Zero-Shot Audio-Visual Compound Expression Recognition Method based on Emotion Probability Fusion

Elena Ryumina[1], Maxim Markitantov[1], Dmitry Ryumin[1], Heysem Kaya[2], and Alexey Karpov[1]

[1]St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
[2]Department of Information and Computing Sciences, Utrecht University, The Netherlands
{ryumina.e, markitantov.m, ryumin.d, karpov}@iias.spb.su, h.kaya@uu.nl

Abstract

A Compound Expression Recognition (CER) as a subfield of affective computing is a novel task in intelligent human-computer interaction and multimodal user interfaces. We propose a novel audio-visual method for CER. Our method relies on emotion recognition models that fuse modalities at the emotion probability level, while decisions regarding the prediction of compound expressions are based on the pair-wise sum of weighted emotion probability distributions. Notably, our method does not use any training data specific to the target task. Thus, the problem is a zero-shot classification task. The method is evaluated in multi-corpus training and cross-corpus validation setups. We achieved F1 scores of 32.15% and 25.56% for the AffWild2 and C-EXPR-DB test subsets without training on target corpus and target task, respectively. Therefore, our method is on par with methods developed training target corpus or target task. The source code is publicly available from https://elenaryumina.github.io/AVCER/.

1. Introduction

A Compound Expression Recognition (CER) as a part of affective computing is a novel task in intelligent human-computer interaction and multimodal user interfaces. It entails the automated identification of compound emotional states in individuals, which may include combinations of two or more basic emotions such as: Fear, Happiness, Sadness, Anger, Surprise, and Disgust.

Over the last two decades, research efforts in the field of automatic expression analysis have predominantly focused on identifying six basic emotions [12, 34]. However, these methods fail to fully capture the complexity of everyday emotional expressions. Individuals often exhibit Compound Expressions (CEs), such as Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful, Sadly Angry, which are combinations of basic emotions. These CEs underscore the need for more comprehensive models capable of capturing the subtleties inherent in human's emotional expressions.

Existing methods for CER predominantly focus on the visual modality [24, 38]. These methods use both dynamic [24, 38, 39] and static [22, 37, 41] deep models, relying on facial action units [12, 22, 23]. The audio models used in the first two audio-visual methods are based on spectrograms [12, 39]. However, to train a model for CER, it is necessary to have relevant data comprising balanced samples for each class, collected under uncontrolled conditions, containing multimodal data, and being large enough to train deep neural network models. Nevertheless, challenges in annotating CEs [12] contribute to the scarcity of such corpora. An exception is the C-EXPR-DB corpus [10–12, 20], a part of which is presented as the test set in the 6th Workshop and Competition on Affective Behavior Analysis in-the-Wild (ABAW) [21][1]. Another CE corpus is the Multi-modal compound Affective database for facial expression recognition in the Wild (MAFW) [24], access to which is limited.

In this paper, we present a novel method for audio-visual CER. Our method does not utilize the CE labelled data as training data; rather, it includes models trained for basic emotion recognition. Decisions regarding predicted CEs are determined by the pair-wise sum of weighted emotion probability distributions. This approach enables us to address the problem of insufficient publicly available data with CEs. Moreover, CEs comprise various pair combinations of basic emotions, rendering the proposed method particularly valuable to the research community.

In summary, our main contributions are as follows:

- We introduce a novel audio-visual CER method based on basic emotion recognition and analysis of emotion probability distributions through multimodal fusion.

