Accurate Identification of Human Emotional States from Images Using Deep Learning

Emmy Yang and Jake Y. Chen, Senior Member, IEEE

***Abstract*— Facial expression recognition is a crucial aspect of human communication, especially for building social relationships. However, machine-based recognition remains a challenging task where implemented SOTA methods achieved testing accuracies from 50% to 60% on FER-2013. Our research proposes an automatic emotion identification system that utilizes emotional state heatmaps (ES-MAPs) and neural network classification algorithms. Using MediaPipe Face Mesh, our system extracts facial landmark coordinates and calculates the distance between all landmarks. A neutral baseline is subtracted from the landmark distances and saved as a heatmap to train a designed CNN model. The use of ES-MAPs minimizes noise introduced by significant variation in image lighting conditions and variation in head rotation. Our proposed system, ESH-Net, achieved 75% in test accuracy, 15% greater than that of other state-of-the-art methods on the HDFE dataset. In addition, ES-MAPs produced better clustering than the original facial images, indicating significant improvement in the separability and consistency of representation of emotional states. This study demonstrates the potential for emotional state heatmaps and deep learning models to significantly improve the accuracy and efficiency of emotion identification, which can greatly assist in assessing patient's emotional state in medical diagnosis and practice.**

***Clinical Relevance*— This proposed system can assist in objectively patient's emotional state for medical diagnosis and remote symptom tracking for affective disorders.**

I. Introduction

Facial expressions are an essential aspect of human communication and can enhance our understanding of human emotions, enable personalized user experiences, improve mental health assessment, and support emotion-driven decision-making in various domains [23].

There is evidence to support universality in facial expressions through studies of facial expression in different ethnicities and cultures, including preliterate cultures, where commonality was found in the expression and recognition of emotions on the face [30], [11], [1], [31]. For examples of commonalities in expressions, refer to the Github documentations. The "Universal facial expressions" are those representing happiness, sadness, anger, fear, surprise, and disgust. Various approaches have been proposed for machine-based emotion recognition from facial expressions, including rule-based systems, feature-based methods, and machine-learning techniques [17], [28].

Machine learning methods, particularly deep learning models like convolutional neural networks (CNNs), have shown promising results by reducing the dependence on face-physics-based models and other pre-processing techniques by enabling "end-to-end" learning to occur in the pipeline directly from the input images [29]. Traditional machine learning algorithms often struggle with complex and dynamic facial expressions, but recent studies indicate that deep learning algorithms can improve facial emotion recognition performance [12], [13], [24], [7]. Studies on self-modified FER-2013 datasets have reported high accuracy rates ranging between low 60% to low 70% using CNN-based models [24], [5], [24], [13], [7], [6]. Despite the effectiveness of CNN-based models, they still face challenges in accurately detecting emotions from images with significant variations in facial expression and lighting conditions [2]. Additionally, the emotional state recognition challenge demonstrates a wide disparity in reported human accuracy, ranging from the low 60% to high 80% in the literature, highlighting the significant variability in performance [22].

We propose a novel approach, ESH-Net, that uses emotional state heatmaps (ES-MAPs) and convolutional neural networks. First, we calculated pairwise distances between 468 facial landmarks in 3-D space to generate emotional state heatmaps that visually represent the facial landmark distributions, enabling a comprehensive analysis of expressions. This use of landmark distances minimizes noise introduced by significant variation in image lighting conditions. Then a neutral baseline is subtracted to calculate the relative change in facial landmark distances. These measurements are represented in a heatmap form called ES-MAPs. This innovative approach, which integrates emotional state heatmaps and deep learning, has the potential to greatly enhance the reliability and precision of facial emotion recognition systems. Our proposed model is then trained and tested on the Japanese Female Facial Expression dataset (JAFFE), Facial Emotion Recognition (FER-2013), and High Resolution Facial Expression (HRFE) datasets.

II. Methods

A. Compiling emotional state data sets

In this study, we utilized two publicly available datasets for emotion identification: JAFFE [31][32] and FER-2013 [30], both downloaded from Kaggle. The JAFFE dataset comprises 213 black-and-white facial images capturing facial expressions from different individuals. These images are each labeled with one of seven fundamental emotions: anger (N=30), disgust (N=29), fear (N=32), happiness (N=31), sadness (N=31), surprise (N=30), and neutral (N=30). The images were obtained under controlled conditions and are labeled with their corresponding emotions. Image dimensions are of 256 by 256 pixels. The JAFFE dataset's limitations may include low coverage of each emotion as well as the lack of color information. The FER-2013 dataset consists of 29,476 black-and-white facial images sourced from various origins. These images are each labeled with one of seven

emotions: anger (N=3,995), disgust (N=1,203), fear (N=4,097), happiness (N=7,215), sadness (N=4,830), surprise (N=3,171), and neutral (N=4,965). FER-2013 exhibits variations unrelated to expression, such as lighting variation, occlusions, and non-frontal head poses. Image dimensions are that of 48 by 48 pixels. The FER-2013 dataset's limitations may include low resolution of images as well as the lack of color information.

