i>Jaw Drop</i>	1089
<b>9</b>	<i>Nose Wrinkler</i>	749+(25)	<b>27</b>	<i>Mouth Stretch</i>	810
<b>10</b>	<i>Upper Lip Raiser</i>	1274+(116)	<b>28</b>	<i>Lip Suck</i>	13+(31)
<b>12</b>	<i>Lip Corner Puller</i>	1187+(81)	<b>29</b>	<i>Jaw Thrust</i>	27
<b>14</b>	<i>Dimpler</i>	105+(22)	<b>30</b>	<i>Jaw Sideways</i>	9
<b>15</b>	<i>Lip Corner Depressor</i>	290+(2)	<b>32</b>	<i>Lip Bite</i>	6
<b>16</b>	<i>Lower Lip Depressor</i>	720	<b>33</b>	<i>Cheek Blow</i>	1
<b>17</b>	<i>Chin Raiser</i>	541	<b>34</b>	<i>Cheek Puff</i>	9+(1)
<b>18</b>	<i>Lip Puckerer</i>	118+(11)	<b>35</b>	<i>Cheek Suck</i>	10+(1)
<b>19</b>	<i>Tongue Visible</i>	34	<b>39</b>	<i>Nostril Compressor</i>	5
<b>20</b>	<i>Lip stretcher</i>	199+(33)	<b>43</b>	<i>Eyes Closed</i>	148+(9)

**Table 1.** Frequency of all used AUs (without action descriptors) coded by manual FACS coders on the RAF-AU dataset. The numbers in parentheses refer to the sample size of one-sided AUs.

mon because perceived emotion is influenced by gaze, head pose, gesture, and facial appearance. Figure 2 shows some examples labeled as null. In the figure, (a) is null because the faintly discernible AU17 (i.e., Chin raiser) is not sufficiently obvious; without a baseline, we weren't confident we were avoiding a mistake. In (b), the face of this elderly person has many perpetual traces, and therefore was hard to annotate without a baseline. The AU4 marker seemed locally obvious. However, this is an image of an elderly person that did not show a clear vertical grain between the eyebrows. Therefore, we didn't annotate this image with AU4. The forehead in (c) is covered and the mouth may be stretched by an outer force. These AUs are either invisible or involuntary.

### 3.2 Profile of RAF-AU

This database contains 4,601 images obtained from RAF-ML. We have provided AU annotations for all faces. There were 26 AUs used in the annotations, without action descriptors. We list all used AUs and their frequency in Table 1. It should be noted that the AU distribution was largely imbalanced. For example, only a few AU39 (i.e., compressed nostrils) were found in this dataset. Some representative examples with frequent AUs in the RAF-AU are shown in Figure 1.

Surprise		Fear		Disgust	
<b>AU25</b>	0.5926	<b>AU25</b>	0.5111	<b>AU10</b>	0.5964
<b>AU5</b>	0.4665	<b>AU12</b>	0.4033	<b>AU4</b>	0.5330
<b>AU26</b>	0.3820	<b>AU27</b>	0.3729	<b>AU17</b>	0.2973
Happiness		Sadness		Anger	
<b>AU12</b>	0.7040	<b>AU4</b>	0.6723	<b>AU25</b>	0.4659
<b>AU25</b>	0.5143	<b>AU25</b>	0.3462	<b>AU9</b>	0.4337
<b>AU27</b>	0.2491	<b>AU1</b>	0.2979	<b>AU10</b>	0.4236

**Table 2.** AUs that contribute most (variance) for these facial expressions in RAF-AU.

	Sur	Fea	Dis	Hap	Sad	Ang	Neu
<b>Sur</b>	1.0000	0.6802	0.1733	0.4908	0.3284	0.3688	-0.3784
<b>Fea</b>	0.6802	1.0000	0.3645	0.8168	0.5754	0.7665	-0.8262
<b>Dis</b>	0.1733	0.3645	1.0000	0.2497	0.6997	0.6266	-0.4654
<b>Hap</b>	0.4908	0.8168	0.2497	1.0000	0.5048	0.6207	-0.7539
<b>Sad</b>	0.3284	0.5754	0.6997	0.5048	1.0000	0.6342	-0.4866
<b>Ang</b>	0.3688	0.7665	0.6266	0.6207	0.6342	1.0000	-0.8170
<b>Neu</b>	-0.3784	-0.8262	-0.4654	-0.7539	-0.4866	-0.8170	1.0000

**Table 3.** The relationship of the facial expressions based on the AU combination for RAF-AU. Sur=Surprise, Fea=Fear, Dis=Disgust, Hap=Happiness, Sad=Sadness, Ang=Anger, Neu=Neutral.