[1]https://affective-behavior-analysis-in-the-wild.github.io/6th/

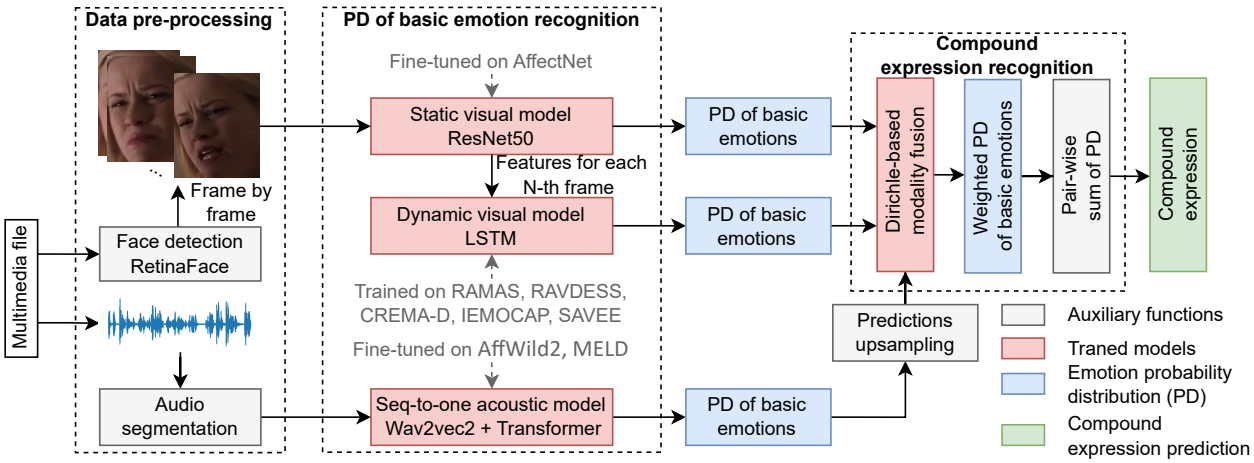

Figure 1. Pipeline of the proposed audio-visual CER method.

- We present a method for audio-visual emotion recognition based on multi-corpus and cross-corpus research.
- We provide new baseline performance scores for the recognition task of the seven basic emotions on the Validation subsets of the AffWild2 [15] and Acted Facial Expressions in The Wild (AFEW) [5] corpora.

The remainder of this paper is organized as follows. In Section 2, we analyze the State-of-the-Art (SOTA) methods for CER. Section 3 outlines our proposed methods. Our research corpora, experimental results and discussions are presented in Section 4. Finally, in Section 5, we summarize the study and consider potential future research.

2. Related Work

In this paper, we conduct research using the C-EXPR-DB corpus for CER. We compare our proposed method with the methods presented in the scope of the 6th ABAW Competition [21]. Several systems for CER were proposed in this challenge, including our contribution [36].

Savchenko [37] presented an audio-visual method based on emotion recognition. The author used the Wav2Vec2 [1] model to extract features from raw audio signals and several static convolutional and transformer models trained on emotion recognition to extract scores and features from the face regions. CER was carried out at the score level and clustering of features.

Yu et al. [43] and Wang et al. [41] proposed visual methods with several static models trained on the target task. In both methods, the authors used the Real-world Affective Faces Database (RAF-DB) annotated with CEs to optimize their models. It is noteworthy that Wang et al. [41] also used the visual language pre-trained models called Claude3[2] to annotate the C-EXPR-DB test samples.

[2]https://www.anthropic.com/claude

Zhang et al. [46] developed an ensemble audio-visual method. Unlike previous methods, this method is based on a dynamic CER model. Convolutional features are extracted from audio spectrograms and face regions, which are then concatenated and fed to transformers. The final CE predictions are based on the voting of predictions of several models. To train the models, the authors used a private corpus annotated with CEs.

In the methods proposed by [41, 43, 46], additional corpora annotated with CEs were used to train models. Therefore, our method is comparable to the methods proposed in [37] because neither uses models trained for the target task and therefore are zero-shot approaches. In general, our method uses a dynamic (spatiotemporal) model next to a static visual model. Moreover, our visual models were not trained on the AffWild2 corpus [15], which is similar to the target corpus. We intentionally did not train visual models on similar corpora to increase and fairly assess their generalizability to new corpora.

3. Proposed Method

A pipeline of the proposed audio-visual CER method is shown in Figure 1. The method accepts a multimedia file as an input. Then, it performs the necessary pre-processing for each modality, including face region detection and audio data extraction. The pre-processed data are input to three models, which output the probability distribution of recognized emotions. A weighting of the probability of three models is then applied. The resulting probabilities are pairwise summed to obtain the probability distribution for CE prediction.

3.1. Video Models

We use the RetinaFace[3] model [4] for face region detection. However, relying only on a detector is insufficient; it is necessary to perform post-processing on the detected face regions, including determining the target person, removing erroneously detected face regions, etc. Since using only a static model is insufficient, for example, transitioning from one expression to another may involve a neutral state or other intermediate states [35]. Our method integrates static and dynamic visual models to recognize CEs.

Static visual model (VS). As a static model for affective state recognition (comprising six basic emotions and a neutral state), we utilize the ResNet50 model [8] pre-trained on face recognition[4]. The model extracts discriminative features from faces useful for transfer learning in our task. We initialize the model with pre-trained weights and then fine-tune it to recognize affective states without freezing its layers. We extend the model for feature extraction and classification by adding two Fully Connected Layers (FCLs) comprising 512 and 7 neurons, respectively. In the proposed method, we use a static model to detect affective states in each frame and extract features from every N^{th} frame (where N is the frame step size). We use these features as inputs for the dynamic model.

