Context-Aware Facial Expression Recognition Using Deep Convolutional Neural Network Architecture

Abha Jain[1], Swati Nigam[1,2(✉)], and Rajiv Singh[1,2]

[1] Department of Computer Science, Banasthali Vidyapith, Radha Kishnpura, Rajasthan 304022, India
swatinigam.au@gmail.com

[2] Centre for Artificial Intelligence, Banasthali Vidyapith, Radha Kishnpura, Rajasthan 304022, India

Abstract. A frame of reference, which includes additional contextual information, can provide a more accurate and comprehensive understanding of the individual's emotional state. This context might encompass factors such as the person's surroundings, body language, gestures, tone of voice, and the specific situation or events taking place. Previous research in this field has often struggled to recognize emotions within a contextual framework. However, by considering contextual elements in addition to facial expressions, we can gain a more nuanced and precise picture of the individual's emotions. In this paper, we used both context-aware datasets (Emotic, CAER, and CAER-S) and only the facial emotion datasets (Affectnet and AEFW) to signify the context. In this Emotic dataset images are labeled with 26 emotional categories. We utilized these datasets to build a convolutional neural network model that effectively examines both the individual and the overall scenario to accurately identify a wide range of information pertaining to emotional states. The features obtained from these two modules are combined using a specialized fusion network. Through this approach, we demonstrate the significance of emotion recognition within a visual context.

Keywords: Context-aware · convolution neural network · emotion

1 Introduction

Context-aware facial emotion recognition technology is the key to unlocking a new era of emotional intelligence. Context-aware facial emotion recognition refers to the ability of a system or technology to accurately detect and interpret facial expressions in real-time, taking into consideration the surrounding context in which the expressions occur. This includes factors such as the individual's environment, social cues, and other relevant contextual information.

Previous research in the subject of computer vision has primarily focused on the examination of facial expressions, often involving the categorization of these expressions into the six (or seven) basic emotions [1–3]. By incorporating context awareness,

B. J. Choi et al. (Eds.): IHCI 2023, LNCS 14531, pp. 127–139, 2024.
https://doi.org/10.1007/978-3-031-53827-8_13

facial emotion recognition systems become more robust and reliable in accurately identifying and understanding emotions displayed on a person's face. Contextual factors, such as social settings, cultural norms, and individual experiences, can influence emotions, making it particularly important to incorporate context awareness into facial emotion recognition systems. [4, 6].

In many instances, when we broaden our perspective beyond an individual and consider the surrounding environment, we can discern additional emotional nuances that would otherwise remain hidden without context. For example, we can observe from the scenario depicted in Figure 1(a) that this individual is experiencing feelings of worry and pressure. But if you consider the contextual boundaries in Figure 1(b), it appears that he is ready to launch an attack on his opponent in a game and is prepared to counter any offensive moves made by the opponent. Moreover, we can infer that his overall emotional state is alarmed, as he appears confident in the actions he is about to undertake. So he is in disquiet about the situation.

Traditional facial emotion recognition systems primarily focus on analyzing facial features and patterns to determine emotions. The way emotions are expressed varies, including through facial expressions [3, 4], speech [6], and body language [7]. However, these systems often overlook the influence of context, which can significantly impact the interpretation and understanding of emotions. For example, a smile at a social gathering may indicate happiness, while the same smile at a business meeting may indicate politeness or agreement rather than genuine joy. Indeed, when taking the context into account, it becomes possible to make reasonable conjectures regarding emotional states even in cases where the person's face is not visible.

In this paper, we address the problem of recognizing emotion states in context. We used two popular datasets, Emotic [4] and CAER [5]. The EMOTIC and CAER (Context-Aware Emotion Recognition Networks) databases comprise images featuring people within their respective contexts, each annotated to reflect the emotional states that an observer can deduct from the overall situation. We structured the networks using a two-stream architecture, which consists of two feature encoding streams: one for facial encoding and the other for context encoding. Our primary concept revolves around the search for pertinent contexts, a factor that aids the model in mitigating ambiguity and enhancing accuracy in emotion recognition. Our study focused on evaluating the efficacy of a convolutional neural network (CNN) model in accurately identifying emotions within a contextual framework.

