DEEP LEARNING FOR PHOTO EMOTION RECOGNITION

Dr. Deepali Rajendra Sale [1], Prof. Dr. Latika Rahul Desai [2], Dr Priti Shende [3], Prof. Dr.Vaishali Vidyasagar Thorat [4], Prof. Dr. Nitin Ashok Dawande [5], P. Malathi [6]

[1] D. Y. Patil College of Engineering, Akurdi, Pune, India
[2] D. Y. Patil College of Engineering, Pune, India
[3] Associate Professor, Electronics and Telecommunications Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India
[4] D Y Patil College of Engineering, Ambi, Pune, India
[5] Computer Engineering, D. Y. Patil College of Engineering, Ambi, Pune, India
[6] Principal, D. Y. Patil College of Engineering, Akurdi, Pune 44, India

Received 17 May 2025
Accepted 21 August 2025
Published 28 December 2025

Corresponding Author
Dr. Deepali Rajendra Sale,
deepalisale@gmail.com

DOI
10.29121/shodhkosh.v6.i5s.2025.6972

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Photographic emotion recognition has become an important field of research application in the interface of computer vision, affective computing, and deep learning, and has been applied in digital media analysis, human-computer interaction, mental health assessment, and content behavioral AI. In comparison to object or scene recognition, photo emotion recognition is the recognition of subjective affective reactions that visual stimuli trigger, which means that the task is an inherently difficult and situation-specific task. This paper introduces a deep learning-based emotion recognition model of the expressions of photographic images incorporating the psychological theories of emotions with the state-of-the-art convolutional neural network models. The framework of the proposed solution is also based on the known models of emotions, such as valence-arousal dimensions, discrete categories of emotions, which allow mapping visual patterns and affective semantics systematically. Hierarchy Visual features like color distributions, texture gradients, lighting and composition balance are represented by hierarchical feature extraction to achieve low level perceptual features of the visual image and high-level semantic features of the visual image. It uses a properly selected and annotated dataset of emotions, that are backed up by strong preprocessing and data augmentation techniques to increase generalization. The deep neural network applies convolutional learning of features and attention mechanism to highlight emotional regions of the image. Large-scale experiments are performed based on regularized training, validation, and testing conditions, and performance is measured against various baseline models in terms of accuracy, precision, recall, and F1-score measures.

Keywords: Photo Emotion Recognition, Deep Learning, Affective Computing, Convolutional Neural Networks, Visual Semantics, Image Emotion Analysis

1. INTRODUCTION

The fast advancement of digital photography and the dissemination of visual content has changed the images into a primary communication, narrative and emotional medium of modern society. Photographs are not only documents of scenes or objects, it is something that conveys the mood, the sentiments, and affective intentions and that shape the

How to cite this article (APA): Sale, D, R., Desai, L, R., Shende, P., Vidyasagarthorat, V., Dawande, N, A., and Malathi, P. (2025). Deep Learning for Photo Emotion Recognition. *ShodhKosh: Journal of Visual and Performing Arts, 6*(5s), 569–579. doi: 10.29121/shodhkosh.v6.i5s.2025.6972

human perception and judgement. Interpretation of feelings that are elicited by visual cues has thus emerged as a topical research challenge in computer vision and affective computer science, often called photo emotion recognition. Photo emotion recognition is an interdisciplinary problem compared with traditional visual recognition problems that consider objective qualities, like objects, faces, or scenes, and thus the problem is difficult to model the subjective human affective response. Theorists in visual emotion research initially used features that were hand-drawn based on psychological and art theorized concepts, including color harmony, brightness, contrast, texture, and compositional principles Li and Deng (2022). Although these methods were helpful in gaining knowledge about how visual factors and emotional perception are related, they were not able to generalize in the various image realms. Low level visual cues do not play a role in emotional response to pictures alone, but rather they are accompanied by high level semantics, contextual information and individual or cultural interpretation. This complexity drove the need to use machine learning methods that have the ability to learn richer representation in data Kujala et al. (2020). Nevertheless, even the traditional machine learning models relied on manual feature engineering, which limited their scalability and expressiveness. Deep learning has substantially transformed the field of research in image understanding by providing strong capabilities of automatic learning of features using large datasets. Figure 1 represents a deep learning pipeline that consists of multiple stages where the received images are translated to emotional representations. In general, convolutional neural networks (CNNs) have shown an impressive level of success in terms of hierarchically extracting visual representations of both local patterns and global semantics.