Dynamic visual model (VD). The model designed for analyzing dynamically changing affective states operates on 2-second segments or 10 frames. To produce 10 frames within two seconds, the frame rate of each video is reduced to five Frames Per Second (FPS). The proposed model comprises two Long Short-Term Memory (LSTM) layers, with 512 and 256 neurons, respectively. It also includes a classification layer consisting of 7 neurons.

To enhance the generalizability of the video models, several augmentation techniques are applied, namely MixUp [45] and Label Smoothing [29]. These techniques help reduce the models' confidence levels in their basic emotion predictions, enabling them to identify multiple emotions with varying degrees of certainty in the frames. All the models are trained using the Adam optimizer with a learning rate of 1e-4 for 30 epochs and the Cosine Annealing Cold Restart Learning Scheduler [26] with five rate restart cycles. More details on the training of the visual models are presented in [34].

3.2. Audio Model

In addition to extracting audio signals from multimedia files, we detect voice activity, which is applied during the training and validation and not during testing, where we need frame-wise predictions. Two approaches are used for this purpose, depending on the corpus used. The first one employs an audio-based Voice Activity Detection (VAD)[5]. The second one relies on the video modality, analyzing video data frame by frame. We extract facial landmarks using MediaPipe [28], subsequently, the mouth landmarks are identified and the region of interest is extracted. We use it to determine whether the target speaker's mouth is open or closed. We employ this method due to the specificity of the training acoustic data, which may include background noises, making it challenging to identify the target speaker. Then, 4-second segments with a step of two seconds are formed on the detected segments of voice activity. In addition, to obtain the target label of a window, we compute the most frequent frame-wise label.

Sequence-to-One acoustic model. We proposed two slightly different models. The backbone of both models is the pre-trained public dimensional emotional model (PDEM) [40] that is based on the Wav2Vec2 model. This model was pre-trained using the regression emotion dimensions (arousal, valence, and dominance) from the MSP-Podcast corpus [27]. On top of the model, we stack two transformer layers with self-attention mechanisms, each with 32 and 16 heads. After the last transformer layer, we aggregate the information along the time axis and apply a FCL with seven or eight neurons, depending on the number of classes. We fine-tune all the layers from the top to the last two (W2V2-7cl) or four (W2V2-8cl) encoding layers of the backbone model for models with seven and eight neurons, respectively.

In the vein of the video model, Label Smoothing [29] is also used for the audio model to reduce the confidence of the model. The remaining training hyper-parameters are identical to those of the video model.

3.3. Modality Fusion

The proposed modality fusion method uses three models to represent emotion probability distributions. Each model exhibits varying prediction confidences for different emotions. Therefore, we employ a hierarchical probability weighting before predicting CEs. The importance of models and probabilities is considered in the first weighting. The weight matrix W is generated using the Dirichlet distribution:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1C} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M1} & w_{M2} & \cdots & w_{MC} \end{bmatrix}, \quad (1)$$

where $W \in \mathbb{R}^{M \times C}$, M is the number of models and C is the number of emotion classes. The weight matrix is generated such that the weights for the three models of each class sum up to one. After the first weighting, new probabilities for each model are calculated using the following formula:

$$\bar{P}_M = P_M \times w_M, \quad (2)$$

[3] https://github.com/hhj1897/face_detection
[4] https://github.com/rcmalli/keras-vggface
[5] https://github.com/snakers4/silero-vad

where $P_M = [p_{M1}, p_{M2}, ..., p_{MC}]$ is the probability vector for the model M, $w_M = [w_{M1}, w_{M2}, ..., w_{MC}]$ is the weight vector for the model M. In the second weighting, only the importance of the models is taken into account. A weight vector V of size M is generated with one value for each model. The vector values are generated in the range [0.01, 0.5] with an increment of 0.005. A final probability vector $\hat{P}$ is obtained using the following formula:

$$\hat{P} = \sum_{i=1}^{M} \bar{P}_i \times v_i, \tag{3}$$

where $v_i \in V$. W and V are weights that remain consistent across all test samples. This hierarchical weighting enhances performance measures for both basic emotion recognition and CE recognition by considering the contribution of each constituent model. The final probability vector is then used for CER.