This research presents a technique that utilizes contextual information along with facial expression to demonstrate the practicality of accurately recognizing the suitable emotion within a given environment. In order to achieve this objective, we have established the concept that a model's emotions and context convey connections and limitations among various elements. This study represents the first known instance of employing deep learning to comprehensively investigate the integration of contextual information and facial information in order to achieve emotion recognition.

Section 2 provides an overview of the proposed context-aware emotion recognition system. Section 3 demonstrates the methodology of integrating contextual information with face expression identification. Section 4 showcases the findings of the experiments conducted. Section 5 provides the concluding remarks of the study.

(a) (b)

Fig. 1. Facial Expression and (b) Facial Expression with contextual information

2 Related Work

A comprehensive literature survey on context-aware facial emotion recognition (FER) reveals a significant body of research and advancements in this area. The following is an overview of some key studies and contributions in the field:

Li et al. (2019) [8] introduced a dynamic attention-based convolutional neural network that effectively captures both local and global context information for the purpose of facial emotion recognition. The model dynamically attends to different facial regions based on their relevance to the emotional context, improving the accuracy of emotion recognition.

Zhang et al. (2020) [9] concentrated on integrating many modalities, including facial expressions, speech, and body motions, in order to enhance context-aware FER. The study develops a deep learning-based framework that effectively combines these modalities to enhance emotion recognition accuracy.

Li et al. (2017) [10] introduced a sophisticated adaptive attention network for accurately identifying face emotions in real-world situations. The model dynamically adjusts attention to different facial regions based on their discriminative power, taking into account contextual information to improve emotion recognition accuracy.

Caon et al. (2013) [11] provided a comprehensive overview of context-aware affective computing, including context-aware FER. It explores different contextual factors, such as social context, environmental context, and temporal context, and their influence on emotion recognition. The study also discusses various approaches and challenges in context-aware affective computing.

Zhao et al. (2019) [12] proposed a context-aware FER framework based on deep neural networks. The study considers both facial expressions and contextual information, such as scene context and temporal dynamics, to improve emotion recognition accuracy.

The model effectively integrates contextual information with facial features for enhanced performance.

These studies highlight the importance of considering contextual factors in facial emotion recognition to improve accuracy and understand emotions in a more comprehensive manner. They demonstrate the effectiveness of various techniques, such as attention mechanisms, multi-modal fusion, and deep learning approaches, in achieving context-aware FER.

The area of context-aware facial expression recognition (FER) is continuously progressing, with ongoing research and improvements. This literature study offers a brief overview of the current corpus of research and establishes a basis for future investigation and advancement in this captivating academic field.

3 Proposed Method

Within this section, we introduce a simple yet powerful structure for the detection of emotions in photos and movies that takes into account the surrounding environment. This paradigm utilizes both facial expressions and environmental information in a complementary and cooperative manner to improve recognition accuracy.

A straightforward approach involves utilizing the holistic visual features, as demonstrated in prior work [13, 14, 28]. However, such a model may not effectively capture important contextual regions. Recognizing that emotions can be better understood by considering both the contextual elements of a scene and facial expressions [15, 16], we introduce an attention inference module designed to estimate contextual information in both images and videos. By temporarily concealing facial regions in the input data and focusing on attention regions, our networks are capable of identifying more discriminative contextual regions. This, in turn, enhances the accuracy of emotion recognition in a context-aware manner.

To establish the proposed set of emotional categories as outlined in Table 1, we conducted a comprehensive collection of affective state vocabulary and concluded 26 groups of words to represent the exact human emotion state [4].

To simplify, let consider an image denoted as "I" and a video $V = \{I_1, \ldots, I_T\}$ comprised of a sequence of "T" images. Our primary objective is to determine the emotion label "y" from a set of "K" emotion labels, $\{y1, \ldots, yK\}$, assigned to either the image "I" or the video clip "V" using deep Convolutional Neural Networks (CNNs). To address this challenge, we introduce a network architecture composed of two distinct sub-networks: a two-stream encoding network and an adaptive fusion network, as depicted in Figure 2. These two-stream encoding networks encompass a face stream and a context stream, each responsible for encoding facial expressions and context information separately. By merging these two sets of features within the adaptive fusion network, our approach achieves optimal performance in the context-aware recognition of emotions.

