DEER: Deep Emotion-sets for fine-grained Emotion Recognition

Sharjeel Tahir | Nima Mirnateghi | Syed Afaq Shah | Ferdous Sohel
Edith Cowan University | Murdoch University

Harnessing the Power of Image-Sets: A Highly Reliable Approach to Emotion Recognition That Outperforms Traditional Techniques—No Data Augmentation Required!

Introduction

In this research, we propose emotion-sets as a unique encoding for face image data (with various people and face angles)to classify emotion classes, as opposed to the conventional single-image-based classification. For each image in an emotion-set, prediction confidence against each emotion is utilized as a vote. The results are generated by a combination of two distinct voting methods, including Majority Voting and Weighted Voting. The proposed method achieves state-of-the-art (SOTA) accuracy on the Facial Emotion Recognition 2013 (FER2013), Cohn Kanade (CK+), and Facial Emotion Recognition Group (FERG) datasets without using techniques like data augmentation, feature extraction, or extra training data, which are used by several SOTA works. Our experimental findings indicate that the proposed emotion-set classification yields more accurate results than the current SOTA FER methods.

Our Key Contributions

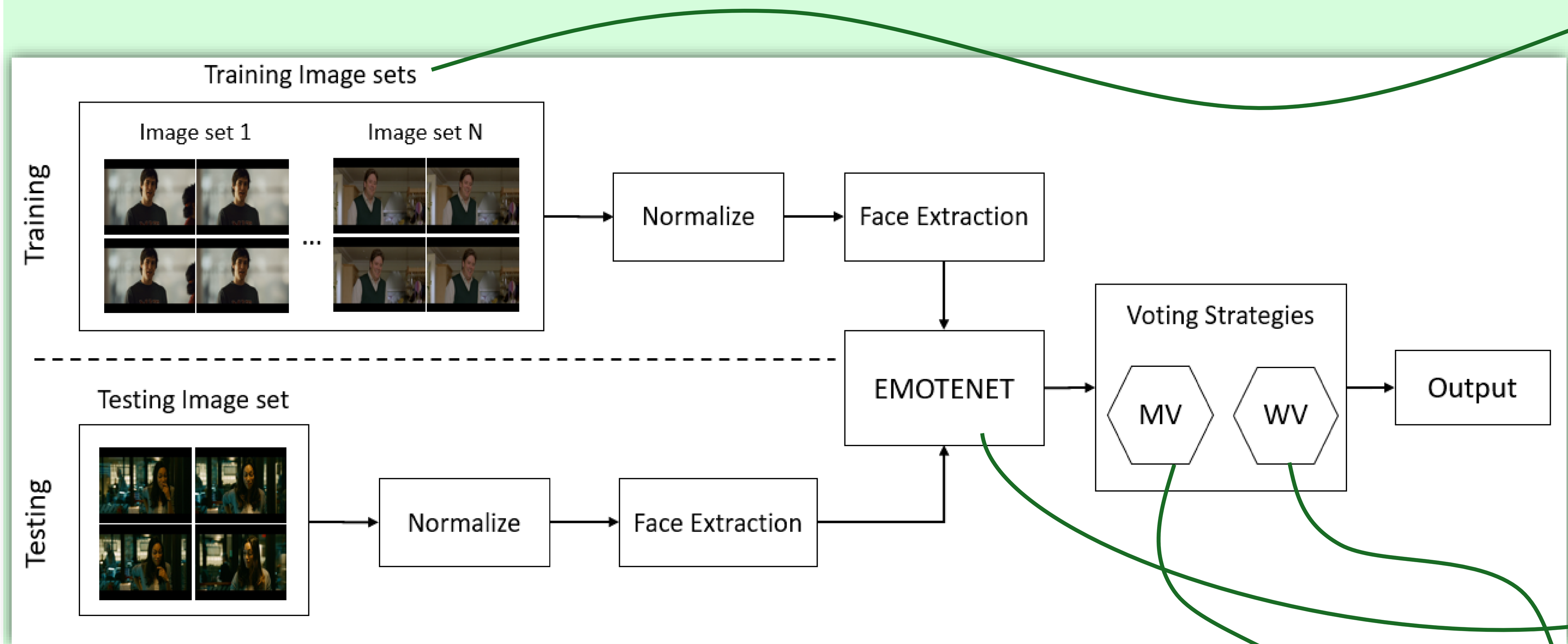
- A novel deep learning-based image set classification for the task of FER has been proposed.
- Our proposed technique outperforms the state-of-the-art frameworks on FER2013, CK+ and FERG datasets with accuracy reaching as high as 100%.