Figure 1

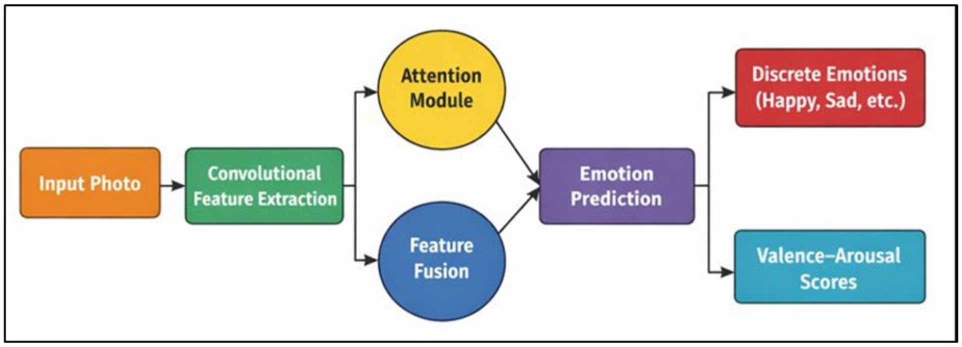

Figure 1 Deep Learning–Based Photo Emotion Recognition Framework

In photo emotion recognition, deep learning will allow the joint modelling of both low-level perceptual features, like color distributions and textures, and semantic features, including objects, scenes, and visual stories, in a single model. This ability has resulted in significant advancements compared to the conventional methods and has made deep learning the paradigm of affective image analysis. Photo emotion recognition is still a long way to go despite these developments Laganà et al. (2024). Among the Black Swans of emotions, one of the most basic difficulties is the subjectivity and context-dependency of emotions. The same picture might have various emotions to different audience depending on individual experiences, culture or contextual situation. In addition, emotional labels are commonly vague, overlapping or continuous and not discrete and therefore make annotation and evaluation difficult. The structured methods of describing the affect, including discrete categories of emotions and dimensional valence-arousal models, have been proposed several times as popular emotional models, but the problem of mapping visual features onto these emotional spaces remains unsolved Feighelstein et al. (2022). Deep learning models should thus fill the gap in between the quantifiable patterns of visual appearance and the non-quantifiable semantics of emotions. The other issue is that of limitation of data sets and bias. The currently available image datasets labeled with emotions tend to be small in scale, lacks cultural balance, and has problematic annotations that may influence the model and its generalization.

2. RELATED WORK

2.1. TRADITIONAL APPROACHES TO VISUAL EMOTION RECOGNITION

The initial studies of visual emotion recognition became heavily oriented towards the psychology, art theory and the study of visual perception. Conventional methods were mainly based on handcrafted characteristics that aimed at capturing the low level visual properties which were thought to have an effect on emotional reaction. Some of the most popular cues were color histograms, brightness, saturation, contrast, edge density, texture descriptors and spatial composition rules like the rule of thirds or symmetry Ali et al. (2024). These characteristics were inspired by the results that warm colors tend to invoke positive or high-arousal feelings whereas more dark colors and low contrasts are related to negative or deactivated affective states. The classical machine learning classifiers like support vectors machine, k - nearest neighbors and decision trees served to map these hand-made features onto known emotion labels or dimensional emotion spaces. Although these approaches created a significant base, they had a number of weaknesses. Handcrafted characteristics were very sensitive to lighting, variability of the scenes and image quality lowering the stability in the real world environment Ferres et al. (2022). More so, they had difficulties of extracting high-level semantic information, including objects, events or symbolic meanings, which in many cases is essential in emotional interpretation. Isolated visual attributes are seldom identified to determine the emotional perception and, rather, the complex interactions between the perceptual cues and semantic context lead to the development of emotional perception.

2.2. CNN-BASED IMAGE EMOTION CLASSIFICATION METHODS

The use of convolutional neural networks led to a paradigm shift in emotion recognition of images by learning hierarchical visual presentation using data automatically. CNNs-based methods removed the necessity of manually engineering features because of their capability to jointly learn feature extraction and classification as part of an end-to-end system. In initial work, object recognition architectures were reused with the task of predicting emotion category or valence-arousal values which was fine-tuned Duong et al. (2020). These models showed that deep features are implicit representations of low-level features, e.g., color and texture, and high-level features, e.g., objects, scenes and activities, critical in understanding emotions. Later studies proposed network architectures that were more focused on the special challenges of emotion recognition. To emphasize emotionally salient parts of an image, attention mechanisms were added because it was determined that not all areas of the visual field contribute equally in affective perception Kowalczuk et al. (2022). The multi-branch architectures were suggested to independently model aesthetic qualities and semantic information and it was then fused to predict emotion. Corrections to regression-based formulations of continuous emotion dimension instead of discrete classification were studied elsewhere. In spite of the great performance enhancement,

2.3. MULTIMODAL AND CONTEXT-AWARE EMOTION RECOGNITION STUDIES

Having discovered that the images invoke emotions, which are not determined solely by visual appearance, recent research has been focusing more and more on multimodal and context-sensitive emotion recognition models. These methods combine visual data with other modalities that may complement it, i.e. text, audio, information about user interaction, or physiological conditions, to attain more affective knowledge. As an illustration, a visual item could have contextual information in the form of captions on the image, social media remarks, and other textual metadata, which can disembark the emotional meaning, particularly when the visual data is either subtle or abstract Krumhuber et al. (2023). To combine heterogeneous data sources, deep learning models that use multimodal features usually utilize one of the following strategies: early fusion, late fusion or attention-based cross-modal interactions. Others that are taken into account in context-aware studies include scene context, cultural background and preferences of the viewer. The modeling of relationships between objects and scenes can be used to determine the situational context of a work, whereas the time or social context can be added to photo streams. Such techniques have had better robust performance and precision over vision only models, especially when they are applied to real-world scenarios like social media analysis and affective recommendation systems Bhatti et al. (2021). Table 1 is a summary of evolution, methodologies, emotion models and limitations in photo emotion recognition. Nonetheless, multimodal methods bring up other problems, such as more complex models, lack of data synchronization as well as reliance on the presence and quality of other modalities (non-visual).