Due to the limitations of JAFFE and FER-2013, we created a high-resolution facial expression dataset, HRFE (High-Resolution Facial Expressions) which includes both high coverage of each emotion state, high image resolution, and RGB color information, by compiling web images. The dataset consists of 1,045 images total. These images are labeled with one of seven emotions: anger (N=150), disgust (N=138), fear (N=151), happiness (N=158), sadness (N=145), surprise (N=143), and neutral (N=160).

B. Representing raw facial images as ES-MAPs

The proposed approach for emotion identification in this study utilizes MediaPipe Face Mesh [33], a 3D face landmark estimation technology. MediaPipe Face Mesh is used to detect for a face and extract the 3D coordinates of the estimated 468 face landmarks. The x- and y-coordinates correspond to the point locations in the 2D plane, and the z-coordinate represents the depth relative to a reference plane passing through the mesh model's center of mass. These landmarks are estimated using a grid of 2D points in feature space and extracting the features under the sampled points in a differentiable manner [33]. We create a transformation matrix using the three-dimensional Euclidean distance between all pairwise combinations of points. This new matrix is normalized using linear normalization and plotted as a heatmap (ES-MAP). The generated ES-MAPs for a subset of FER-2013 are available at DOI: 10.5281/zenodo.8068171. For further details on ES-MAP generation, refer to the Github documentations. (https://github.com/emmyyangqy/ESH-Net).

C. Clustering emotional states

The advantage of the ES-MAP generation will be evaluated with the degree of clustering observed using Uniform Manifold Approximation (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE). These techniques provide insight on the feature representations learned by the emotion identification model and compare the differentiation and separation of different emotional states between ES-MAPs and their original facial images.

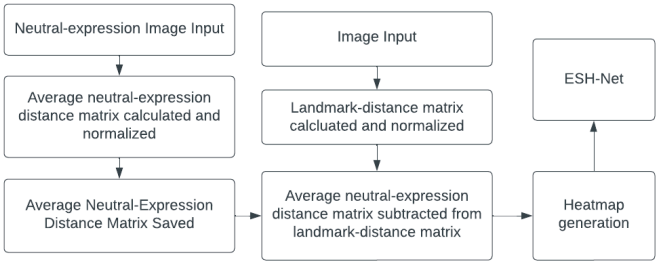

Figure 1. Simplified flowchart representation of data processing, heatmap generation, and ESH-Net model

The consistency of ES-MAPs within the emotion label is evaluated by comparing the heatmap color patterns. An averaged consensus heatmap for each emotion was generated by averaging 50 random heatmaps within the respective emotion. Close-match heatmaps are heatmaps where the heatmap's grid-like pattern and colors are located in similar locations as those of the consensus heatmaps.

D. Building the ESH-Net classification model using CNN

ESH-Net is based on Convolutional Neural Networks (CNNs). ESH-Net is a five-layer CNN designed using Keras with Tensorflow as its backend and was trained using Adam optimizer and categorical cross-entropy loss function.

The architecture was designed with a focus on optimizing the network's ability to learn the relationships between the input images and the categorical classification outcomes. The architecture was selected based on the results of a systematic hyperparameter tuning process involving experimentation with different architectures, activation functions, and optimizers. For further details on the ESH-Net architecture, refer to the Github documentations.

E. Model performance evaluation and validation

To evaluate model performance, we trained ESH-Net, ConvNet, and ResNet using both generated ES-MAPs and facial images (control) as input. We conducted 10 runs, training the model on HDFE, FER-2013, and JAFFE. The training and testing data was randomly split with a 3 to 7 ratio respectively for each dataset. Each run consisted of 100 epochs, and final testing accuracy and loss was recorded at the end of each run.

III. Results

A. ESH-Net has higher performance in accuracy compared to SOTA methods

Our proposed ESH-Net system achieved a testing accuracy of 75% on the HRFE dataset for the seven universal emotions. These accuracy rates surpass the model accuracies achieved by other implemented state-of-the-art methods. Our proposed model is compared to ConvNet [21] and ResNet [6] with various datasets. Table 1 and Table 2 show the comparison results between the state-of-the-art models and our proposed system.