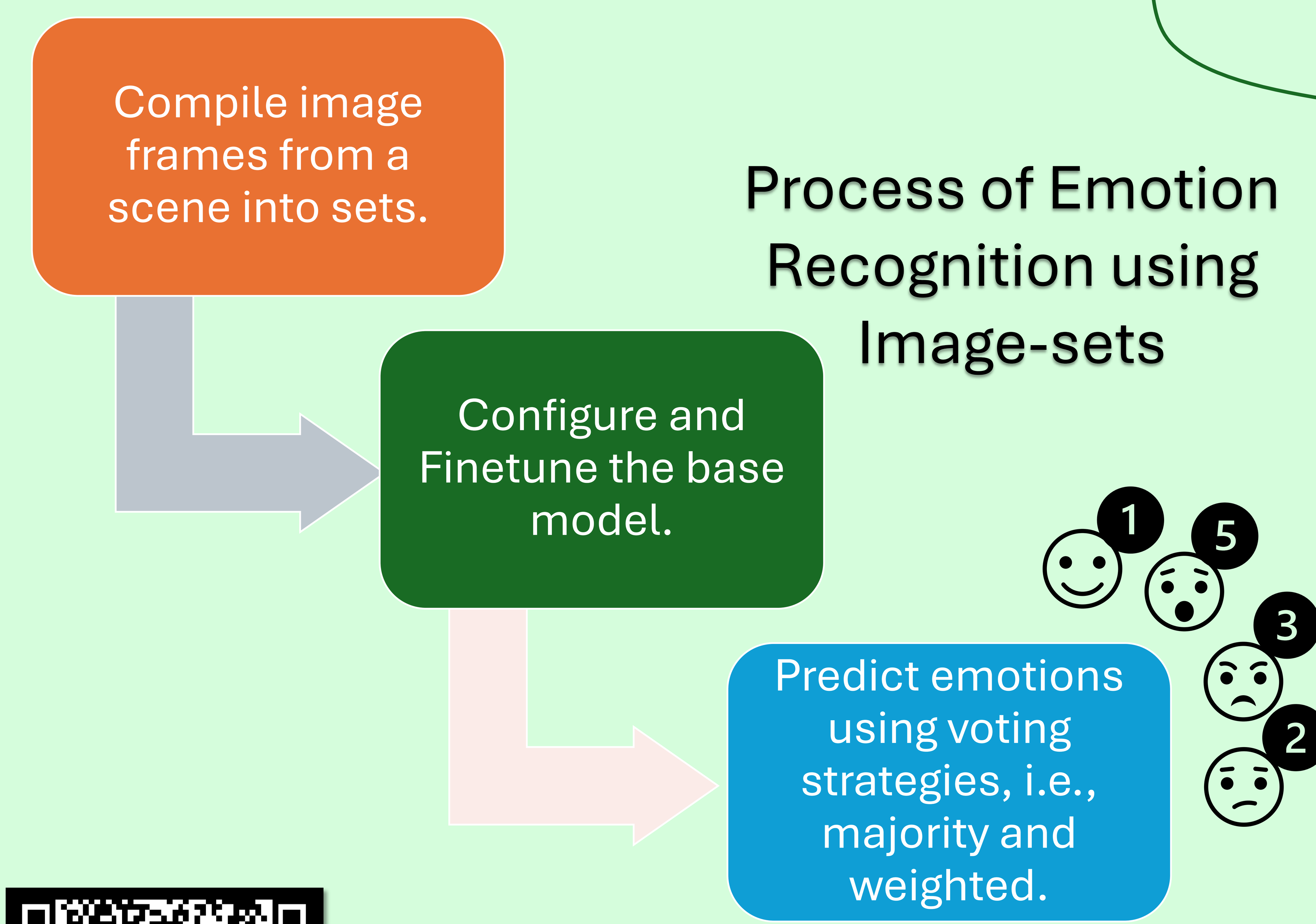
What is image set classification?

In image set classification, multiple images of a given subject are grouped together to form sets. These sets are then used for the training and testing. It enables the model to capture additional information while providing robustness against issues such as, occlusion, pose variance, illumination and others within the images. Because there are deeper features to be extracted from multiple images of each subject, the model can generate higher correlation between images of similar classes.

Conclusion and Future Proposal

To the best of our knowledge, previously, the works directed in this field have used single-image or video-based inputs. We show, with detailed experimentation and analysis, how the proposed image set classification can improve the accuracy and efficiency of FER. Our proposed framework outperforms the SOTA records on FER2013, CK+ and FERG datasets, in addition to achieving superior results on the SFEW dataset. One of the major hurdles we observed in the process are scarcity of images collected in natural settings. Hence, future works can be dedicated for the collection of more in-the-wild facial data.

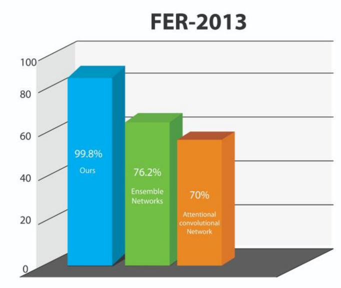




Process of Emotion Recognition using Image-sets

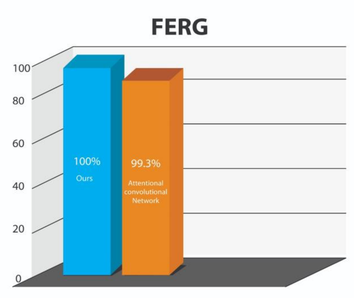
Scan here if you are curious
(❁´◡`❁)