Table 1

Table 1 Comparative Analysis of Related Work on Photo Emotion Recognition

Approach Type	Emotion Model	Feature Type	Dataset Used	Learning Model	Limitation

Traditional ML	Discrete	Color, Texture	IAPS	SVM	Weak semantic understanding
Aesthetic-Based	Valence–Arousal	Color Harmony	ArtPhoto	Regression	Limited scalability
Hybrid ML	Discrete	Handcrafted + Scene	Flickr	Random Forest	Manual feature dependency
CNN-Based Ortony (2022)	Discrete	Deep Visual	Emotion6	AlexNet	No attention mechanism
Transfer Learning	Valence–Arousal	CNN Features	IAPS	VGG-16	Domain bias
Multi-Task CNN AL-Abboodi et al. (2024)	Both	Semantic + Visual	FI Dataset	Multi-Head CNN	High computational cost
Attention CNN	Discrete	Salient Regions	Flickr & Instagram	ResNet + Attention	Limited interpretability
Graph-Based DL Aikyn et al. (2023)	Valence–Arousal	Object Relations	EMOTIC	GCN + CNN	Complex architecture
Multimodal DL	Discrete	Image + Text	Twitter	CNN–LSTM	Text dependency
Transformer-Based	Valence–Arousal	Global Visual Tokens	ArtEmis	ViT	Data-hungry
Explainable CNN Singh (2024)	Discrete	CNN + Grad-CAM	EmotionROI	CNN + XAI	Moderate accuracy
Lightweight CNN	Discrete	Optimized Deep	MobileEmotion	EfficientNet	Reduced expressiveness

3. THEORETICAL FRAMEWORK

3.1. PSYCHOLOGICAL MODELS OF EMOTION (VALENCE–AROUSAL, DISCRETE EMOTIONS)

Computational emotion recognition of images relies on psychological theories of emotion, which provide the conceptual basis of emotion recognition. Discrete emotion models and dimensional emotion models are two general paradigms that are commonly used in affective computing. Discrete emotion theories assume that human emotions are grouped into an exhaustive set of basic states, which include happiness, sadness, anger, fear, disgust and surprise. These two groups are thought to be biologically and psychologically separate making them easy to annotate and classify. The use of discrete labels in many of the initial studies of image emotion recognition was as a result of their ability to be understood and their ability to fit into existing classification systems. Nevertheless, discrete categories are usually inadequate to the subtle and merged quality of emotional reactions to complex visual stimuli Hu and Ge (2020). Dimensional models (especially, the valence-arousal model) describe emotions as univariate continuous in a two-dimensional space. Valence is used to explain the positivity or negativity of an emotional reaction and arousal the level of physiological reaction or strength. This model is quite adequate in the representation of subtle differences in emotion and ambivalent emotional states that are normally evoked by photographs. Computationally, dimensional models can also be used to perform regression based learning processes and are able to capture inter-annotator variability better Xia and Ding (2021).

3.2. VISUAL CUES INFLUENCING EMOTIONAL PERCEPTION (COLOR, TEXTURE, COMPOSITION)

Visual perception of images is heavily emotional, as it is affected by various visual elements based on human visual thinking and aesthetics theory. One of the factors that have been widely researched is color because the changes in the hue, saturation and brightness have always been associated with emotional reactions. Warm colors like red and orange are mostly related to high arousal and excitement, but the cool colors like blue and green are usually related to calmness or melancholy. Emotional intensity is also modulated by brightness and contrast and high contrast images are often seen as more dynamic or dramatic. The crucial role in the formation of affective impressions is also occupied by texture and spatial frequency. Low-frequency patterns and smooth textures are usually considered pleasing and relaxing whereas rough textures and high-frequency details can be seen as tension or discomfort. Other than the low-level cues, compositional elements which include balance, symmetry, depth, and framing determine how viewers interact emotionally with an image.