3.4. Rule-based Decision-Making Method

We apply two rules to make decisions regarding the predicted CEs. The first rule (Rule 1) allows for certain emotion predictions and is based on masking probabilities that fall below the minimum threshold (1/7, since C=7) for emotion prediction. This rule is exclusively applied to the outputs of Dirichlet-based fusion method, e.g., $\bar{P}$ probability vectors. The probability vector is updated according to the following condition:

$$\bar{P}_z = \begin{cases} 0, & \bar{p}^E < 1/7 \\ \bar{p}^E, & otherwise \end{cases}, \tag{4}$$

where $\bar{p}^E$ is the probability of the emotion E. For the Rule 1, according to the probabilities of the basic emotions E_1 and $E_2 \in$ {Neutral (Ne), Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su)}, the CE probability $\bar{cp}^{E1,E2}$ is calculated using a simple pair-wise probability sum:

$$\bar{cp}^{E1,E2} = \bar{p}_z^{E1} + \bar{p}_z^{E2}. \tag{5}$$

The second rule (Rule 2) is based on weighting the frequency of emotions occurring in CEs. In the CE Recognition Challenge, we aim to develop a method for recognizing seven CEs: Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful, and Sadly Angry. The occurrence frequency of each emotion differs among these pairs. For example, the emotion of Surprise is more frequent than the others; therefore, this emotion does not allow distinguishing between the respective CEs. We assumed that the use of emotion weights can enhance the importance of the probability of less represented emotions. The weight vectors, CW_1 and CW_2, determine the weights of the first and second emotions in pairs of CEs. The weight vectors, whose paired values sum to one (see Table 1), are determined a priori considering the frequency of the basic emotions in CEs.

Table 1. Weights of basic emotions used for CER. E_1 and E_2 refer to the first and second emotion in a pair, CW_1 and CW_2 to the weights for the first and second emotion in a pair.

CE class	E_1	CW_1	E_2	CW_2
Fearfully Surprised	Fear	5/7	Surprise	2/7
Happily Surprised	Happiness	6/8	Surprise	2/8
Sadly Surprised	Sadness	4/6	Surprise	2/6
Disgustedly Surprised	Disgust	6/8	Surprise	2/8
Angrily Surprised	Anger	5/7	Surprise	2/7
Sadly Fearful	Sadness	4/9	Fear	5/9
Sadly Angry	Sadness	4/9	Anger	5/9

To exemplify, using Rule 2, the probability value for the Fearfully Surprised CE, $\bar{cp}^{Fe,Su}$, is calculated based on the established weights (see Table 1) and on the probabilities of the basic emotions using the formula:

$$\bar{cp}^{Fe,Su} = \bar{p}^{Fe} \times cw_1^{Fe,Su} + \bar{p}^{Su} \times cw_2^{Fe,Su}, \tag{6}$$

where $cw_1^{Fe,Su} \in CW_1$ and $cw_2^{Fe,Su} \in CW_2$. This rule is applied to the outputs of both Dirichlet-based ($\bar{P}$) and the hierarchical modality fusion ($\hat{P}$) methods.

We also perform rule-free CER. To do this, we calculate each CE probability by the pair-wise sum of emotion probabilities ($\bar{P}$ or $\hat{P}$) using equation 5.

4. Experiments

In this section, we describe the research corpora, present the experimental results, and discuss them.

4.1. Research corpora

The proposed audio-visual CER method is based on the models designed for recognizing six basic emotions and a neutral state.

Table 2. Summary information about the samples of the corpora used for research.

Corpus	# Samples Train / Val. / Test subset	# Hours Train / Val. / Test subset
AffectNet [30]	283901 / 3500 /–	– / – / –
RAMAS [31]	1867 / – / –	≈05:00 / – / –
RAVDESS [25]	4152 / – / –	≈04:50 / – / –
CREMA-D [3]	3843 / – / –	≈01:10 / – / –
IEMOCAP [2]	5422 / – / –	≈07:00 / – / –
SAVEE [7]	480 / – / –	≈00:30 / – / –
AffWild2 [15]	248 / 70 / 228	≈11:00 / ≈04:20 / ≈10:00
MELD [32]	9988 / – / –	≈09:00 / – / –
AFEW [5]	– / 383 / –	– / ≈00:15 / –
C-EXPR-DB [12]	– / – / 56	– /– / ≈00:16

We use several corpora for training, validating and testing the developed emotion recognition models. To train a

Table 3. Experimental results of basic emotions and CE recognition.