The use of ES-MAP for training is superior to the use of facial images (control) in HRFE and FER-2013 data sets, while inferior for the JAFFE data set. Overall, Our ESH-Net outperforms ConvNET in all cases, expect when trained on FER-2013 facial images, whether we use ES-MAP representations or not, suggesting the overall improvement of our methods.

TABLE I. Testing Accuracy with ES-MAP

Dataset	Emotional State Heatmaps (ES-MAP)		
	ESH-Net (ours)	***ConvNet***	***ResNet50***
HRFE	0.75 ±0.01	0.43 ±0.06	0.68 ±0.02
FER-2013	0.66 ±0.01	0.62 ±0.01	n/a
JAFFE	0.92 ±0.01	0.79 ±0.08	n/a

TABLE II. TESTING ACCURACY WITH CONTROL

Dataset	Facial Images (Control)		
	ESH-Net (ours)	***ConvNet***	***ResNet50***
HRFE	0.55 ±0.02	0.50 ±0.04	0.65 ±0.01
FER-2013	0.56 ±0.01	0.62 ±0.01	n/a
JAFFE	0.91 ±0.01	0.75 ±0.12	n/a

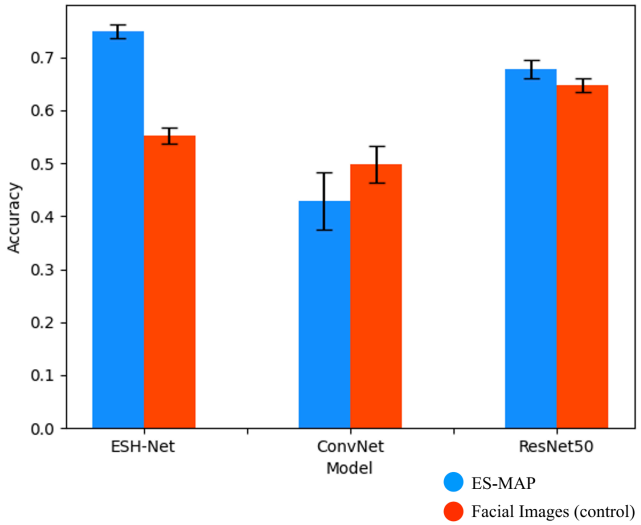

Figure 2. Testing Accuracy of ESH-Net and SOTA models with ES-MAP and Facial Images

B. Improved performance on higher resolution images

Each of these three datasets used for training and testing contains images of various dimensions. The HDFE dataset contains 700 by 700 pixel RGB images. The FER-2013 dataset constrains 48 by 48 pixel grayscale images. The JAFFE dataset contains 256 by 256 pixel grayscale images.

The high-resolution images minimized information loss, thus containing more detailed and fine-grained facial features that might be compromised in lower-resolution images due to down sampling or compression artifacts. The increased level of detail allows the facemesh to capture subtle variations and nuances in facial expressions, which are crucial for the creation of accurate and nuanced facial feature representation by ES-MAPs. This finer level of granularity provided by HRFE allowed for the ES-MAP-trained ESH-Net to significantly improve its performance on emotion recognition tasks compared to the facial-image-trained ESH-Net.

C. Emotional State Heatmap Consistency

We assessed the consistencies between the generated heatmaps and their corresponding averaged heatmap for each specific emotion. The detailed figure, Fig. 3, showcases the consensus heatmap matrix for all six emotional states, accompanied by the closely-matched images and outlier images. We observed that there was overall agreement among the generated heatmaps for a particular emotion. Instances of divergent or inconsistent heatmaps indicates the limitations of the system's ability to overcome noise and high variation in facial expression introduced by the data.