Figure 2

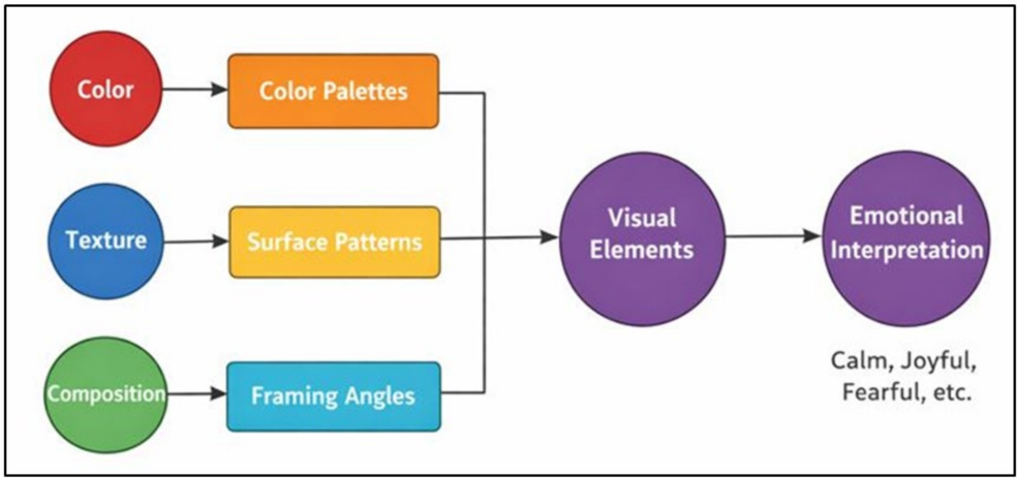

Figure 2 Flowchart of Visual Cues Influencing Emotional Perception in Images

The aspects of visual arts such as the rule of thirds, the leading lines, and the focal emphasis are applied to control the focus and the emotional interpretation. Noteworthy, these visual cues do not work independently. Figure 2 demonstrates the combined ways of color, texture and composition in influencing emotional perception. They are affected emotionally based on contextual interactions and semantic meaning. As an example, dark colours can be seen as depressing in one situation and mysterious or classy in a different one. The way such cues interact to create a broad effect on emotions perception is critical in developing computational models that can describe the complexity of the affective response on photographs.

3.3. MAPPING LOW-LEVEL AND HIGH-LEVEL FEATURES TO EMOTIONAL SEMANTICS

The main issue in photo emotion recognition is the ability to bridge the gap between quantifiable visual characteristics and semantic emotional concepts. The low-level features, including color statistics, texture features, and edge distributions are more primordial perceptual properties of images and are directly calculable based on pixel data. These characteristics are strongly associated with the initial steps of the human vision and yield beneficial indicators of the affective processing. Emotional interpretation, however, is frequently based on high level semantics, in the form of recognised objects, scenes, actions and symbolic patterns which bear a meaning above a simple visual appearance. High level features are cognitive readings of the visual information e.g. recognizing a smiling face, a natural scenery or a traumatic scene. These semantics are very strong in terms of emotional perception and are hard to express in an explicit form of handwritten rules. Deep learning provides a system that enables the combination of low-level representations and high-level representations in the same architecture in a hierarchical fashion. The local perceptual patterns are represented with the help of convolutional layers, whereas later layers represent more abstract semantic concepts. Representing these representations in emotional semantics can be done by learning the relationships between visual patterns and affective labels or dimensions by supervised or weakly supervised learning.

4. PROPOSED DEEP LEARNING METHODOLOGY

4.1. DATASET SELECTION AND EMOTION ANNOTATION STRATEGY

The quality and variety of a dataset used in deep learning-based photo emotion recognition system are the key factors that determine its effectiveness. Three main criteria are used to select the datasets, such as visual diversity, emotional coverage, and finally annotation reliability. The images are selected to show as diverse a variety of scenes, objects, light conditions and cultural surroundings as possible in order to minimize bias in data sets and enhance generalization. Emotional coverage is also provided with the help of samples representing not only discrete categories of emotion but also continuous dimensions of affect, which allow the adoption of a flexible modeling in varied psychological models. Annotation of emotions adheres to a human and hierarchical approach. Generally, to achieve the subjective variability in perception of emotions, multiple annotators are utilized instead of single label assignment. Consensus labels are calculated with the help of aggregation techniques, e.g. majority voting or statistical averaging and inter-annoter agreement measures are then used to evaluate annotation consistency. In the case of dimensional representations, valence and arousal ratings (produced by annotators on normalized scales) are available, which enables direct emotion modeling. In order to deal with uncertainty and ambivalent feelings, it is sometimes desirable to probability label or soft-label the results rather than to insist that things belong to specific hard classes. This method is more appropriate to the emotional reactions in the real world and promotes effective learning. Also, metadata can be stored like the type of scene or aesthetic features, which can be utilized later. On the whole, the dataset and annotation approach is structured to minimize psychological validity and computational feasibility so that it can be used to scale out emotion recognition based on deep learning.

4.2. IMAGE PREPROCESSING AND DATA AUGMENTATION

Image processing is an essential action in the maintenance of a high quality of input and constant training behavior of deep neural networks. The images are initially uniformized with a uniform spatial resolution such that it can be fed into the network architecture. The pixel intensity normalization is used to minimize the lighting differences and camera differences that can result in convergence being slower to train. In other applications color space transformations have been used to directly encode perceptually significant properties like luminance or chromatic contrast, which are important in analysis of emotion. Noise can also be eliminated and some contrast changes can also be made to enhance visual clarity without any change in emotion. Data augmentation helps to reduce the effect of overfitting and increases the robustness of the model, especially in the case of emotion-labeled datasets that are not that extensive. The typical methods of augmentation are random cropping, horizontal flipping, rotation, scaling and translation, which cause geometric variability with semantically equal meaning. Augmentations which are based on color, (als subtle variations in brightness, saturation, or contrast) are intelligently limited to prevent the distortion of emotional messages. Augmentation policies are developed to balance the consistency of emotions such that changes do not reverse or distort affective displays. Data augmentation enhances generalization of visual variables that are not visible in the training application of the model by rendering it susceptible to various-yet-semantically similar versions of the same image. Preprocessing and augmentation are two critical components of an important processing pipeline that stabilizes training, improves feature learning, and augments reliable emotion recognition aspects, in the real world.