3.1 Model Architectures

We present a comprehensive model, as illustrated in Figure 6, that can simultaneously predict both emotion and contextual characteristics. Our networks incorporate a facial

expression encoding module, which is comparable to existing approaches used for determining facial expressions [9, 10, 17]. In order to create the input for the face stream, we first detect and separate the facial areas using easily accessible face detectors [10]. Moreover, supplementary feature extraction modules have been created as a condensed iteration of the low-rank filter convolutional neural network first shown in [5]. The main benefit of this network lies in its ability to offer great precision while simultaneously reducing the number of parameters and computational complexity. The initial network comprises 16 convolutional layers with 1-dimensional kernels, which effectively simulate 8 layers by employing 2-dimensional kernels. Afterwards, a fully connected layer is added, creating a direct link to the SoftMax layer. In our revised version, we remove the fully connected layer and instead transmit the features obtained from the activation map of the final convolutional layer. The selection is predicated upon the objective of preserving the crucial geographical data necessary for the work.

The attributes obtained from these two modules are then merged using a specialised fusion network. The fusion module initiates the process by implementing a global average pooling layer on each feature map, thereby reducing the dimensionality of the data greatly. Subsequently, a primary fully connected layer functions as a dimensionality reduction layer for the pooled features, yielding a vector with 256 dimensions. Subsequently, a bigger fully linked layer is added, allowing the training process to acquire distinct representations for each task, in accordance with the concepts described in [5]. This layer is utilized for the identification of emotion categories, encompassing a total of 26 distinct emotional states. Each convolutional layer is thereafter followed by Batch normalization and rectifier linear activation.

The three modules' parameters are simultaneously learned using stochastic gradient descent with momentum. The batch size has been adjusted to 52, which is twice the number of unique categories in the dataset. Our method employs uniform sampling per category to ensure that each discrete category is represented by at least one instance in every batch. Based on empirical evidence, we have determined that this strategy produces better outcomes in comparison to randomly rearranging the training set.

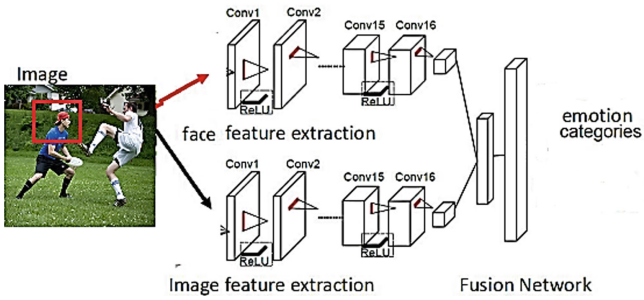


Fig. 2. Propose Model Architecture for Context aware Facial Emotion Recognition

The overall loss function used for model training is defined as a weighted combination of two distinct losses: L_comb = λ_disc * L_disc + λ_cont * L_cont. Here, λ_disc, cont represents the weight that determines the importance of each loss component,

while L_disc and L_cont denote the losses associated with the tasks of learning discrete categories and learning continuous dimensions, respectively.

We approach this multiclass-multilabel problem by framing it as a regression task. To address the class imbalance inherent in the dataset, we employ a weighted Euclidean loss function. Through empirical analysis, we have determined that this particular loss function outperforms alternatives such as Kullback-Leibler divergence or a multi-class multi-classification hinge loss. To be precise, the loss is defined as follows:

$$L_{disc} = \frac{1}{N} \sum_{i=1}^{N} w_i \left(\hat{y}_i^{disc} - y_i^{disc} \right)^2 \tag{1}$$

where N represent the number of categories (N = 26 as per case), yˆ disc i is the caculated estimated result for the i-th category and y_i^{disc} is the original-truth label. The parameter w_i is the weight assigned to each category. Weight values are defined as $w_i = 1/(\ln(c+pi))$, where pi is the probability of the i-th category and c is a parameter to control the range of valid values for w_i. Using this weighting scheme, the values of w_i are bounded as the number of instances of a category approach to 0. This is particularly relevant in our case as we set the weights based on the occurrence of each category in every batch.