ID	Model	Training corpus/corpora	Test corpus						
			AffWild2 (7cl)		AFEW (7cl)		C-EXPR-DB (7cl), F1		
			F1	UAR	F1	UAR	Rule 1	Rule 2	W/o rules
1	Static visual model, ResNet50	AffectNet	34.71	40.33	42.83	43.75	20.34	19.58	22.97
2	Dynamic visual model, LSTM	RAMAS, RAVDESS, CREMA-D, IEMOCAP, SAVEE	39.71	42.44	41.82	43.59	13.08	13.57	16.48
3	Models 1 & 2 (Dirichlet-based weighing)	–	40.95	45.64	43.37	43.98	17.53	17.62	20.87
4	Models 1 & 2 (hierarchical weighing)	–	41.38	46.19	43.74	44.84	–	19.44	20.98
5	Seq-to-one acoustic model, W2V2-7cl	AffWild2, MELD	31.11	30.76	22.88	26.81	11.97	12.85	14.03
6	Seq-to-one acoustic model, W2V2-8cl	AffWild2, MELD	31.56	33.64	22.83	25.95	10.37	9.93	12.10
7	Models 1 & 2 & 5 (Dirichlet-based weighing)	–	44.51	49.51	43.86	44.66	19.15	22.03	**25.27**
8	Models 1 & 2 & 5 (hierarchical weighing)	–	42.77	49.98	43.09	43.55	–	17.56	18.95
9	Models 1 & 2 & 6 (Dirichlet-based weighing)	–	**46.79**	**51.79**	**44.81**	**45.66**	21.14	22.01	**25.26**
10	Models 1 & 2 & 6 (hierarchical weighing)	–	39.36	46.60	35.76	37.70	–	14.87	14.54

static video model, we use the AffectNet corpus [30]. This corpus comprises an extensive collection of static facial images displaying spontaneous emotions. We use the RAMAS [31], RAVDESS [25], CREMA-D [3], IEMOCAP [2] and SAVEE [7] corpora to train the dynamic visual model. In contrast to AffectNet, these corpora were collected in office conditions, but contain dynamically changing expressions. Therefore, the annotation quality is considered reliable, the facial images are more or less frontal and not occluded. The multi-corpus training is deemed to improve the model's generalization ability to new data.

To train the acoustic model, we conduct a multi-corpus training using the AffWild2 [13–19, 44] and MELD [32] corpora. These corpora comprise recordings collected in uncontrolled conditions and include various paralinguistic elements in speech (such as laughter, shouting, etc.), making the data more relevant to real-world scenarios in contrast to the aforementioned corpora [6].

Validation of the acoustic and visual models and the optimization of the modality fusion weights are conducted on the Validation subset of the AffWild2 corpus (the version used in ABAW 2024). To avoid overfitting the models and fusion weights for each corpus, we use the AFEW Validation subset [5] as an additional validation corpus. Finally, the CER method is tested on the sequestered test subset of the C-EXPR-DB corpus. To eliminate noise in the training data, we select samples for which the annotation confidence is above 60% for the RAMAS, CREMA-D, and IEMOCAP corpora. The negative impact of such data on the generalizability of the models is described in [33]. The general information about the corpora used is summarized in Table 2.

4.2. Experimental Results

To evaluate the effectiveness of the proposed method, we use standard metrics for unbalanced data such as macro F1-score (F1) and Unweighted Average Recall (UAR). The experimental results are presented in Table 3. Depending on the corpus, the visual models show different performance;

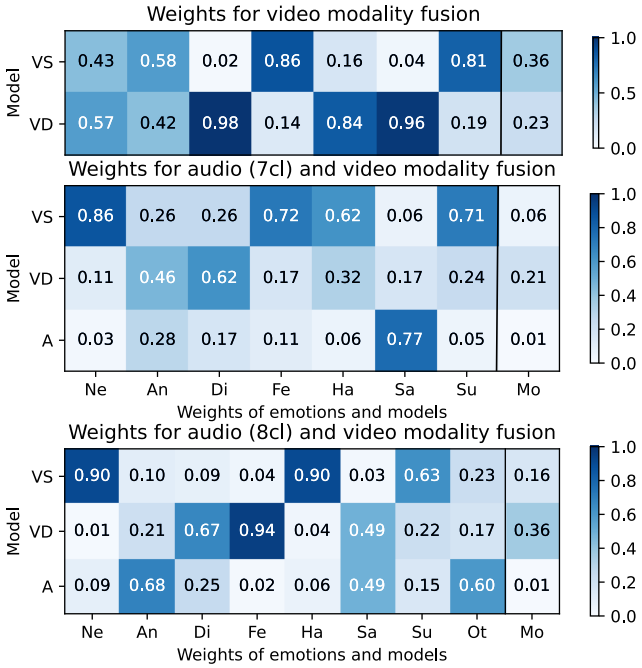

Figure 2. Weights for different modality fusion. VS, VD, and A refer to static visual, dynamic visual, and acoustic models, respectively. Ne, An, Di, Fe, Ha, Sa, Su, Ot refer to the weights of seven and others emotions used for Dirichlet-based weighting, Mo to the weights of models used for hierarchical weighting.