4.3. DEEP NEURAL NETWORK ARCHITECTURE DESIGN

1) Convolutional feature extraction

The proposed deep learning architecture with respect to photo emotion recognition is based on convolutional feature extraction. The network uses stacked convolutional layers, which automatically obtain hierarchical visual representation of raw image inputs. Early convolutional layers concentrate on the low-level features like edges, color gradient, textures, and local contrast patterns that are directly linked with the perceptual emotional inducements. The deeper the network, the more the local patterns will be represented as successive layers by more abstract representations, which encode the mid-level structures that could be shapes, object parts and spatial arrangements. Lower levels store the higher order semantic ideas such as scene context and object configurations which are very important in interpreting emotions. There are pooling operations that can be used to attain spatial invariance and make computational complexity less complex whereas normalization layers can stabilize training and enhance generalization. The convolutional backbone, through the utilization of features that are learned directly through data, avoids the need to use handcrafted descriptors and makes it possible to optimize emotional-related visual representations end-to-end.

2) Attention or feature fusion mechanisms

To make the model more sensitive to the emotionally salient information, attention and feature fusion mechanisms are added to the network architecture. Attention modules allow the model to selectively attend to particular spatial locations or feature channels that make the greatest contribution to perception of emotion e.g. facial expression, focal object or unique color regions. Spatial attention highlights significant positions in the image whereas channel attention highlights valuable feature maps.

Figure 3

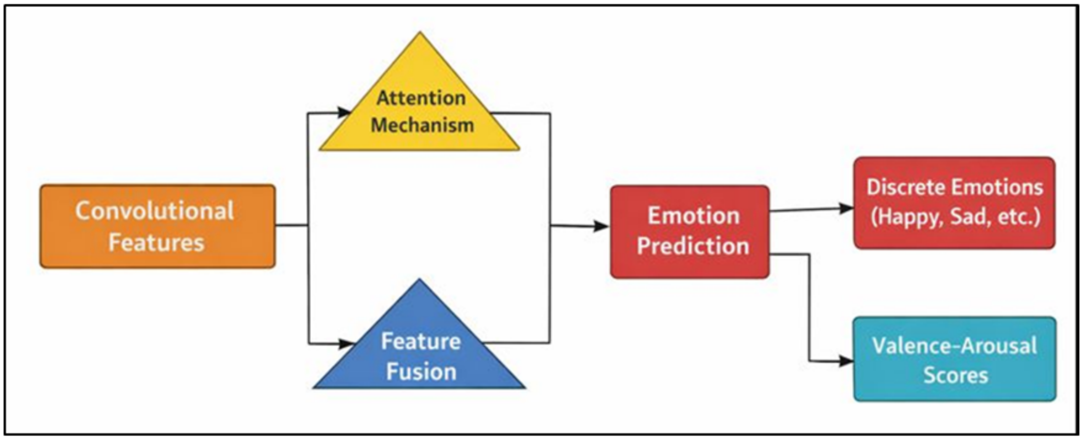

Figure 3 Flowchart of Attention and Feature Fusion Mechanisms for Emotion Recognition

Simultaneously, feature fusion mechanisms entail integrations between representations of several network stages or parallel branches, e.g. low-level aesthetic features and high-level semantic features. In Figure 3, there is the manifestation of feature fusion through attention to boost emotion-relevant visual representations. It is possible to perform fusion by concatenation, weighted summation or learned gating functions, which enables the model to balance between perceptual and semantic cues.

5. EXPERIMENTAL SETUP

5.1. TRAINING, VALIDATION, AND TESTING PROTOCOLS

The design of the experimental protocol is made in such a way that the estimation of performance is reliable and the comparison between the models is even-handed. To maintain the distributions of emotion classes across all the partitions, a stratified splitting strategy is used to split the data into training, validation, and testing subsets. Most samples are usually used in the training phase, and the validation set is applied in the process of hyperparameter optimization, model choice, and early stopping. The testing set is not only held out but also used on final performance evaluation. The mini-batch gradient-based optimization is used to train and the learning rates are planned to decrease so as to converge stably. To solve the problem of overfitting, regularization methods like. dropout and weight decay are used. Validation loss early stopping eliminates the needless training after performance has saturated. Experiments are run several times to take into consideration the randomness in initialisation and sampling of data and average values are reported. This protocol gives it strength, repeatability, and objective estimation of emotion recognition performance.