It is essential to merge the derived characteristics from two modules in order to effectively identify the emotion by utilizing both facial and contextual information simultaneously. The feature extraction modules are initiated by utilizing pre-trained models from two distinct extensive classification datasets, specifically ImageNet [18] and Places [19]. ImageNet contains a diverse collection of photos that represent common items, including people. This makes it a helpful tool for understanding the visual content of the image area that includes the person of interest. Conversely, Places is a deliberately designed dataset for advanced visual comprehension tasks, specifically for recognizing scene categories. Therefore, by pretraining the image feature extraction model using the Places dataset, it guarantees the inclusion of global (high-level) contextual information.

4 Experiments and Discussion

In this section, we discuss the two benchmark datasets and their effectiveness in the proposed context-aware Facial Expression Recognition (FER) system [20]. Initially, we provide an overview of the benchmark datasets rather than the details of the experimental setup. Subsequently, we compare the performance of the other model on these benchmark datasets with their approach and their efficiency and effectiveness on the same dataset.

4.1 Benchmark Datasets: Emotic and CAER

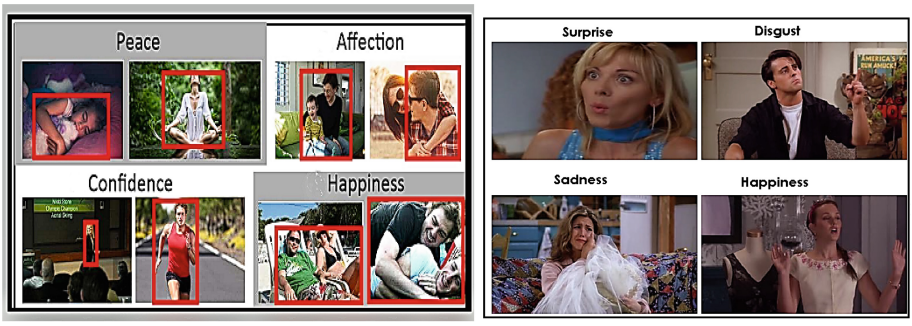
The EMOTIC database [4] consists of images sourced from MSCOCO [20], Ade20k [21], and the Google search engine. The collection consists of 18,316 pictures, each containing 23,788 individuals with annotations. Figure 1 exhibits instances of images contained in the database, accompanied by their corresponding comments. The “EMOTIC” framework has 26 distinct emotional categories, which cover a wide range of emotional states. The categories are elaborated and delineated in Table 1.

The table's definitive list of categories includes the six fundamental emotions (category 7, 10, 18, 20, 22, and 23) [22]. Category 18, designated as "Aversion," functions as a more comprehensive category that includes the basic feeling of disgust.

CAER is a compilation of extensive video snippets extracted from television programmes, which are then annotated to facilitate the recognition of emotions in a context-aware manner. Every video clip underwent manual annotation, categorizing them into six distinct emotions: "anger", "disgust", "fear", "happy", "sad", and "surprise", in addition to a category labelled as "neutral". The collection comprises 3,201 video segments, totaling around 1.1 million frames.

Furthermore, Lee and Kim [5] have derived approximately 70,000 static images from CAER, resulting in the formation of a static image subset referred to as CAER-S. Figure 1(b)illustrate the images from CAER-S. This dataset considers only images with one emotional label and ignores images with more than two annotations.

Table 2 conducts a comparison and gives a description of context-aware datasets CAER CAER [5] and Emotic [4] datasets and several other widely used datasets, including CAER-S [5], Affect-Net [23], AFEW [24], and Video Emotion datasets [25] (Fig. 3).



