A Fast Multiple Cue Fusing Approach for Human Emotion Recognition

Willams de Lima Costa
Voxar Labs, Centro de Informática
Universidade Federal de Pernambuco
Av. Jorn. Aníbal Fernandes, Recife, Brazil
wlc2@cin.ufpe.br

Estefania Talavera Martínez
University of Twente
Drienerlolaan 5, 7522 NB Enschede, Netherlands
e.talaveramartinez@utwente.nl

Lucas Silva Figueiredo
Unidade Acadêmica de Belo Jardim
Universidade Federal Rural de Pernambuco
PE-166, 100, Belo Jardim, Brazil
lsf@cin.ufpe.br

Veronica Teichrieb
Voxar Labs, Centro de Informática
Universidade Federal de Pernambuco
Av. Jorn. Aníbal Fernandes, Recife, Brazil
vt@cin.ufpe.br

Abstract

Classifying perceived emotion from visual cues is crucial for developing intelligent systems to understand user behaviour and respond appropriately. This work proposes a deep learning framework for emotion recognition that combines multiple visual cues to accurately identify emotions in images and videos. Our approach achieves competitive results on the well-known CAER-S dataset, with an accuracy of 89.76%, while being more than nine times faster than the other best-performing approach. By extracting and fusing representations extracted from different visual cues, our model provides a more nuanced understanding of emotion, which can help to reduce ambiguity in predictions. This allows for in-the-wild applications where occlusion by the environment may affect the reception of emotional signals.

1. Introduction

Emotions are a foundation for developing social links and for being understood by others. According to researchers, humans can decode emotional signals intuitively, an ability that starts to develop at four years of age [2] and is acquired through interaction with peers. As emotional skills are a key resource for human interaction, we can also argue that interaction between humans and machines could also leverage the understanding of emotion. Therefore, applications should not discard the emotional aspects of the user but rather focus on understanding them [18].

Recognizing human behaviour enables a deeper understanding of users in the most diverse scenarios, and the recognition of perceived emotion is one of these many aspects. Especially in the context of smart environments, which can also be extended by smart cities, understanding their user is crucial. Given its focus on building attractive and lively spaces, physical sensors and Internet-of-Things frameworks are currently deployed to measure various aspects of a region, such as weather, pollution, and traffic. However, they still lack the comprehension of the citizen as an individual and key aspects such as how they feel near specific areas in the city and how urban interventions could impact their perception of those spaces [6]. For instance, sudden changes in the emotional state, going from *Happy* to *Fear*, could indicate criminal or violent activity in a public space.

However, many of these emotional states are not communicated explicitly through speech. Instead, they come from nonverbal communication, encompassing facial expressions, gestures, body language, speech tonality, and other forms of expressions [12, 17, 20]. These cues are often presented simultaneously, and humans rely on multiple nonverbal cues to interpret emotion. Consequently, to accurately recognize perceived emotion, systems must decode multiple nonverbal cues to some extent. Additionally, taking into account the situational context in which the person is placed can help models make more precise assessments since it can directly influence a person's emotional state. For example, a person may experience happiness in social gatherings or when in contact with nature while experiencing sadness in hospitals.

In this work, we propose the usage of multiple nonverbal behavioural cues to recognize perceived emotion, namely, facial expressions, context, and body language. Our goal is to recognize perceived emotion in unconstrained scenarios, allowing for different human-computer interaction applica-

tions in multiple contexts, especially in smart environments. Our approach is called Emotion Recognition on Adaptive Multi-cues (EmotionRAM), and comprises extracting features related to nonverbal cues extracted automatically to predict recognized emotions, as we show in Figure 1.

2. Emotion recognition in computer vision

Facial expression is rich in descriptive features that correlate with emotion, leading researchers to focus on extracting these features for Facial Expression Recognition (FER) tasks. EmotiW [3] competitions have promoted less restrictive approaches that can be deployed in the wild. However, such techniques are still limited by scenarios where the face is occluded or not aligned with the camera [15, 23–25]. To overcome this limitation, researchers such as Lee *et al.* [14], and Le *et al.* have explored context to extend FER. However, by design, these architectures place body language inside context by allowing a single encoding stream to decode background information and body language, leading to poor results related to body language. Techniques for body expression were also proposed in the past; however, focusing on perception extracted from movement, such as gait [1, 19].