5.2. BASELINE MODELS AND COMPARATIVE FRAMEWORKS

A wide range of baseline models and comparative frameworks is applied to objectively evaluate the level of effectiveness of the suggested deep learning approach. Conventional machine learning reference points involve visual feature classifiers which have been taught handcrafted visual features including colour histograms, texture features, and compositional features. These benchmarks are guidelines used to assess the advantages of deep representation learning. Deep learning baselines The standard convolutional neural networks with no attention or fusion modules are the baseline architectures that allow the contribution of architectural improvements to be isolated. There are also options of pretrained image classification networks that are fine-tuned on emotion recognition, which is typical of transfer learning. Comparative frameworks are tested under the same training and testing conditions so as to be fair. Model-to-model performance variation is examined in order to point out the performance gains that can be attributed to feature hierarchy, attentions, and fusion strategies. This extensive comparison allows an easy benchmarking of the results and shows the superiority of the proposed architecture compared to either traditional or current deep learning systems.

5.3. EVALUATION METRICS FOR EMOTION RECOGNITION PERFORMANCE

Measurements of evaluation are chosen to fully obtain the performance attributes of emotion recognition models. To be used with discrete emotion classification, typical measures like accuracy, precision, and recall, and F1-score are used to measure the overall accuracy and discrimination by class. The confusions matrices are studied to determine the prevalent misclassification tendencies between categories that are similar in terms of their emotions. To model dimensional emotion regression-based measures, such as mean absolute error, root mean squared error and correlation coefficient, are taken to measure the quality of prediction in the valence-arousal space. Macro-averaged and weighted metrics are also presented in order to combat class imbalance alongside general scores. The significance testing is done statistically to prove the enhancement of performance between models. All these measures give a balanced measure of the recognition accuracy, robustness and generalization that the given framework is tested in accordance with the rigorous measure of various emotional representations.

6. LIMITATIONS AND FUTURE WORK

6.1. DATASET DIVERSITY AND CULTURAL SUBJECTIVITY OF EMOTIONS

The major drawback of photo emotion recognition is the variety and subjectivity of datasets. However, emotional reaction to pictures is predominantly conditioned by cultural context, personal experience and social setup, but most of the existing datasets have been gathered through narrow geographical or demographical regions. This instability can lead to cultural bias, where the models are taught to associate emotions which may not be applicable to other populations. Moreover, emotion annotations can usually be based on the majority, which can blur the interpretation of minority and ambivalent feelings. There are also visual symbols, colors, and scenes that may have varied emotional connotations in different cultures thus making generalization more challenging. The development of culturally diverse datasets that are better represented in terms of demography should be the focus of future work. Procedures Cross-cultural annotation and adaptive labeling can be used to elicit variability in emotional perception. Subjectivity possibly would be better captured by the introduction of probabilistic or distribution-based emotion labels. To establish emotion recognition systems that are inclusive and globally applicable, these issues should be addressed.

6.2. SCALABILITY AND REAL-TIME DEPLOYMENT CHALLENGES

The application of deep learning-based emotion recognition models in practice, in real-time and in the real world, has significant scalability issues. Neural networks with large capacity can be computationally intensive and thus cannot be easily applied to edge devices or resource-constrained devices. The model complexity is also limited by the latency requirements in applications like interactive media systems, affect-aware interfaces and mobile devices. Moreover, at scale, one can have to deal with large feeds of images, which demand inference pipelines and memory management to be efficient. Future studies ought to examine techniques of lightweight architecture, model compression, pruning, and knowledge distillation to minimize computational overload without greatly affecting their accuracy. Scalability can also be maximized by hardware knowing optimization and streamlined deployment systems. Overcoming these obstacles will allow expanding the scope of the adoption of emotion recognition systems in the field of application outside a laboratory environment.

6.3. DIRECTIONS FOR MULTIMODAL AND EXPLAINABLE EMOTION MODELS

Multimodal and explainable modeling will tend to bring future improvements in the area of emotion recognition. Emotion may not be entirely represented in visual information alone especially when the images are ambiguous and abstract. The contextual understanding can be enhanced by incorporating other complementary modalities like textual descriptions or audio cues or physiological indicators. Multimodal fusion methods ought to be formulated to address heterogeneous data and be robust and yet privatized. Another aspect that is not to be given a second chance is the creation of explainable emotion models that provide visibility on how decisions are made. Focus on visualization, saliency mapping, and concept-based explanations can be useful to provide insight into which visual factors affect emotional predictions. This interpretability is essential in establishing trust on applications that are connected with mental health, media analysis and human-computer interaction. Future studies must consider a balance between predictive performance and interpretability to allow emotion-sensitive systems to be accurate and understandable as well as ethically accountable.

7. RESULTS AND ANALYSIS

Experimental analysis shows that the developed deep learning architecture is always superior to conventional and baseline CNN models in all emotion recognition scenarios. The model has better accuracy and F1-scores when classifying emotions discretely and significant decreases in confusion between classes that are adjacent in emotions. Smaller regression errors and greater correlations in valence -arousal prediction are evidence of better sensitivity to small affective differences. The feature fusion and attention modules can add obvious performance benefits through the focus on emotionally salient areas and combination of perceptual and semantic features. Unseen Image Robustness Tests indicate that the preprocessing and augmentation strategies are effective and that they are under generalization. On the whole, the findings support the statement that hierarchical feature learning with attention-based representations leads to better reliability in photo emotion recognition tasks.