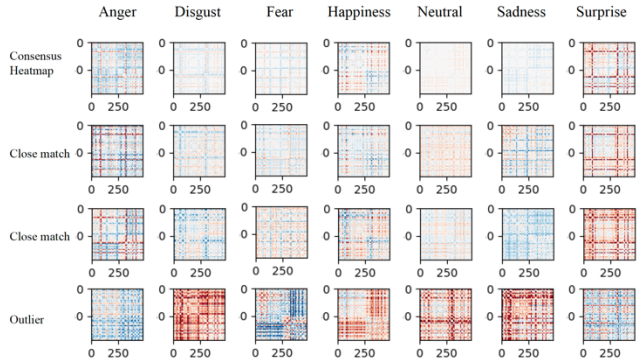

Figure 3. Comparison of averaged consensus emotional state heatmaps with closely matched heatmaps and outliers

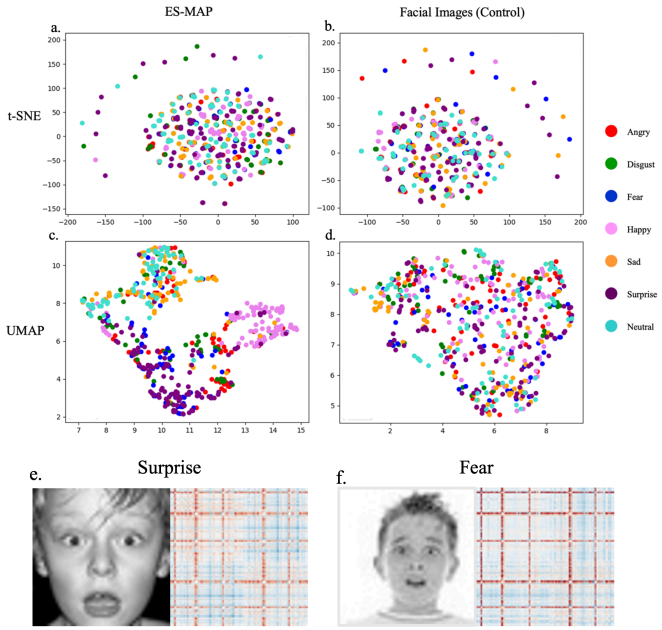

Figure 4. 4a-d: UMAP and t-SNE plots comparing clustering of ES-MAPs and facial images (control). 4e-f: Two labeled images of different emotions having similar facial expressions; FER-2013

C. Emotional state heterogeneity

To assess the advantage of ES-MAPs over original facial images, we conducted clustering analysis using UMAP and t-SNE. Fig. 4 demonstrates that ES-MAPs exhibit improved clustering in UMAP plots but exhibit similar degrees of clustering compared to facial images (control) in tSNE plots. The UMAP plots of ES-MAPs showed increased distinction and separation between emotional states, happiness being the most distinct, indicating that ES-MAPs captured subtle differences in emotional expression more effectively. These findings suggest that the ES-MAP generation approach improves the separability of emotional states within the HDFE dataset. Despite increased distinction of clusters for different emotions, certain clusters (Anger and Fear) appear to overlap in a higher degree. This may indicate that the emotion's expression contains similar facial patterns, and that ES-MAP generation cannot significantly capture the differences between them.

IV. Conclusion

In this paper we have presented, ESH-Net, an approach for emotion identification in static facial images. MediaPipe Facemesh was used for image feature extraction to generate ES-MAPs and a CNN neural network architecture was employed for classification. This approach surpasses existing state-of-the-art solutions and systems in testing accuracy.

Based on our results, we have shown that the use of ES-MAPs with ESH-Net can significantly increases the testing accuracy on facial emotion recognition compared to the use of facial images resulting in better performance. In addition, our analysis indicates that the ES-MAPs generate a more consistent and reliable representation of emotional states compared to the raw facial images. This is evidenced by the significant improvement in clustering observed with Uniform Manifold Approximation (UMAP). It should be noted that many different emotions have very similar facial expressions and can difficult to differentiate in both facial image and ES-MAP formats as seen from Fig. 4e-f.

Our study demonstrates the potential for ESH-Net to significantly improve the accuracy and efficiency of emotion identification. We believe that these findings will have significant implications for the development of more effective and reliable emotion identification systems, with the potential to improve a range of applications in areas such as psychology, marketing, and human-computer interaction.

During the analysis of the image labels in this study, it is important to acknowledge the possibility of incorrect labeling and outlier images. These labeling errors can introduce noise and uncertainty into the dataset, potentially impacting the performance and reliability of the emotion identification system. The FER-2013 dataset used in this study for model training and testing has been modified to clear some of these incorrect labels.

References

[1] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," in Proceedings of the 18th International Conference on Pattern, 2006.

[2] X. Wu, R. He, and Z. Sun, "A Light CNN for Deep Face Representation with Noisy Labels," IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, pp. 823–828, 2017.

[3] Z. Wu, C. Shen, and A. V. D. Hengel, "A light CNN for deep face representation with noisy labels," IEEE Transactions on Information Forensics and Security, vol. 12, no. 10, pp. 2454–2462, 2017.

