Multi-Label Compound Expression Recognition: C-EXPR Database & Network

Dimitrios Kollias
Queen Mary University of London, UK
d.kollias@qmul.ac.uk

1. C-EXPR-DB

Data collection All videos of Compound-Expression-DataBase (C-EXPR-DB) have been downloaded from YouTube. For finding videos with compound expressions, we searched YouTube with different expression related keywords, reactions and actions-causes that trigger these expressions. The keywords were either one of the basic or compound categories or synonym words for them, eg for the 'happily surprised' category we searched for 'pleasant surprise', 'joyful surprise', for the basic expressions we searched for 'smile', 'giggle', 'cry', 'rage', 'scared', 'frightened', 'terrified', 'shocked', 'astonished', 'disgust' etc. We also searched for keywords that may cause-trigger the compound expressions, such 'unexpected gift', 'gets pranked', 'gets scared', 'range cry', reactions to movies/trailers/series (such as Game of Thrones)/podcasts/tweets/events (such as United Nations or Oscar Awards)/jokes/(un)pleasant stories/flirt/rejection etc; additionally we searched for people performing activities that can induce such expressions (e.g., riding a rolling coaster).

Data Ethical Considerations The data collection has been conducted under the scrutiny and approval of the Institutional Ethical Committee (IEC).

Sample Images of C-EXPR-DB Fig. 1 shows some sample images of C-EXPR-DB. In these images the in-the-wild nature of the database can be seen, with high variations in head and body poses, gestures, lightning and illumination conditions, backgrounds. Let us mention that all available modalities-components, such as facial expressions, head and body poses, gestures, audio, context-background, have been taken into account when the experts were annotating. Additionally, in each video the person (being annotated) may show multiple compound expressions; this can also facilitate research for generation of compound expressions (since there exist different compound expressions of the same person). Finally, let us mention that our A/V database can foster research on multi-modality learning and on understanding compound expressions from speech signals.

Data Partition C-EXPR-DB is split into subject independent training, validation and test sets.

Annotations Each frame of C-EXPR-DB has been annotated by seven expert annotators; some annotators had psychology background and the rest were computer scientists with a working understanding of facial expressions who had annotated datasets before. Each video has been annotated by all annotators. In terms of AU annotations, these have been provided by a FACS coder and an automatic method similar to the one described in the section 'Generation of Missing Labels for the Auxiliary Task' in Section 2.

Regarding the annotation procedure: all annotators were at first instructed both orally and through a multi-page document on the process to follow for the annotation and how to use the annotation platform. The annotation process provided the freedom to the annotators to watch the videos, pause, re-wind, and then mark the start of an expressional state. The provided document included a list of some well identified expressional cues for all expressions, providing a common basis for the annotation task. For accurately performing frames' annotation, experts exploited all available modalities, namely facial expressions, audio/sound, context, body pose and gesture. On top of that, the experts used their own appraisal of the subject's expressional state for creating the annotations.

The experts also labeled the frames on which the subject is speaking or not speaking. Again instructions have been provided on the multi-page document. The experts labeled speech even if they could only hear it, such as in cases when the mouth was covered, or the head was turned away; other vocal expressions which usually accompany expression, such as laughter (hah), surprise (ooh), disgust (ugh), etc., have also been labeled as speech; in cases where the experts were unable to make a clear judgment, due to bad video/audio quality, or loud background noise, their annotation was 'uncertain'.

In terms of compound annotations, we want to mention that our work follows the definition that compound means that the expression category is constructed as a combination of two basic ones, as introduced in psychology and in var-