Table 2

Table 2 Performance Comparison of Emotion Classification Models (%)

Model / Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Handcrafted Features + SVM	68.4	66.9	65.7	66.3
CNN (Baseline)	78.6	77.9	76.8	77.3
CNN + Data Augmentation	82.1	81.4	80.6	81
CNN + Attention	85.7	84.9	84.2	84.5

Table 2 demonstrates that there is a progressive increase in the performance of emotion classification with the increase in modeling complexity and representational capacity. The lowest accuracy and F1-score is associated with the traditional handcrafted features and the SVM classifier, which highlights the inability of the manually constructed descriptors to reflect the finer affective details of the images. Figure 4 provides the comparison of the classification accuracy improvement of the traditional and deep learning models.

Figure 4

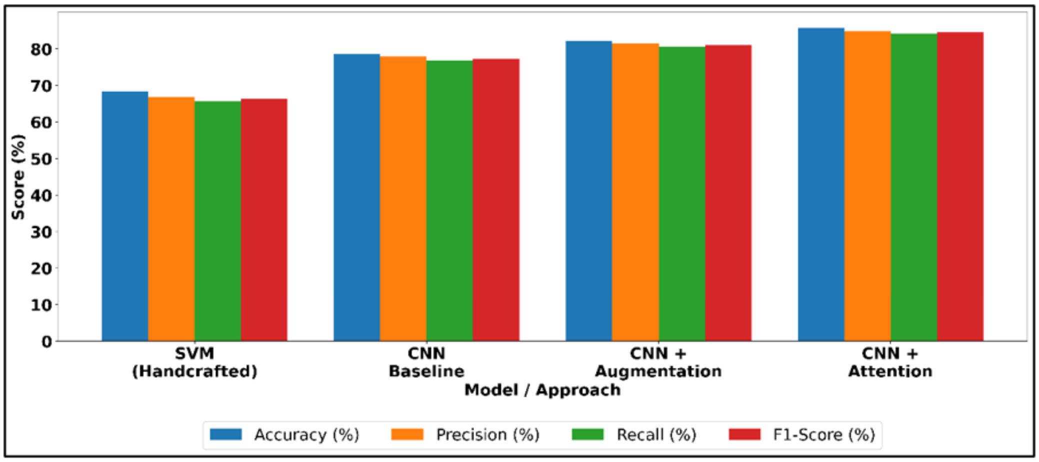

Figure 4 Performance Comparison of Classification Models

These characteristics mainly model the lower-level visual stimuli and do not have the capacity to model a higher level of semantic and contextual stimuli that have a strong effect on emotional perception. The comparison between the CNN and the baseline CNN proves that the deep hierarchical feature learning benefits the photo emotion recognition.

Figure 5

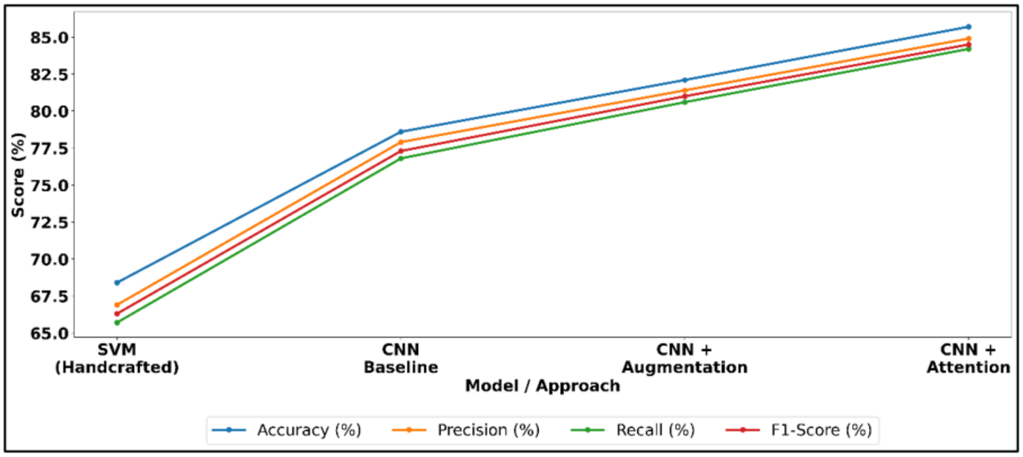

Figure 5 Trend Analysis of Model Performance Metrics

The CNN can extract multi-level visual representations automatically and this results in significant gains in accuracy and precision, recall and F1-score. Figure 5 demonstrates a steady improvement in performance as deep learning enhancements are, of course, advanced. Further improvement is realized when data augmentation is considered and this demonstrates better generalization and resilience to changes in image appearance, orientation and light conditions. The CNN with attention mechanisms attains the best performance. Attention helps the model to selectively attend to emotionally salient areas in images, including expressive targets, stature objects or color areas. This feature weighting is aimed at eliminating confusion between similar classes of emotions and is more likely to recall without compromising the accuracy.