[4] G. P., P. I., and H. I, Advances in Hybridization of Intelligent Methods: Models, Systems and Applications. Cham, Switzerland: Springer, 2018.

[5] X. Zhang, S. Yin, W. Wang, and X. Liu, "Attention-based deep learning network for facial expression recognition," Neural Computing and Applications, vol. 32, pp. 605–618, 2020.

[6] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," Neurocomputing, vol. 422, pp. 165–175, 2021, doi: 10.1016/j.neucom.2020.10.015.

[7] X. Zhang, L. Yin, J. F. Cohn, and S. Canavan, "Can Facial Landmarks Help Deep Learning for Facial Expression Analysis?," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 469–474.

[8] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in 3rd IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205. doi: 10.1109/AFGR.1998.670949.

[9] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat, no. PR00154), pp. 46–53, 2000.

[10] J. Fiedorowicz, Course of illness and the development of vascular disease in individuals with bipolar disorder. Dissertation. The University of Iowa, 2012.

[11] V. V. Edwards, Cues. Portfolio, 2022.

[12] D. Kollias, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," International Journal of Computer Vision, vol. 127, no. 6–7, pp. 907–929, 2019.

[13] D. Kollias, "Deep affect prediction: Data, challenges, and recent advances," IEEE Signal Processing Magazine, vol. 37, no. 5, pp. 110–125, 2020.

[14] H. K., Z. X., R. S., and S. J, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[15] D. Arumugam, D. S. P. B. E, and M. E, "Emotion Classification Using Facial Expression," International Journal of Advanced Computer Science and Applications (IJACSA, vol. 2, no. 7, 2012.

[16] M. W. Sullivan and M. Lewis, "Emotional Expressions of Young Infants and Children: A Practitioner's Primer." Infants & Young, 2003.

[17] M. Pantic and L. J. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images," IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics, vol. 34, no. 3, pp. 1449–1461, 2004.

[18] N. Scott, "Facial Cues to Mental Health Symptoms. 228." XXXX.

[19] P. Ekman, Facial expression and emotion. American, 1993.

[20] L. K., Z. M., and P. Z, "Facial expression recognition with CNN ensemble," in Proceedings of the 2016 International Conference on Cyberworlds (CW, Chongqing, China, Sep. 2016, pp. 163–166.

[21] T. Debnath, M. Reza, A. Rahman, A. Beheshti, S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," Scientific Reports, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-11173-0.

[22] U. Tariq, M. A. Naeem, H. Abbas, and A. Jalil, "Human and machine performance comparison for facial expression recognition," Cognitive Computation, vol. 13, no. 2, pp. 294–305, 2021.

[23] E. C. Grant, "Human Facial Expression," Man, vol. 4, no. 4, pp. 525–692, 1969, doi: 10.2307/2798193.

[24] P. Liu, X. Han, J. Yuan, and B. Li, "Hybrid deep learning for facial expression recognition," Journal of Visual Communication and Image Representation, vol. 51, pp. 114–123, 2018.

[25] P. R., B. V., and B. V, "Local multi-head channel self-attention for facial expression recognition," Information, vol. 13, no. 419, 2022, doi: 10.3390/info13090419.

[26] Y. Shi, J. Wei, Y. Tang, and Y. Liu, "Machine learning in depression diagnosis: A review," Journal of Affective Disorders, vol. 274, pp. 193–205, 2021.

[27] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing Action Units for Facial Expression Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 97–115, 2001.

[28] P. Ekman, "Universals and cultural differences in the judgments of facial expressions of emotion," Journal of Personality and Social Psychology, vol. 53, pp. 712–717, 1987, doi: 10.1037/0022-3514.53.4.712.

[29] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial expression intensity estimation," *Image Vis. Comput*, vol. 259, pp. 143–154, 2017.

[30] I. Goodfellow, D. Erhan, P. Carrier, A. Courville, M. Mirza, and B. Hamner, "Challenges in Representation Learning: A report on three machine learning contests." 2013. [Online]. Available: https://arxiv.org/abs/1307.0414v1

[31] M. Lyons, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets (IVC." 2020. [Online]. Available: https://arxiv.org/abs/2009.05938

[32] M. Lyons, "Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset." 2021. [Online]. Available: https://arxiv.org/abs/2107.13998

[33] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann, "Attention Mesh: High-fidelity Face Mesh Prediction in Real-time." 2020. [Online]. Available: https://arxiv.org/abs/2006.10962