	Sur	Fea	Dis	Hap	Sad	Ang	Neu
<b>Sur</b>	1.0000	0.2706	-0.5585	0.1544	-0.4168	-0.3756	-0.2598
<b>Fea</b>	0.2706	1.0000	-0.4242	-0.2193	-0.1224	-0.1628	-0.3568
<b>Dis</b>	-0.5585	-0.4242	1.0000	-0.2673	0.0686	0.0739	0.1447
<b>Hap</b>	0.1544	-0.2193	-0.2673	1.0000	-0.3023	-0.2986	-0.0387
<b>Sad</b>	-0.4168	-0.1224	0.0686	-0.3023	1.0000	-0.2060	0.0817
<b>Ang</b>	-0.3756	-0.1628	0.0739	-0.2986	-0.2060	1.0000	-0.1858
<b>Neu</b>	-0.2598	-0.3568	0.1447	-0.0387	0.0817	-0.1858	1.0000

**Table 4.** The correlation coefficients between the facial expressions for RAF-AU. Sur=Surprise, Fea=Fear, Dis=Disgust, Hap=Happiness, Sad=Sadness, Ang=Anger, Neu=Neutral.

Among these AU annotations, some were one-side AUs. Facial action units are not always the same for left and right halves of faces; they may also vary in intensity. One-sided AUs were labeled as L or R, depending on which half of the face acted. Among the 4,601 faces, we found 219 right-only and 304 left-only AUs. For the unweighted condition (the basic unit was facial expression, and it did not matter if it contained one or more one-sided AUs), we found 163 right-only and 232 left-only AUs. This finding suggests that the left side of the face is more expressive than the right. Psychologists may be interested in these types of characteristics and thus should further investigate this topic. In addition to the left- and right-only AUs, some AUs only contained top or bottom elements. For example, AU23 indicates tightened lips, and sometimes this is only expressed by half (top or bottom). For these instances, we annotated T for top only and B for bottom only. There were 131 T and 179 B faces in RAF-AU.

By annotating for both AU and perceived emotion, we were able to explore which AUs contributed the most to blended emotions. Through a linear transformation, the AUs and perceived emotions were made equivalent in the same space. That is, the AUs of each image were linearly transformed to obtain the expression corresponding to the image. Through matrix transformation, we will be able to get the linear relationship between AUs and perceived emotions. Accordingly, the AUs of each image can be linearly transformed to obtain the expression possibilities corresponding to the image. Table 2 shows the three AUs contributing the most to each facial expression. We explored the inner relationship of the expression itself by using the expression matrix and calculating the relationship of the facial expression based on the AU combination (see Table 3). In addition, we provide the correlation coefficients for the facial expressions in RAF-AU (see Table 4).

## 4 Baseline Evaluation on RAF-AU

### 4.1 Pre-processing and dataset split

We first filtered out all 253 images annotated with “null” as they may contain irrelevant information which would largely distract the AU detection. Then, we chose 13 kinds of AUs that occur more than 8% base rate in RAF-AU dataset for experiment and analysis. Specifically, each image was annotated +1 or -1 if an AU is present or absent, and 0 for one-sided AUs. During pre-processing, manually annotated five facial landmarks provided in RAF-ML [9] were employed to register all images to a reference face using an affine transformation, resulting 100\*100 cropped images. Then gray-scale samples were transformed for the following feature extraction. For dataset split, we divided RAF-AU into training part and test part, where the size of the training part is four times larger than the test part.

### 4.2 Feature extraction and classification

For the comparison purpose, we implemented two handcrafted features and three deep learning features. For hand-crafted feature extraction, we have tried histogram of orientated gradients (HOG) [31] and Local binary patterns (LBP) [32]. For HOG, we used

Layer	1	2	3	4	5	6	7	8	9
Type	Conv	ReLU	MPool	Conv	ReLU	MPool	Conv	ReLU	Conv
Kernel	3	-	2	3	-	2	3	-	3
output	64	-	-	96	-	-	128	-	128
Stride	1	1	2	1	1	2	1	1	1
Pad	1	0	0	1	0	0	1	0	1
Layer	10	11	12	13	14	15	16	17	18
Type	ReLU	MPool	Conv	ReLU	Conv	ReLU	FC	ReLU	FC
Kernel	-	2	3	-	3	-	-	-	-
output	-	-	256	-	256	-	2000	-	13
Stride	1	2	1	1	1	1	-	1	-
Pad	0	0	1	0	1	0	-	0	-

**Table 5.** The network configuration parameters in the AU-CNN.

this shape-based segmentation dividing the image into 10\*10 pixel blocks of four 5\*5 pixel cells with no overlapping. By setting 10 bins for each histogram, we got a 4000-dimensional feature vector per aligned image. For LBP, we selected the 59-bin  $LBP_{8,2}^{u2}$  operator, and divided the 100\*100 pixel images into 100 regions of 10\*10 grid size, which was empirically found to achieve relatively good performance for expression classification.