Our approach differs from the state-of-the-art by allowing a more representative extraction of cues without imposing constraints on learning the correlations between them and emotions. Furthermore, we directly expand the work by Lee *et al.* [14] and Le *et al.* [13] by adding a new nonverbal cue for emotion recognition related to body language and extending the context encoding stream with the self-calibrated convolution module, building towards in-the-wild applications and reproducibility.

3. EmotionRAM for Emotion Recognition

Given an image I, we aim to infer an emotion y among a set of K emotion labels by using a convolutional neural network model. The proposed network architecture extracts features of three streams: *face encoding stream*, *context encoding stream* and *body encoding stream*. By combining these features in an *adaptive fusion network*, the proposed method can infer emotion from multiple non-verbal cues. In Figure 1, we present the proposed architecture, and each module is detailed below.

3.1. Multi-cue encoding streams

3.1.1 Face encoding stream

Given that in some images we have multiple faces present, we implement a *face selector algorithm* that selects the leading performer's face based on their placement on the scene. Given a set of detected faces $F = \{F_1, ..., F_n\}$, we rank each candidate F_c and select a face for input F_s based on its bounding box area, which points to if the person is on background or on foreground, selecting those on the foreground, and the 2D position (x, y) of the bounding box centroid on the scene.

We crop the bounding-box region of F_s and use it as input to the *face encoding stream*, as shown in Figure 1. This module consists of five convolutional layers with 3x3 kernels with sizes 32, 64, 128, 256, and 256, followed by batch normalization (BN) and rectified linear unit (ReLU) activation, and four max-pooling layers with a kernel size of 2. We spatially average the final feature layer using an average-pooling layer.

3.1.2 Context encoding stream

Contextual emotional information extraction is challenging due to the variability of contextual information and the potential occlusion of critical details. To address this, enhancing the encoding stream for context could improve emotion classification accuracy by enabling more representative feature extraction. This module consists of four convolutional layers with 3×3 kernels with sizes 32, 64, 128, and 256, followed by batch normalization layers, ReLU activations, and four max-pooling layers with a kernel size of 2×2. Finally, we add an adaptive self-calibrated convolution [16] with kernel size 3×3 and a ReLU activation layer.

Adaptive self-calibrated convolutions. We propose the usage of self-calibrated convolutions [16] to allow output features to be enriched. The adaptive self-calibrated convolution module receives an input with channels size C and outputs a features map with channels size C'; a restriction when using the original self-calibrated convolutions we overcame.

Attention inference module. An unsupervised attention inference module allows the *context encoding stream* to focus on salient contexts. This involves using output features as input, applying a Softmax operation to reduce channels from 256 to 1, and applying attention to the output feature to boost important feature participation. The module addresses the issue of context containing excessive background information, allowing selection of the most important features.

3.1.3 Body encoding stream

We employ a *body encoding stream* to investigate body pose as a nonverbal communication input by extending the Simple Baselines pipeline [26]. We remove the last deconvolutional layer to extract features before the classification of joints. Our intuition with this proposal is to allow the following layers to learn the correlation between features and emotions instead of leveraging the 2D annotations. We also use Mask R-CNN to create segmentation masks of the people present in the scene and remove people present in the background. We select a mask from the scene based on the overlap between the face selected for input F_s.