the dynamic model outperforms the static model for AffWild2, and vice versa for AFEW and C-EXPR-DB. Using rules on C-EXPR-DB reduces CER performance for both unimodal models and the multimodal fusion systems. This result is unexpected, since the proposed rules are intended to enhance CER performance by balancing the contribution of each emotion in a pair of emotions. When the two visual models are combined, the hierarchical weighting fusion is slightly better than the Dirichlet-based weighting for both tasks. However, counter to our expectations, when the two

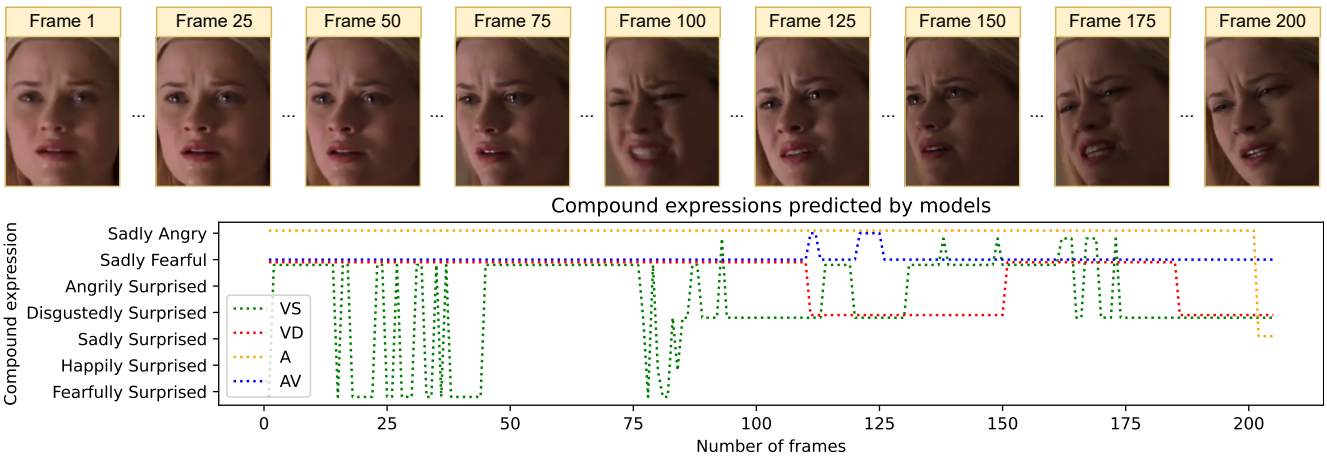

Figure 3. An example of CE prediction using video from the C-EXPR-DB corpus. VS, VD, A and AV refer to static visual, dynamic visual, acoustic, and audio-visual models, respectively.

visual models are combined, the CER performance drops by 1.99% compared to using only the static model.

We then evaluate the performance of the acoustic models on the research corpora. Both acoustic models demonstrate almost identical performance on AffWild2 and AFEW, but on C-EXPR-DB, Model 5, trained to recognize seven emotions, outperforms Model 6, trained to recognize eight emotions, by 1.93%.

Finally, we fuse the acoustic and visual models by comparing various weighting schemes. The results show that the Dirichlet-based fusion outperforms the hierarchical weighting fusion on all research corpora. While Model 9 outperforms Model 7 for emotion recognition, they exhibit equal performance on the C-EXPR-DB corpus. The fusion results of the three models indicate that the acoustic model significantly contributes to improving the method's performance. At the same time, we consider Model 9 with an acoustic model trained for eight emotion recognition as the most generalizable to new corpora, since it demonstrates the maximum average performance across the research corpora. Thus, we present novel cross-corpus recognition benchmark F1 performances of 46.79% and 44.81% for seven-class emotion recognition tasks on the AffWild2 and AFEW corpora, respectively. It is worth noting that the latest cross-corpus UAR performance obtained by the visual model on the AFEW corpus is 39.56% [9], which is 5.28% lower than our visual model's performance (39.56% vs. 44.84%).

To understand the significance of each model in the final CE predictions, the fusion model weights are presented in Figure 2. The analysis of the fusion weights involving only the two visual models indicates a preference towards the dynamic model for predicting Di, Ha, and Sa emotions, while favoring the static model for the predicting Fe and Su emotions. The hierarchical weighting reduces the contribution of the dynamic model. This weight distribution suggests that considering the CE weighting (see Table 1), the method bases its decision on the dynamic model for predicting, for example, the Disgustedly Surprised and Happily Surprised classes, whereas it relies on the static model for predicting classes like Fearfully Surprised and Sadly Fearful.