For deep learning feature extraction, we first employed the already trained baseD-CNN and DBM-CNN provided in [9], then 2000-dimensional deep features learned from raw data can be extracted from the penultimate fully connected layer of these two DCNNs for both training set and test set in RAF-AU. We also tried a multi-label CNN for AU detection. Let us assume that there are  $k$  AU categories and  $n$  training samples. Given a set of ground truth label  $\mathbf{y} \in \{-1, +1, 0\}^k$  and the corresponding prediction results  $\mathbf{p} \in \mathbb{R}^k$  for all  $k$  AU labels, the goal is to minimize the following multi-label cross entropy loss:

$$L = \frac{-1}{n} \sum_{i=1}^n \sum_{j=1}^k \left\{ [y_i^j > 0] \log p_i^j + [y_i^j < 0] \log(1 - p_i^j) \right\}, \quad (2)$$

where  $[ \cdot ]$  is an indicator function that returns 1 if the Boolean expression is true, and 0 otherwise. And  $y_i^j$  is the ground truth for the  $i$ -th sample of  $j$ -th AU,  $p_i^j$  is the predicted probability for the  $i$ -th sample of  $j$ -th AU. We then also extracted the output of the penultimate fully connected layer as the final feature representation, resulting in 2000-dimensional vectors. Table 5 displays detailed network architectures of this AU-CNN trained on the RAF-AU dataset.

During AU detection, support vector machine with linear kernel implemented in LIBSVM [33] was utilized for the one-versus-all binary classification. Given a training set  $\{(x_i, y_i), i = 1, \dots, n\}$ , where  $x_i \in \mathbb{R}^d$  and  $y_i \in \{+1, -1\}$  (samples with one-sided AU has been omitted during this AU detection), then the testing sample can be

AUs	Number	HOG	LBP	BaseDCNN	DBM-CNN	AU-CNN
<b>1</b>	1028	76.83	78.72	72.31	71.74	82.03
<b>2</b>	701	84.67	87.76	83.54	80.67	90.49
<b>4</b>	1808	72.95	74.27	75.67	78.73	86.28
<b>5</b>	975	78.19	76.43	83.25	87.84	88.86
<b>6</b>	404	84.61	80.68	81.98	84.29	87.41
<b>9</b>	749	84.09	88.49	86.79	86.76	90.03
<b>10</b>	1274	78.98	76.63	80.85	81.56	87.89
<b>12</b>	1187	80.51	82.27	86.23	86.17	88.48
<b>16</b>	720	78.47	77.84	77.71	79.67	86.11
<b>17</b>	541	81.81	79.55	80.07	78.29	88.58
<b>25</b>	2830	85.13	86.24	89.87	91.18	95.33
<b>26</b>	1089	66.27	67.72	75.12	78.82	86.31
<b>27</b>	810	91.48	93.11	93.42	93.65	95.77
AVG	-	80.31	80.75	82.06	83.03	<b>88.73</b>

**Table 6.** Performances of the AUC-ROC for AU detection on RAF-AU dataset using different features.

detected by optimizing:

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max(1 - y_i w^T x_i, 0). \quad (3)$$

The penalty parameter  $C$  of SVM was fixed to 1 for all different features.

### 4.3 Evaluation metrics

In terms of evaluating the performance on AU detection, two different metrics were employed: the area underneath the receiver-operator characteristic (ROC) curve (AUC-ROC) and the F1 score. The rank metric ROC curve visualizes the trade-off between sensitivity and specificity by plotting both values as a function of a varying classification threshold. And the threshold metric F1 score is defined as  $F1 = \frac{2RP}{R+P}$ , where  $R$  and  $P$  denote recall (the number of correctly recognized samples divided by the actual number of all samples with the target AU) and precision (the number of correctly recognized samples divided by the total number of samples detected with the target AU), respectively. We then computed the average over all 13 AUs (AVG) to measure the overall performance.

### 4.4 AU detection results

In Table 6 and Table 7, we show the AUC-ROC and F1 score results for each of the 13 AUs in RAF-AU using four different features, respectively. We also list the statistic

AUs	Number	HOG	LBP	BaseDCNN	DBM-CNN	AU-CNN
<b>1</b>	1028	33.70	50.00	29.60	35.42	60.47
<b>2</b>	701	45.00	50.23	38.50	32.49	65.59
<b>4</b>	1808	58.86	57.89	62.35	65.47	73.44
<b>5</b>	975	45.43	37.29	51.53	59.52	69.69
<b>6</b>	404	26.09	26.80	17.78	18.18	58.21
<b>9</b>	749	52.94	59.29	52.24	51.48	67.44
<b>10</b>	1274	49.30	49.65	49.52	51.99	68.41
<b>12</b>	1187	54.72	60.29	64.76	62.09	69.62
<b>16</b>	720	21.65	27.03	25.37	31.63	59.38
<b>17</b>	541	7.27	20.74	18.60	16.26	25.64
<b>25</b>	2830	85.07	85.16	87.50	88.87	92.14
<b>26</b>	1089	20.00	17.90	42.11	48.84	64.65
<b>27</b>	810	68.69	72.11	70.99	71.23	82.67
AVG	-	43.75	47.26	46.99	48.73	<b>65.95</b>