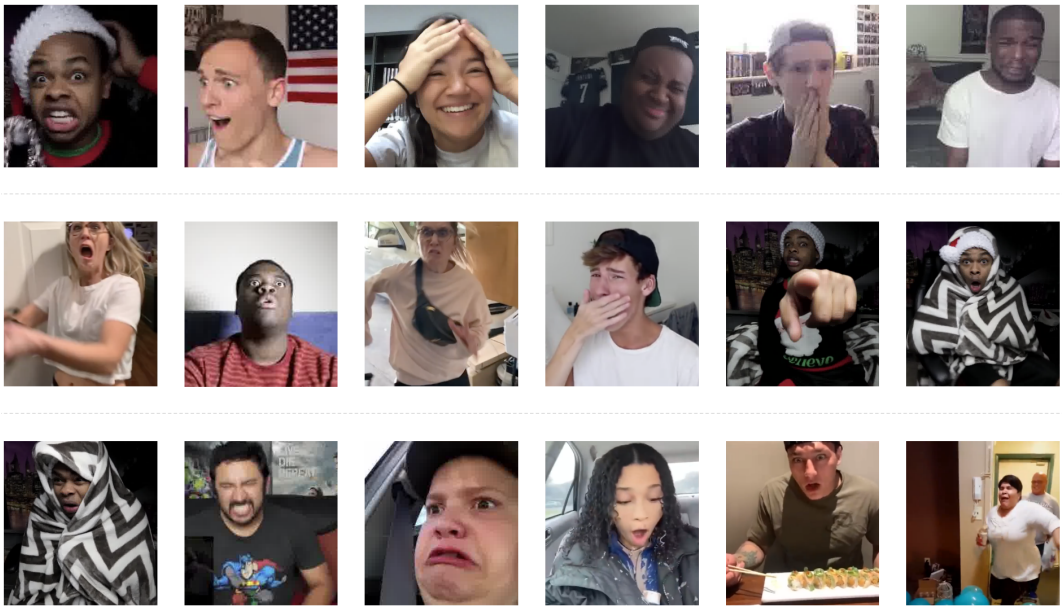

Figure 1. Some sample images from C-EXPR-DB

ious human studies. In this framework, we annotated the database in terms of the 12 compound emotions most typically expressed by humans. We need to mention that there are many challenges that still have to be tackled to fully understand compound emotions and their expressions, e.g., if contempt was found to also be consistently produced by people of distinct cultures, then many more compound categories would be possible.

The category 'other' refers to affective states/classes other than the utilized compound expression ones. We did not explicitly annotate these cases (as we were focusing on only the 12 compound expressions) and thus we put them all into this 'other' category. Such examples include cases where basic expressions or the neutral state occurred, or cases where other expressions than the basic ones occurred (e.g. boredom, confusion, shame, contempt, jealousy, guilt), or cases where there was no desirable agreement between the annotators.

Annotators' Agreement For the compound expression annotations, we kept only the annotations on which at least five (out of seven) experts agreed. For the valence-arousal annotations, the final label values were the mean of the annotations of the seven experts. The mean inter-annotation correlation was 0.68 for valence and 0.66 for arousal.

Facial Attributes and their distribution Table 1 shows the chosen attributes that have been annotated in C-EXPR-DB.

Table 1. C-EXPR-DB's facial attributes: some attributes are binary, whereas others take multiple values

gender	ethnicity	age group	bald
hair colour	hair style	lips size	nose size
eyeglasses	makeup	mustache	beard
wearing hat	wearing lipstick	wearing jewelry	attractive

2. The Proposed Method: C-EXPR-NET

Generation of Missing Labels for the Auxiliary Task In our method, AU labels are needed. These can be either manual or automatic ones. For our experiments, when no manual labels exist, we create a teacher DNN that provides AU annotations. For this, we merge many AU annotated databases so as: i) to have samples annotated in terms of all 17 AUs (namely AU 1,2,4,5,6,7,9,10,11,12,15,17,20,23,24,25,26) and ii) to have an adequate amount of samples that can produce a good AU detection performance. We merge GFT [5], BP4D [10], FERA [11], DISFA [8] and EmotioNet databases (all - except for EmotioNet- are video databases) and we train a CNN-RNN model on them. The CNN part is a modified version of ResNet-50 (that we pre-train on face recognition databases) followed by a GRU (two layers, each with 128 units). As the EmotioNet is a static database, during network training and inference, when loading images from that database, we duplicate each image 'sequence length' times (sequence length is the chosen length for the RNN). Finally we apply a post-processing step for each AU, in which if there exists only one activated AU inside a sequence of 5

consecutive frames; this is corrected to non-activation, and vice-versa. The performance of this model was shown to be either state-of-the-art, or close, over most databases. Consequently, we used this model as a teacher, to provide the AU labels, y_{AU}, of the images x shown in Fig. 1 of the main paper, in which we illustrate the proposed methodology (C-EXPR-NET).

Relatedness between expressions and their associated/activated AUs Table 2 shows the 6 basic expressions along with the AUs that are activated when the particular expression is expressed. This relatedness was found in the psychological study of [4].

Table 2. Relatedness between expressions and their associated/activated AUs from [4]

expression	Activated AUs
happiness	12, 25, 6
sadness	4, 15, 1, 6, 11, 17
fear	1, 4, 20, 25, 2, 5, 26
anger	4, 7, 24, 10, 17, 23
surprise	1, 2, 25, 26, 5
disgust	9, 10, 17, 4, 24

Late Fusion Let r_{AU} and r_{expr} be the AU and expression logits, respectively, i.e., vectors of raw (non-normalized) AU and expression predictions that the fully connected (fc) layers of the Expression and AU Blocks give. The final AU and expression predictions of the multi-modal late fusion (P_{AU} and P_{expr}) are a weighted average of r^v_{AU} and r^v_{expr}, r^a_{AU} and r^a_{expr} of the networks processing the visual (v) and audio (a) modalities (C-EXPR-NET-V and C-EXPR-NET-A), respectively; each weight ($t^v_i, t^a_i, {t'}^v_i, {t'}^a_i$) is equal to the corresponding network's performance on the validation set:

$$P_{AU_i} = \frac{t^v_i \cdot r^v_{AU_i} + t^a_i \cdot r^a_{AU_i}}{t^v_i + t^a_i},$$

$$P_{expr_k} = \frac{{t'}^v_k \cdot r^v_{expr_k} + {t'}^a_k \cdot r^a_{expr_k}}{{t'}^v_k + {t'}^a_k} \quad (1)$$

where $i = 1, \ldots, 17$ and $k = 1, \ldots, 6$.