The acoustic model trained on seven classes contributes less to the final prediction than the model trained on eight classes. For example, the first model has a strong impact in predicting only the Sa emotion, whereas the second model demonstrates a higher contribution in predicting An and competes with the dynamic video model in predicting Sa. Nevertheless, in both cases, the hierarchical weighting leads to a complete disregard of the acoustic model, consequently leading to a decrease in emotion recognition performance. Thus, when combining three models, using hierarchical weighting of emotion probability distribution proves ineffective. The latter fusion (see Figure 2, bottom sub-figure) demonstrates that the method bases its decision on the acoustic model when predicting, for example, the Angrily Surprised and Sadly Angry classes, while relying on the static model to predict the Happily Surprised class. The contribution of the dynamic model is considered when predicting other CEs.

We also show an example of CER using Model 9 in Figure 3. From the frames depicted, it is clear that the woman is experiencing negative CEs; none of the models make a mistake by predicting Happily Surprised. The VS model predicts four CEs, including Fearfully Surprised, Disgustedly Surprised, Sadly Fearful, and Sadly Angry. On the other hand, the VD model predicts two of them, namely Sadly Fearful and Disgustedly Surprised, while the acoustic model (A) predicts only Sadly Angry. In this case, the fusion of the emotion probabilities of all three models predicts two CEs, Sadly Fearful and Sadly Angry. Subjectively, we have a tendency to assume that the video represents CEs

Table 4. Performance comparison (F1, %) of the SOTA methods.

Method	Validation subset		Test subset	
	AffWild2 (7cl)	AffWild2 (8cl)	AffWild2 (8cl)	C-EXPR-DB (7cl)
Zhang et al. [46]	–	55.55	50.05	55.26
Savchenko [37]	38.30	43.40	34.14	27.08
Yu et al. [42, 43]	–	44.43	35.34	22.40
Ryumina et al. [36] (Ours)	–	–	–	22.01
Wang et al. [41]	–	–	–	18.45
Ours (w/o rules)	46.79	35.98	32.15	25.56

such as: Sadly Fearful, Sadly Angry, Angrily Disgusted (this CE is not in the target classes).

A comparison between the results obtained using Model 9 without applying rules and the SOTA results proposed as a part of the 6th ABAW Competition is presented in Table 4. Only two different methods including ours [36, 37] do not use models trained for the CER task, the rest methods [41, 43, 46] employ a task-specific training. For the seven emotion recognition on the Validation subset of the AffWild2 corpus, our method outperforms the method [37] by 8.49% (46.79% vs. 38.30%). However, for the eight emotion recognition on both the Validation and Test subsets of the same corpus, our method performs poorer than all methods. This outcome indicates that our model is reliable in predicting basic emotions, but is inferior in predicting non-basic emotions. This is because the visual models are not trained to recognize other emotions in the samples of the AffWild2 corpus, unlike the other proposed methods [37, 41, 43, 46]. For the same reason, we achieved 1.52% (25.56% vs. 27.08%) lower F1 in CER on the C-EXPR-DB corpus compared to the method proposed in [37]. This is due to the similar recording conditions and annotation methodology used for both the target C-EXPR-DB corpus and the AffWild2 corpus.

Thus, our results and comparisons with the SOTA methods show that we have developed one of the best audio-visual methods for basic emotion recognition. Additionally, we can claim that the developed method is suitable for the zero-shot CER as it comprises three emotion prediction models, with each assigned responsibility for predicting its respective class during CER.

5. Conclusions

In this paper, we propose a novel audio-visual method for CER. The method integrates three models, including the static and dynamic visual models, as well as the acoustic model. Each model predicts the emotion probabilities for six basic emotions and the neutral state. The emotional probabilities are then weighted using the Dirichlet distribution. Finally, a pair-wise sum of weighted emotion probability distributions is applied to determine the compound emotions. Additionally, we provide novel cross-corpus recognition benchmark F1 performances of 46.79% and 44.81% for seven emotion recognition on the Validation subsets of the AffWild2 and AFEW corpora, respectively.

The experimental results obtained for CER demonstrate that each model is responsible for predicting specific CEs. For example, the acoustic model is responsible for predicting the Angry Surprised and Sadly Angry, the static visual model is responsible for predicting the Happily Surprised class, and the dynamic visual model predicts other CEs well. Using our proposed method, we obtain an F1 score of 25.56% for CER on the C-EXPR-DB corpus. In our future research, we aim to improve the generalization ability of the proposed method by adding the text modality and increasing the number of heterogeneous training corpora for multi-corpus and cross-corpus studies.

6. Acknowledgements

This research was partially supported by RSF in the framework of the project No. 22-11-00321 (works by E. Ryumina, M. Markitantov, D. Ryumin, and A. Karpov).