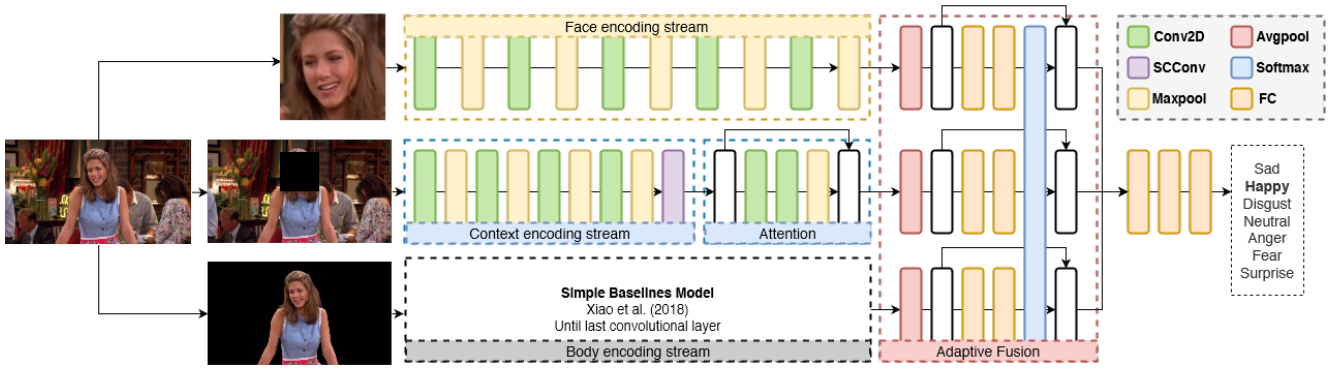


Figure 1. Our proposed architecture for multi-cue emotion recognition. Given an input image in an unconstrained scenario, we use an off-the-shelf face detector algorithm [10] to get the localization of the face on the image. First, we crop the face and use it as input for the *face encoding stream*, responsible for extracting features from the face. Next, we fill the cropped region with a black rectangle and use this new image as input for the *context encoding stream*. Since the facial crop is occluded, this stream is "forced" to search for features from other image regions during training (i.e., the background context). Finally, we apply a segmentation technique [7] to remove background noise and persons that are not acting directly on the scene. We use this segmentation mask as input for an off-the-shelf human keypoint extractor [26]. The features extracted from these three streams are fused in an adaptive way allowing the emotion classification.

3.2. Adaptive fusion networks

The direct concatenation of multiple nonverbal cues fails to provide adequate performance. We propose the fusion of features from each encoding stream, X_F, X_C, and X_B using a fusion network that adaptively chooses the weight for each cue. By learning a set of attention weights $\lambda = \{\lambda_F, \lambda_C, \lambda_B\}$, we weigh the contribution of each stream as $X_A = \Pi(X_F \odot \lambda_F, X_C \odot \lambda_C, X_B \odot \lambda_B)$, given Π as the concatenation operator and $\odot$ the Hadamard product.

4. Experiments

We perform our experiments on the well-known CAER-S dataset [14] to allow direct comparison with the state-of-the-art. This dataset is based on video clips from TV shows and each frame is categorized as one of the seven basic emotional states [4], namely: *Angry*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise* and *Neutral*. We compare our proposed multi-cue learning framework against baseline works [9,11,22] and state-of-the-art approaches on emotion recognition [5, 13, 14, 27, 28]. Moreover, we assess the contribution of each proposed module using an ablation study on the considered cues and discuss the contribution of our approaches to deal with the limitations of the CAER-S [14] dataset, such as the face selector algorithm.

5. Results and discussion

We compare our results with different approaches in Table 1. The proposed method was 16.25% better when comparing our implementation of the baseline approach - CAER-Net-S [14] that obtained an accuracy of 73.51%. We also performed consistently better against traditional deep neural networks approaches for image tasks, such as AlexNet [11] (47.36%), VGG-Net [22] (49.89%), and ResNet [9] (57.33%). We also compared to [13], by using their available training and evaluation code on their GitHub page, changing their data loader method to point to our references of the CAER-S dataset. However, in this reproduction, we could not achieve their reported performance (89.88%) and obtained an accuracy of 83.08%. Our method performs slightly worse than their reported result by a difference of 0.12%, but 6.68% better than what we obtained when reproducing their framework. Even though the difference in performance might not seem significant, the fact