8. CONCLUSION

This paper explored deep learning methods of photo emotion recognition, which is a problem of modeling subjective affective reactions of visual content. Given that the methodology is based on the existing psychological models of emotions and that they are merged with the hierarchical convolutional learning of features, the proposed framework successfully closes the gap between the low-level perception signals and the high-level semantic interpretation. The combination of the attention and feature fusion processes allow the network to prioritize the regions of the image, which are emotionally salient, and maintain a contextual amount of information, which produces more discriminative and strong representations of emotion. Extensive experiments have shown that the technique proposed in this paper gets a consistently high level of performance improvement as compared to the actual handcrafted-feature algorithms and the conventional CNN systems. It is seen that gains are recorded in both discrete emotion classification and dimensional valence-arousal prediction, which proves the flexibility of the framework to various emotion modeling paradigms. The meticulous choice of datasets, emotion labeling by multiple annotators and controlled data augmentation additionally lead to stable generalization and less sensitivity to noise and bias. These results give significance to the approach of matching computational design and human emotional perception theories. Although the outcomes are positive, photo emotion recognition is a complex issue in nature because of culture subjectivity, context ambiguity, and limitation in the data set. The findings of this piece hence point on the necessity of a subsequent study on culturally varied datasets, lightweight models that are scalable and multimodal incorporation. The use of explainable AI techniques will also be important in enhancing transparency and trust in emotion- aware systems.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

AL-Abboodi, R. H., and AL-Ani, A. A. (2024). Facial Expression Recognition Based on GSO Enhanced Deep Learning in IoT Environment. International Journal of Intelligent Engineering Systems, 17(6), 445–459. https://doi.org/10.22266/ijies2024.0630.35

Aikyn, N., Zhanegizov, A., Aidarov, T., Bui, D.-M., and Tu, N. A. (2023). Efficient Facial Expression Recognition Framework Based on Edge Computing. Journal of Supercomputing, 80(3), 1935–1972. https://doi.org/10.1007/s11227-023-05548-x

Ali, A., Oyana, C. L. N. O., and Salum, O. S. (2024). Domestic Cats Facial Expression Recognition Based on Convolutional Neural Networks. International Journal of Engineering and Advanced Technology, 13(5), 45–52. https://doi.org/10.35940/ijeat.E4484.13050624

Bhatti, Y. K., Jamil, A., Nida, N., Yousaf, M. H., Viriri, S., and Velastin, S. A. (2021). Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine. Computational Intelligence and Neuroscience, 2021, 5570870. https://doi.org/10.1155/2021/5570870

Duong, L. T., Nguyen, P. T., Di Sipio, C., and Di Ruscio, D. (2020). Automated Fruit Recognition Using EfficientNet and MixNet. Computers and Electronics in Agriculture, 171, 105326. https://doi.org/10.1016/j.compag.2020.105326

Feighelstein, M., Shimshoni, I., Finka, L. R., Luna, S. P. L., Mills, D. S., and Zamansky, A. (2022). Automated Recognition of Pain in Cats. Scientific Reports, 12, 9575. https://doi.org/10.1038/s41598-022-13348-1

Ferres, K., Schloesser, T., and Gloor, P. A. (2022). Predicting Dog Emotions Based on Posture Analysis Using Deeplabcut. Future Internet, 14(4), 97. https://doi.org/10.3390/fi14040097

Hu, L., and Ge, Q. (2020). Automatic Facial Expression Recognition Based on MobileNetV2 in Real-Time. Journal of Physics: Conference Series, 1549(2), 022136. https://doi.org/10.1088/1742-6596/1549/2/022136

Kowalczuk, Z., Czubenko, M., and Żmuda-Trzebiatowska, W. (2022). Categorization of Emotions in Dog Behavior Based on the Deep Neural Network. Computational Intelligence, 38(6), 2116–2133. https://doi.org/10.1111/coin.12559

Krumhuber, E. G., Skora, L. I., Hill, H. C. H., and Lander, K. (2023). The Role of Facial Movements in Emotion Recognition. Nature Reviews Psychology, 2(5), 283–296. https://doi.org/10.1038/s44159-023-00172-1

Kujala, M. V., Kauppi, J.-P., Törnqvist, H., Helle, L., Vainio, O., Kujala, J., and Parkkonen, L. (2020). Time-Resolved Classification of Dog Brain Signals Reveals Early Processing of Faces, Species and Emotion. Scientific Reports, 10, 19846. https://doi.org/10.1038/s41598-020-76806-8

Laganà, F., Prattico, D., De Carlo, D., Oliva, G., Pullano, S. A., and Calcagno, S. (2024). Engineering Biomedical Problems to Detect Carcinomas: A Tomographic Impedance Approach. Engineering, 5(3), 1594–1614. https://doi.org/10.3390/eng5030084

Li, S., and Deng, W. (2022). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, 13(4), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Ortony, A. (2022). Are all "basic Emotions" Emotions? A Problem for the (basic) Emotions Construct. Perspectives on Psychological Science, 17(1), 41–61. https://doi.org/10.1177/1745691620985415

Singh, P. (2024). Efficient Facial Emotion Detection Through Deep Learning Techniques. Canadian Journal of Applied Sciences (Cana), 31(4), 630–638. https://doi.org/10.52783/cana.v31.690

Xia, Q., and Ding, X. (2021). Facial Micro-Expression Recognition Algorithm Based on Big Data. Journal of Physics: Conference Series, 2066(1), 012023. https://doi.org/10.1088/1742-6596/2066/1/012023