Training implementation details regarding our proposed methodology At first, the Backbone Network (pre-trained on ImageNet [1]) is trained for compound expression classification; two fully connected layers are put on top of it followed by the output layer. Dropout with (keep) probability 0.8 was applied to the output of the backbone and dropout with probability 0.5 was applied to the first fully connected layer. After training is finished, the fully connected and the output layers are discarded. Then, the Backbone Network is kept fixed (i.e., fixed weights). Then the Expression Branch is being utilized and trained independently. In parallel, the AU Branch is being utilized and trained independently. Once these (i.e., Backbone Network, Expression and AU Branches) are trained independently, they are used for initializing their corresponding counterparts in the whole architecture. Finally, the entire method is being trained in an end-to-end manner.

Regarding the AU Branch: The Intra-AU encoder is a multi-layer transformer encoder that has the same configuration as BERT-Large [3] and is used to encode each AU description; it encodes contextual information for tokens within each sentence. Each layer of the Intra-AU encoder is the same as the vanilla transformer encoder layer [9]. In more detail, the number of layers is set to 24, the hidden size is set to 1024, and the number of heads is set to 16. The Intra-AU encoder is pre-trained and freezed during training of the AU Branch.

The Inter-AU encoder is designed to exchange information across multiple AU embeddings. Like Intra-AU encoder, we also apply the multi-layer transformer network to encode the embeddings from the Intra-AU encoder. In more detail, the number of layers is set to 2, the hidden size is set to 1024, and the number of heads is set to 6. The Inter-Encoder module outputs embeddings, $\mathcal{E}_i$, $i = 1, .., 17$, one embedding per AU descriptor. The size of each such embedding is projected to match the depth dimensions of the AU Block output feature map.

We use the Adam optimizer with initial learning rate of 10^-3, and the learning rate is decayed with momentum 0.85. The batch size used is 200. Training was performed on a Tesla V100 32GB GPU; training time was about 3 hours for RAF-DB and 20 hours for C-EXPR-DB.

3. Experimental Results

Utilized databases To evaluate the proposed method, we perform extensive experiments on RAF-DB and C-EXPR-DB. C-EXPR-DB is one of the paper's contributions and has been presented in the related Section of the paper.

The Real-world Affective Faces database (RAF-DB) [7] contains about 30,000 facial images from thousands of individuals. Each image has been individually labeled about 40 times for the six basic expressions and the neutral one, as well as for eleven compound expressions, namely Happily surprised, Happily disgusted, Sadly fearful, Sadly angry, Sadly disgusted, Fearfully angry, Fearfully surprised, Fearfully disgusted, Angrily surprised, Angrily disgusted, Disgustedly surprised. The training set for the six basic expressions consists of around 12,000 images and the test set of around 3,000 images. The training and test sets for the eleven compound expressions consist of 3,150 and 800 images, respectively.

Data Pre-Processing We used RetinaFace [2] to detect the precise location of the face region in all images of RAF-DB, as well as to detect five facial landmarks, namely the two eyes, the tip of the nose and two corners of the mouth (RAF-DB and C-EXPR-DB). Given these 5 facial landmarks, we performed similarity transformation so as to align the detected and cropped faces in all images.

All cropped and aligned images were resized to $96 \times 96 \times 3$ pixel resolution, their intensity values were normalized to $[-1, 1]$ and random horizontal flip has been applied as a data augmentation step during training.

Performance Metrics The Unweighted Average Recall (UAR) is the performance metric (for compound expression recognition) for RAF-DB. UAR -in other words the average accuracy or in other words the mean diagonal value of the confusion matrix- has been selected by the authors who introduced RAF-DB [6, 7] for evaluating the performance on this database; almost all research works evaluate their results given this criterion and thus we also report this so as to be consistent.

When the classes are mutually exclusive, UAR is defined as:

$$\text{UAR} = \frac{1}{N} \sum_{k=1}^{N} \frac{TP^k}{TP^k + FN^k} \tag{2}$$

where N is the total number of expressions; TP^k is the total true positives for the expression k and FN^k is the total false negatives for the expression k. The UAR score reaches its best value at 1 and its worst score at 0.

For evaluating the performance (for compound expression recognition) of the models on C-EXPR-DB, we choose the F1 score as the performance metric, as this metric is known to be sensitive to imbalanced datasets. When evaluating the performance of the models for AU detection (on RAF-DB and C-EXPR-DB) we report the average F1 score over all action units due to its sensitivity to imbalanced datasets and because it is the main evaluation criterion for AU detection in all known and widely used databases (e.g., DISFA [8], BP4D [10], BP4D+ [11]).

The F_1 score is a weighted average of the recall and precision. The F_1 score reaches its best value at 1 and its worst score at 0. The F_1 score is defined as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

The F_1 score for C-EXPR-DB is computed based on a per-frame prediction (since an expression category or an action unit is specified in each frame).

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 4

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[4] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 3

[5] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 581–588. IEEE, 2017. 2

[6] 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. 4

[7] 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. IEEE, 2017. 3, 4

[8] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013. 2, 4

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[10] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2, 4

[11] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 2, 4