Methods	**Acc. (%)**
ImageNet-AlexNet [11]	47.36
ImageNet-VGG-Net [22]	49.89
ImageNet-ResNet [8]	57.33
Fine-tuned AlexNet [11]	61.73
Fine-tuned VGGNet [22]	64.85
Fine-tuned ResNet [8]	68.46
GRERN [5]	81.31
EfficientFace [28]	81.48
CAER-Net-S [14]	73.51
CAER-Net-S* [14]	81.50
MA-Net [27]	88.42
GLAMOR-Net [13]	77.90
GLAMOR-Net* [13]	76.33
GLAMOR-Net (ResNet-18) [13]	**89.88**
GLAMOR-Net (ResNet-18)* [13]	83.08
EmotionRAM (face+context)	88.10
EmotionRAM (face+context+body)	**89.76**

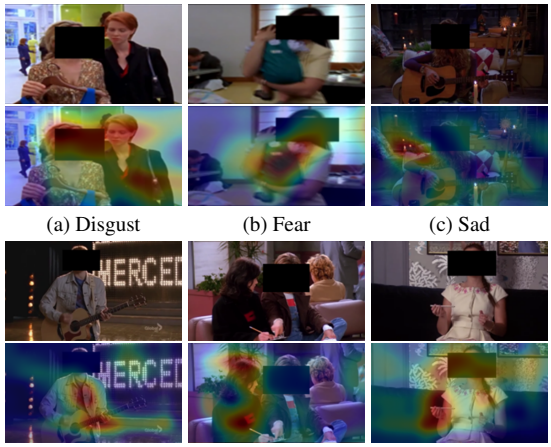
* our reproduction

Table 1. Quantitative evaluation of EmotionRAM in comparison with baseline methods on the CAER-S dataset.

Method	Min.	Avg.	Acc. (%)
EmotionRAM	**6.1490ms (≈162 fps)**	**7.0110ms (≈142 fps)**	89.76
GLAMOR [13]*	20.6775ms (≈48 fps)	28.7653ms (≈43 fps)	77.90
GLAMOR (ResNet-18) [13]*	59.9367ms (≈17fps)	71.9979ms (≈15 fps)	**89.88**

* our reproduction

Table 2. Inference time of our EmotiRAM framework compared against the state-of-the-art models.



(a) Disgust (b) Fear (c) Sad

(d) Happy (e) Neutral (f) Surprise

Figure 2. Visualization of features learned from context. Only six emotions are shown for brevity. On the top row, we see the image used as input for the module, and on the bottom row, the output from Grad-CAM [21] with respect to the last convolutional layer of the attention inference module. The areas with the highest activations are the areas with higher significance in relation to the prediction.

that we rely on an additional cue for describing the scene might allow for a better understanding of the situation. We should also consider that adding a third encoding stream may be helpful in real situations that are not present in the dataset where the face might be occluded or the context is insignificant.

Focusing on the application scenario, we compare the computational cost of the top-performing models by measuring the minimum and average inference time on a deep learning environment powered by an NVIDIA GeForce RTX 2080 Ti with 12GB of RAM. The results of this evaluation are shown in Table 2. On a consumer-grade notebook with an NVIDIA GeForce GTX 1060 with 6GB of VRAM, our model requires an inference time of 13.9586ms (≈ 72 fps) on minimum and 16.6082ms (≈ 60 fps) on average. This experiment shows that our model can infer faster than the architectures proposed in [13], while keeping a competitive accuracy score. Furthermore, given its lower computational cost, our model is more suitable for deployment on consumer-grade computers or energy-friendly edge devices.

We also applied a visual investigation based on Grad-CAM [21] in which we investigate how the context encoding stream acts towards the prediction of emotion, as we show in Figure 2. In Figure 2a, we can see that the *context encoding stream* learned to take into consideration interactions with other people by focusing on the second performer on the scene and started leveraging their emotions, since emotions in a group tend to point towards the same direction, something that is deeply motivated by the field of behavioural psychology. The same idea is reinforced in Figure 2b, in which the *context encoding stream* focuses on the way that a person is holding an infant - by placing them close to their body in a protective manner as if they were in a dangerous situation.