(a) (b)

Fig. 3. (a) Sample Image from (a) Emotic and (b) CAER-S dataset

Table 1. Emotion Categories as per EMOTIC Dataset

Sr. No.	Emotion Labels	Feelings
1	Peace	a state of well-being and calmness by the absence of worry, the presence of positive thoughts or sensations, and a feeling of satisfaction
2	Affection	"Fond feelings" encompass with love, care, and affection
3	Esteem	refers to the positive opinion and a sense of high regard, respect, or admiration one holds for someone
4	Anticipation	involves a sense of expectation and hope for something positive or significant to occur
5	Engagement	Act of showing the genuine interest or focusing one's attention and energy on the task at hand, often with enthusiasm and dedication
6	Confidence	It involves having a robust belief in one's capacity to tackle challenges, make informed decisions, and accomplish goals
7	Happiness	feelings of joy, contentment, and satisfaction
8	Pleasure	a state of enjoyment, delight, or satisfaction that arises from experiencing something pleasurable or enjoyable
9	Excitement	state of enthusiasm, eagerness, or a heightened sense of anticipation
10	Surprise	astonishment or disbelief in response to unexpected or startling events
11	Sympathy	someone who is experiencing pain, suffering, or hardship. It involves a sense of care
12	Doubt/Confusion	states marked by uncertainty or lack of clarity
13	Disconnection	a feeling of detachment or isolation from others or from one's surroundings
14	Fatigue	physical and emotional state characterized by extreme tiredness, weakness
15	Embracement	state characterized by open acceptance, warmth, and a willingness to embrace someone or something
16	Yearning	a deep and intense desire or longing for something
17	Disapproval	indicating a negative judgment or lack of acceptance
18	Aversion	is a strong feeling of dislike or avoidance
19	Annoyance	is a mild form of irritation or frustration caused by something that disrupts or disturbs one's peace
20	Anger	feelings of displeasure, hostility, and a desire to react to a perceived injustice, frustration, or provocation

(*continued*)

Table 1. (*continued*)

Sr. No.	Emotion Labels	Feelings
21	Sensitivity	to the capacity to perceive and react to stimuli, emotions, or external influences with awareness
22	Sadness	marked by feelings of sorrow, unhappiness, and a sense of loss or disappointment
23	Fear	triggered by the perception of a threat, danger, or harm
24	Pain	sensation characterized by discomfort, distress, or suffering, often resulting from injury, illness, or emotional distress
25	Suffering	Is state of experiencing physical or emotional pain, distress, or hardship
26	Disquietment	state marked by restlessness, unease, or a lack of tranquility

Table 2. Description of the different datasets

Dataset name	Image/video	Size of data	Setting	Annotation type	Context
Emotic [4]	Image	18316	Web	26 categories	Yes
Affecnet [23]		450,000	Web	8 categories	No
CAERS [5]	Image	70,000	TV show	7 Categories	Yes
AFEW [24]	Video	1,809 Clip	Movie	7 categories	No
CAER [5]	Video	13,201Clip	TV Show	7 categories	Yes

4.2 Experimental Setup

In this implementation, OpenCV was employed to crop the face images. We implemented this fusion model using the Pytorch library. We used the pretrained model Resnet 18 with the Places dataset. We conducted training on three variations of the CNN model: one exclusively for facial data, another solely for contextual information, and a third that combined both. These configurations are illustrated in Figure 6, utilising different input types and utilising distinct loss functions. Afterwards, we evaluated the performance of these models using the testing set. For every case, we determined the training parameters by considering the validation set. Table 2 displays the average precision (AP), which indicates the extent of accuracy obtained by the test set across different categories, as represented by the area under the precision-recall curves. The results in the first three columns are obtained by employing a unified loss function (Lcomb) with CNN architectures that only process the face (F, first column), solely the image (C, second column), and both the body and the image simultaneously (F + C, third column).

Incorporating information from both the body and image inputs yields the best results for all categories except “esteem.” This underscores the effectiveness of incorporating information from both sources for discrete category recognition. Notably, the results

obtained using only the image context (C) generally perform less favorably when compared to the other two inputs (F, C, and F+C). This observation aligns with the understanding that within the same scene, different individuals may exhibit varying emotions, even though they share most of the context.