**Table 7.** Performances of the F1 score for AU detection on RAF-AU dataset using different features.

regarding the AU occurrence, i.e., the number of positive samples for each AU. It can be seen that the AU distribution in RAF-AU is more imbalanced than those lab-controlled datasets. With regard to individual AU detection, our baselines can yield relatively good performance on frequently occurring AUs (e.g., AUs 25, 27, 5, 4 and 12). However, we also observe a significant drop in the performance on other less common AUs. When comparing different features, the deep learning feature AU-CNN can achieve comparable and better AU detection rates in terms of both average AUC and F1 score. Nevertheless, when compared to the accuracy achieved on other lab-controlled datasets, there is still room for improvement on this challenging realistic dataset which contains various naturalistic illuminations, occlusions, hear poses and obvious imbalanced distribution.

## 5 Conclusions

In real life, facial expressions occur in the wild, and thus their emotional meaning cannot be absolutely defined. The crowdsourcing approach provides a subjective label based on human perception. Thus, we update the previous RAF-ML database by providing AU coding and removing certain confusing samples. The present database, RAF-AU, provides both AU- and judgement-based annotations from objective and subjective approaches, respectively. Thus, this database provides a fuller picture of given facial expressions. Based on these annotations, we were able to investigate the relationship between objective description and subjective understanding. We also provide a set of baselines for RAF-AU depending on different features. And the deep learning feature

obtained using a multi-label AU detection CNN achieve the best detection rate. Further research should be conducted to study this relationship, with more samples and across various cultures.

## 6 Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under Grants No. 61871052. We thank Yong Ke, Huaxiu Li for their assistance in AU coding.

## References

1. Harrigan, J., Rosenthal, R., Scherer, K.R., Scherer, K.: New handbook of methods in non-verbal behavior research. Oxford University Press (2008)
2. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system: The manual on cd rom. A Human Face, Salt Lake City (2002) 77–254
3. Russell, J.A., Rosenberg, E.L., Lewis, M.D.: Introduction to a special section on basic emotion theory. *Emotion Review* **3** (2011) 363
4. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE (2000) 46–53
5. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE (2010) 94–101
6. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5562–5570
7. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* **PP** (2017) 1–1
8. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* **28** (2018) 356–370
9. Li, S., Deng, W.: Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision* **127** (2019) 884–906
10. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision* **126** (2018) 550–569
11. Yan, W.J., Wu, Q., Liang, J., Chen, Y.H., Fu, X.: How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior* **37** (2013) 217–230
12. Friesen, W.V., Ekman, P., et al.: Emfac-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco **2** (1983) 1
13. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. *Psychiatry* **32** (1969) 88–106
14. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* **111** (2014) E1454–E1462

15. Ruiz, A., van de Weijer, J., Binefa, X.: From emotions to action units with hidden and semi-hidden-task learning. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society (2015) 3703–3711
16. Durán, J., Reisenzein, R., Fernández-Dols, J.M.: Coherence between emotions and facial expressions. *The science of facial expression* (2017) 107–129
17. Kollia, D., Sharmanska, V., Zafeiriou, S.: Face behavior à la carte: Expressions, affect and action units in a single network. *CoRR* **abs/1910.11111** (2019)
18. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. *Pattern recognition* **36** (2003) 259–275
19. Ko, B.: A brief review of facial emotion recognition based on visual information. *sensors* **18** (2018) 401
20. Li, S., Deng, W.: Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020) 1–1
21. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06), IEEE (2006) 211–216
22. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo, IEEE (2005) 5–pp
23. Kollia, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B.W., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vis.* **127** (2019) 907–929
24. Zafeiriou, S., Kollia, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal 'in-the-wild' challenge. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society (2017) 1980–1987
25. Kollia, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. In: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9–12, 2019, BMVA Press (2019) 297
26. Kollia, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first ABAW 2020 competition. *CoRR* **abs/2001.11409** (2020)
27. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* **32** (2014) 692–706
28. Sneddon, I., McRorie, M., McKeown, G., Hanratty, J.: The belfast induced natural emotion database. *IEEE Transactions on Affective Computing* **3** (2011) 32–41
29. Barrett, L.F.: Feelings or words? understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology* **87** (2004) 266
30. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2852–2861
31. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, on. Volume 1, IEEE (2005) 886–893
32. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24** (2002) 971–987
33. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2** (2011) 27:1–27:27 Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.