Figure 2c and Figure 2d may be directly compared since this is a classic question regarding context. In Figure 2c, we have a prediction for *Sad*, and the *context encoding stream* focuses on the background of the scene, such as the presence of candles and how the illumination leans towards a darker theme, while in Figure 2d we have a more uplifting scene, with a more relaxed pose on a well-lit stage. In Figure 2e, we have another group interaction in which the *context encoding stream* also considered how the other person was feeling, pointing out a simple conversation with neutral feelings. By investigating the scene, we may notice that they are placed in an uninteresting situation of filing a document, and the network can correctly predict the emotion for this interaction. Finally, in Figure 2f, we see a case in which person-object interaction is taken into consideration, in which the *context encoding stream* focuses on the object on the performer's hand (a pregnancy test) and on the way that they are holding it.

6. Conclusion

In this work, we present an approach for emotion recognition using multiple nonverbal cues as input. Our proposed model learns to weigh each input adaptively by attributing higher weights to more descriptive ones in each scenario. Our approach is comparable to the state-of-the-art in the accuracy metric of the CAER-S dataset, with a 0.12% difference from the highest-rated method. Our model also requires lower computational cost than the current state-of-the-art, being more than three times faster on average, allowing for deployment on energy-friendly edge devices in the application scenario (e.g., department stores or public parks). Finally, our model differs from the state-of-the-art by leveraging more nonverbal cues, showing promising results by tackling real-life challenges such as ambiguity and occlusion.

For future works, we plan to build a new dataset that tackles the limitations of CAER-S, especially the lack of variability regarding ethnicity and culture. We also plan on investigating new approaches for improving context leveraging.

References

[1] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. *AAAI Conference on Artificial Intelligence*, pages 1342–1350, 2020. 2

[2] R Thomas Boone and Joseph G Cunningham. Children's decoding of emotion in expressive body movement: the development of cue attunement. *Developmental psychology*, 34(5):1007, 1998. 1

[3] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. *ACM International Conference on Multimodal Interaction*, pages 427–432, 2016. 2

[4] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. 3

[5] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. Graph reasoning-based emotion recognition network. *IEEE Access*, pages 6488–6497, 2021. 3

[6] Benjamin Guthier, Rajwa Alharthi, Rana Abaalkhail, and Abdulmotaleb El Saddik. Detection and visualization of emotions in an affect-aware city. In *Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities*, pages 23–28, 2014. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016. 3

[10] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 3

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 3

[12] Soonil Kwon et al. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167:114177, 2021. 1

[13] Nhat Le, Khanh Nguyen, Anh Nguyen, and Bac Le. Global-local attention for emotion recognition. *Neural Computing and Applications*, pages 1–15, 2021. 2, 3, 4

[14] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. *IEEE International Conference on Computer Vision*, pages 10143–10152, 2019. 2, 3

[15] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, pages 2439–2450, 2018. 2

[16] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10096–10105, 2020. 2

[17] Dipika S Patel. Body language: An effective communication tool. *IUP Journal of English Studies*, page 2, 2014. 1

[18] Rosalind W Picard. *Affective Computing*. MIT Press, 2000. 1

[19] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv:1906.11884*, page 1, 2019. 2

[20] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2019. 1

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 4

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 3

[23] Wen Su, Haifeng Zhang, Yuan Su, and Jun Yu. Facial expression recognition with confidence guided refined horizontal pyramid network. *IEEE Access*, 9:50321–50331, 2021. 2

[24] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 2

[25] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2

[26] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *European Conference on Computer Vision*, pages 466–481, 2018. 2, 3

[27] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, pages 6544–6556, 2021. 3

[28] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. *AAAI Conference on Artificial Intelligence*, pages 3510–3519, 2021. 3