Results on SOTA Datasets



Method	Year	Accuracy
VGGNet	2021	73.28%
Ensemble of 7 CNNs	2019	76.2%
MSAU-Net	2020	78.3%
EMOTENET (Ours)	**2023**	**99.6%**

Method	Pre-trained Dataset	Year	Accuracy
Island Loss	FER2013	2018	52.52%
Identity-aware CNN	FER2013	2017	50.98%
Multiple deep CNNs	FER2013	2015	55.96%
RAN-ResNet18	MSCeleb	2019	54.19%
RAN(VGG,ResNet)	MSCeleb	2019	56.4%
MSAU-Net	MSCeleb	2020	57.4%
EMOTENET (Ours)	MSCeleb	2023	56.2%



Method	Year	Accuracy
DeepExpr	2016	89.02%
Ensemble Multi-feature	2018	97%
Adversarial NN	2018	98.20%
Attentional CNN	2019	99.30%
EMOTENET (Ours)	**2023**	**100%**

Overall Results

Average of 5x folds

Dataset	MV	WV
SFEW	52%	56.2%
FER-2013	99.6%	98.1%
FERG	100%	100%
CK+	100%	100%

Image-set Formulation

Let X is an image set that contains multiple images M, where the number of images within an image set is T.

$$X_i = \{M_1, M_2, M_3...., M_T\}$$

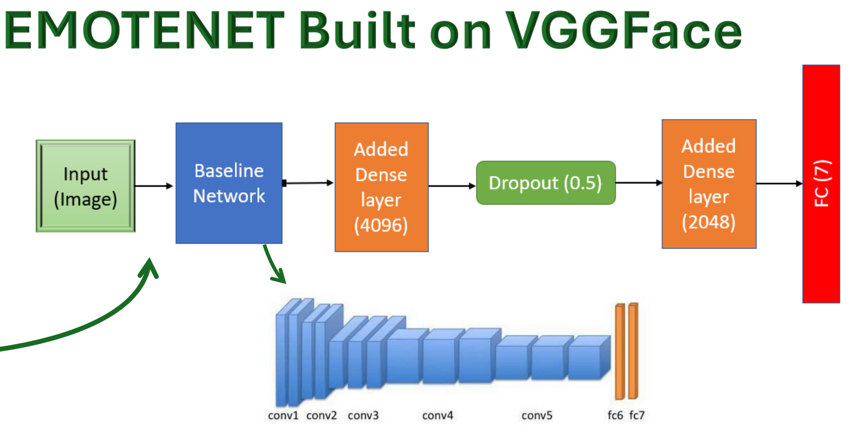
where $i = \{1, 2, 3, ..., N\}$. Similarly, a gallery G, where the total number of image sets is N can be represented as;

$$G_\Delta = \{X_1, X_2, X_3, ..., X_N\}$$

where $\Delta = \{1, 2, 3, ..., \delta\}$. Therefore, the total number of images in a gallery can be given as;

$$G = \{M_1, M_2, M_3, ..., M_{N*T}\}$$

EMOTENET Built on VGGFace



Voting Strategies

In majority voting, each image M casts a vote V_M on the basis of maximum probability P_M for the given image,

$$V_M = argmax P_M(M)$$

In the subsequent step, the votes cast by each image M, of the testing image set X_{test}, are compared. The class with the highest number of votes is declared as the nominated class y_{test} for the image set.

$$MV_{X_{test}} = mode(Y_{test})$$

Where Y_{test} is the predicted class for each test image set X_{test}.

In Weighted Voting, each image M casts a vote for all classes. The vote is then assigned with a weight as per its probability, P_M. Let V_w is the vote for an image, where β is a constant, it can be given as:

$$V_{w(M_T)} = e^{-\beta P_{M_T}}$$

Hence, the predicted weighted vote WV for the testing image set X_{test} can be deduced as;

$$WV_{X_{test}} = \sum_{M=1}^{T} V_{w(M_T)}$$

Ablation Experiments

To validate the generalization abilities and robustness of the proposed technique, the data was exposed to four problems and here's how it went:

- Salt & Pepper Noise
- Gaussian Noise
- Image Resolution
- Image-set Size

S. No.	Amount	MV	WV
1	0.05	95.91%	97.27%
2	0.1	91.15%	89.11%
3	0.15	81.63%	78.91%
4	0.2	69.40%	69.40%

S. No.	Image Size	MV	WV
1	48×48	100%	100%
2	56×56	100%	100%
3	64×64	100%	99.31%
4	72×72	94.55%	94.55%

S. No.	Mean	MV	WV
1	0.5	98.63%	95.91%
2	0.1	98.63%	99.31%
3	0	98.63%	99.31%
4	-0.1	99.31%	99.31%
5	-0.25	95.23%	95.23%
6	-0.5	51.20%	48.97%

S. No.	Training Image Set Size	MV	WV
1	90%	99.31%	99.31%
2	80%	97.95%	99.31%
3	70%	95.23%	96.59%
4	60%	99.31%	99.31%
5	50%	97.27%	96.59%
6	20%	90.47%	93.87%
7	10%	80.95%	78.23%

Average of 3x folds