This paper focuses on the issue of identifying emotional states within a given setting. The EMOTIC database is a collection of images featuring people in various real-life settings, rather than controlled conditions. The images are labelled based on the individuals' discernible emotional states, utilising a combination of two distinct types of annotations: the 26 emotional categories suggested and elucidated in this study, together with a CNN model designed for the purpose of estimating emotions within a given environment. The model utilises cutting-edge methods for visual recognition and serves as a standard for the task of measuring emotional states in a given scenario. A technology capable of perceiving emotions in a manner like to humans has a multitude of possible applications in fields such as human-computer interaction, human-assistive technologies, and online education, among others.

The primary objective of this study is to precisely determine emotional states in a particular context. The EMOTIC database is a collection of photos captured in unregulated environments, featuring persons in their personal surroundings. The photographs are annotated to portray the perceived emotional states of the individuals portrayed. This task involves the use of two distinct forms of annotations: the 26 emotional categories, which are elucidated and delineated in this investigation, and the three customary continuous emotional dimensions (valence, arousal, and dominance). Moreover, a Convolutional Neural Network (CNN) model is shown to precisely forecast emotions in particular contextual settings. The model incorporates cutting-edge techniques in visual recognition and establishes a benchmark for predicting contextual emotional states.

The utilisation of an advanced technology that can precisely discern emotions, akin to human perception, holds significant potential in various domains, including human-computer interaction, assistive technologies, and online education, among others (Fig. 4 and Table 3).

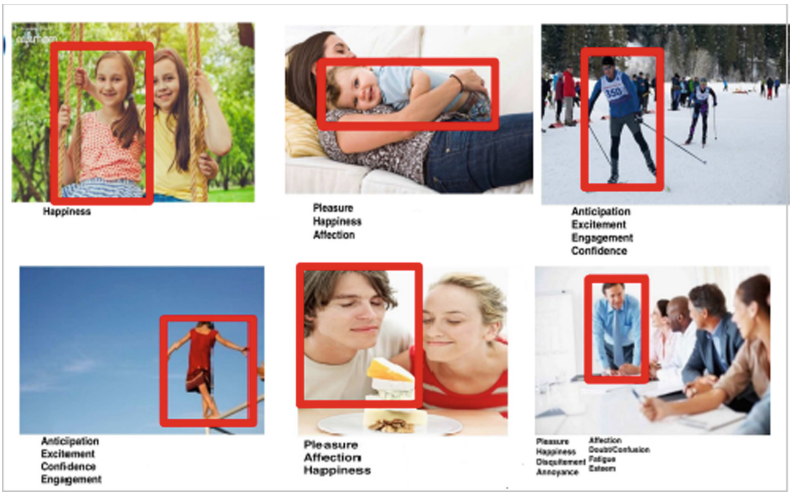


Fig. 4. Implemented model show the annotated emotion for given images

Table 3. Precision value for Emotic dataset

Sr. No.	Category	Face	Context	Face + Context
1	Peace	21.65	20.46	20.4
2	Affection	19.17	16.67	21.65
3	Esteem	9.34	17.56	18.45
4	Anticipation	56.34	49.28	52.49
5	Engagement	82.16	78.92	80
6	Confidence	73.59	65.6	69
7	Happiness	54.9	49.45	52.81
8	Pleasure	46.33	44.56	49
9	Excitement	74.78	68.45	70
10	Surprise	21.89	19.55	20
11	Sympathy	11.56	11.67	15.43
12	Doubt/	33.49	32.1	33.36
	Confusion			
13	Disconnection	14.47	12.49	16.25
14	Fatigue	9.53	8.34	10.76
15	Embracement	2.59	3.05	3.24
16	Yearning	8.66	8.01	8.92
17	Disapproval	11.97	6.32	10.04
18	Aversion	7.89	3.43	9.81
19	Annoyance	11.64	6.09	16.93
20	Anger	7.76	5.69	11.29
21	Sensitivity	5.14	4.77	5.08
22	Sadness	9.06	6.11	19.34
23	Fear	15.55	14.68	16.77
24	Pain		2.09	
25	Suffering	10.35	5.77	8.96
26	Disquietment	18.24	15.78	19.88

5 Conclusions

The primary objective of this study is to precisely discern emotional states in a particular context. The EMOTIC database is a collection of photographs captured in unregulated environments, displaying persons in their personal surroundings. The photographs are annotated to represent the perceived emotional states of the individuals represented. This is done using two types of annotations: the 26 emotional categories, which are introduced and explained in this study, and the three classic continuous emotional dimensions

(valence, arousal, and dominance). Moreover, this research presents a convolutional neural network (CNN) model that can precisely forecast emotions in different contextual settings. This model sets a benchmark for assessing contextual emotional states by using advanced techniques in visual recognition. The applicability of a system capable of discerning emotions in a manner akin to human perception is significant in various domains, including human-computer interaction, assistive technology, and online education.