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 2

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 4, 5

[3] Houwei Cao, David G Cooper, Michael K Keutmann, et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4): 377–390, 2014. 4, 5

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, et al. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 3

[5] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *International Commission on Mathematical Instruction (ICMI)*, pages 546–550, 2019. 2, 4, 5

[6] Denis Dresvyanskiy, Maxim Markitantov, Jiawei Yu, et al. Multi-modal arousal and valence estimation under noisy conditions. *CVPRW*, page in print, 2024. 5

[7] Sanaul Haq, Philip JB Jackson, and James Edge. Audio-visual feature selection and reduction for emotion classification. In *Int. Conf. on Auditory-Visual Speech Processing (AVSP)*, pages 185–190, 2008. 4, 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[9] Ryosuke Kawamura, Hideaki Hayashi, Noriko Takemura, and Hajime Nagahara. Midas: Mixing ambiguous data with soft labels for dynamic facial expression recognition. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 6552–6562, 2024. 6

[10] Dimitrios Kollias. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *CVPR*, pages 2328–2336, 2022. 1

[11] Dimitrios Kollias. ABAW: Learning from synthetic data & multi-task learning challenges. In *ECCV*, pages 157–172, 2023.

[12] Dimitrios Kollias. Multi-label compound expression recognition: C-EXPR database & network. In *CVPR*, pages 5589–5598, 2023. 1, 4

[13] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, pages 1–15, 2019. 5

[14] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, pages 1–20, 2021.

[15] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second ABAW2 competition. In *CVPR*, pages 3652–3660, 2021. 2, 4

[16] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, pages 1–11, 2019.

[17] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, et al. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *IJCV*, pages 1–23, 2019.

[18] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 794–800, 2020.

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, pages 1–15, 2021. 5

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *CVPR*, pages 5888–5897, 2023. 1

[21] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, et al. The 6th affective behavior analysis in-the-wild (ABAW) competition. *arXiv preprint arXiv:2402.19344*, pages 1–10, 2024. 1, 2

[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017. 1

[23] Ximan Li, Weihong Deng, Shan Li, and Yong Li. Compound expression recognition in-the-wild with AU-assisted meta multi-task learning. In *CVPR*, pages 5734–5743, 2023. 1

[24] Yuanyuan Liu, Wei Dai, Chuanxu Feng, et al. MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *ACM MM*, pages 24–32, 2022. 1

[25] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 4, 5

[26] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, pages 1–16, 2017. 3

[27] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017. 3

[28] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, pages 1–9, 2019. 3

[29] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458, 2020. 3

[30] 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, 2017. 4, 5

[31] Olga Perepelkina, Evdokia Kazimirova, and Maria Konstantinova. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer*, pages 501–510, 2018. 4, 5

[32] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536, 2019. 4, 5

[33] Elena Ryumina, Oxana Verkholyak, and Alexey Karpov. Annotation confidence vs. training sample size: Trade-off solution for partially-continuous categorical emotion recognition. In *Interspeech*, pages 3690–3694, 2021. 5

[34] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514: 435–450, 2022. 1, 3

[35] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, and Alexey Karpov. Ocean-ai framework with emoformer cross-hemiface attention approach for personality traits assessment. *Expert Systems with Applications*, 239:122441, 2024. 3

[36] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, et al. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, pages 1–7, 2024. 2, 7

[37] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, pages 1–10, 2024. 1, 2, 7

[38] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. MAE-DFER: Efficient masked autoencoder for self-supervised dy-

namic facial expression recognition. In *ACM MM*, pages 6110–6121, 2023. 1

[39] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. HiCMAE: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *arXiv preprint arXiv:2401.05698*, pages 1–19, 2024. 1

[40] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE TPAMI*, pages 1–13, 2023. 3

[41] Jiahe Wang, Jiale Huang, Bingzhao Cai, et al. Zero-shot compound expression recognition with visual language model at the 6th abaw challenge. *arXiv preprint arXiv:2403.11450*, pages 1–4, 2024. 1, 2, 7

[42] Jun Yu, Zhihong Wei, and Zhongpeng Cai. Exploring facial expression recognition through semi-supervised pretraining and temporal modeling. *arXiv preprint arXiv:2403.11942*, pages 1–8, 2024. 7

[43] Jun Yu, Jichao Zhu, and Wangyuan Zhu. Compound expression recognition via multi model ensemble. *arXiv preprint arXiv:2403.12572*, pages 1–6, 2024. 2, 7

[44] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, et al. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *CVPRW*, pages 1980–1987, 2017. 5

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, pages 1–13, 2017. 3

[46] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, et al. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, pages 1–11, 2024. 2, 7