References

1. Ekman, P.: Cross-cultural Studies of Facial Expression. Darwin and Facial Expression, pp. 169–220. Malor Books, Los Altos (2006)
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**, 124–129 (1971)
3. Fridlund, A.J.: Human facial expression: an evolutionary view. Nature **373**, 569 (1995)
4. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: CVPR (2017)
5. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. IEEE explore (2019)
6. Soleymani, M Pantic, M.; Pun, T. Multimodal Emotion Recognition in Response to Videos. IEEE Trans. Affect. Comput. 2012,3, 211–223
7. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Trans. Affect. Comput. **10**, 60–75 (2019)
8. Li, X., Peng, X., Ding, C.: Sequential interactive biased network for context-aware emotion recognition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–6. IEEE (2021)
9. Zhang, D., et al.: Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14338–14346 (2021)
10. Li, Y., Lu, G., Li, J., Zhang, Z., Zhang, D.: Facial expression recognition in the wild using multi-level features and attention mechanisms. IEEE Trans. Affect. Comput. (2020)
11. Caon, M., Angelini, L., Yue, Y., Khaled, O.A., Mugellini, E.: Context-aware multimodal sharing of emotions. In: Kurosu, M. (ed.) HCI 2013. LNCS, vol. 8008, pp. 19–28. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39342-6_3
12. Wu, S., Zhou, L., Hu, Z., Liu, J.: Hierarchical context-based emotion recognition with scene graphs. IEEE Trans. Neural Netw. Learn. Syst. (2022)
13. Mehendale, N.: Facial emotion recognition using convolutional neural networks (FERC). SN Appl. Sci. **2**(3), 1–8 (2020). https://doi.org/10.1007/s42452-020-2234-1
14. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: FER-net: facial expression recognition using deep neural net. Neural Comput. Appl. **33**, 9125–9136 (2021)
15. Johannßen, D., Biemann, C.: Neural classification with attention assessment of the implicit-association test OMT and prediction of subsequent academic success. In: KONVENS (2019)
16. Shenoy, A., Sardana, A.: Multilogue-net: a context aware RNN for multi-modal emotion detection and sentiment analysis in conversation (2020). arXiv preprint arXiv:2002.08267
17. Bendjillali, R.I., Beladgham, M., Merit, K., Taleb-Ahmed, A.: Improved facial expression recognition based on DWT feature for deep CNN. Electronics **8**, 324 (2019)
18. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In CVPR (2009)
19. B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. CoRR, abs/1610.02055, 2015

20. Ismatov, A., Enriquez, V.G., Singh, M.: FaceHub: facial recognition data management in blockchain. In: Lee, S.-W., Singh, I., Mohammadian, M. (eds.) Blockchain Technology for IoT Applications. BT, pp. 135–153. Springer, Singapore (2021). https://doi.org/10.1007/978-981-33-4122-7_7
21. Lin, T., etal.: Microsoft COCO: common objects in context. CoRR abs/1405.0312 (2014)
22. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through ade20k dataset (2016)
23. Prinz, J.: Which emotions are basic. Emot. Evol. Rational. **69**, 88 (2004)
24. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
25. Dhall, A., Goecke, R., Lucey, S., Gedeon, T., et al.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia (2012)
26. Jiang, Y.G., Xu, B., Xue, X.: Predicting emotions in user-generated videos. In: AAAI (2014)
27. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
28. You, Q., Jin, H., Luo, J.: Visual sentiment analysis by attending on local image regions. In: AAAI (2017)
29. Ko, B.C.: A brief review of facial emotion recognition based on visual information. Sensors **18**, 401 (2018)