Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Intelligent Evaluation, Emotion Analytics,
and AI-Based Assessment in Digital Art 2025 6(3s)

EMOTION RECOGNITION IN DIGITAL ART USING DEEP LEARNING

Dr. Anmol Suryavanshi [1], Nidhi Sharma [2], Manisha Chandna [3], Lakshya Swarup [4], Sunitha BK [5], Subramanian Karthick [6]

[1] Assistant Professor, Department of Computer Engineering, Shri Shivaji Vidya Prasarak Sanstha's Bhausaheb Shivajirao Deore College of Engineering, Dhule, Maharashtra, India
[2] Professor, Department of Development Studies, Vivekananda Global University, Jaipur, India
[3] Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India
[4] Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh, Solan, 174103, India
[5] Assistant Professor, Department of Management Studies, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India
[6] Department of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, 411037 India

Received 01 March 2025
Accepted 06 July 2025
Published 20 December 2025

Corresponding Author
Dr. Anmol Suryavanshi,
eranmol89@gmail.com

DOI
10.29121/shodhkosh.v6.i3s.2025.6779

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

ABSTRACT

The concept of emotion recognition in visual media has been an area of increasing discussion due to the emergence of digital art as a powerful creative art form in the digital realm. Digital art in contrast to photographs can have stylized and exaggerated, or non-realistic visual qualities, which adds further complexity to affective interpretation. This paper examines a framework of deep-learning-based emotion detection in digital art using a combination of the principles of the psychological theory of emotions and computer vision advancements. We discuss the affective perception of human viewers, basing on the simple emotions described by Ekman, the wheel of emotions by Plutchik, and the circumplex model offered by Russell to form a powerful labeling system that can be applied to artistic images. This is a selective collection of digital artworks which is compiled based on multiple online collections, and then annotated systematically according to structured guidelines assembling form of reduced subjectivity. The given model is a hybrid of CNN-based local feature extraction and transformer-based global attention mechanisms that can extract both fine-grained stylistic cues and larger compositional patterns which are characteristic of digital art. The experimental findings prove that the hybrid architecture compares to standalone CNN and ViT baselines in classifying emotional categories and especially in artworks that have an abstract or non-photorealistic style.

Keywords: Emotion Recognition, Digital Art, Deep Learning, Vision Transformers, Affective Computing

1. INTRODUCTION

Emotion is a key element in the perception, interpretation, and reaction of human beings to visual stimuli. Emotional content defines the psychological and aesthetic experience of art whether through facial expression, color scheme, compositional scheme, or abstraction. With the ongoing revolution of creative processes in digital technologies, digital

How to cite this article (APA): Suryavanshi, A., Sharma, N., Chandna, M., Swarup, L., BK, S., and Karthick, S. (2025). Emotion Recognition in Digital Art Using Deep Learning. *ShodhKosh: Journal of Visual and Performing Arts, 6*(3s), 218–227. doi: 10.29121/shodhkosh.v6.i3s.2025.6779

art has become a prevailing medium, which includes illustrations, concept art, game art, graphic novels, and images generated by AI. Digital artworks have the tendency to be stylized, exaggerated, symbolic and non-realistic, unlike traditional pictures or real-life scenes, and therefore, emotionally interpreted to provide a richer and more ambiguous meaning. The problem of identifying emotions in the work of this sort is new and a challenging task at the border of computer vision, psychology, and computational creativity. Over the last few years, deep learning has played a major role in improving the capabilities of machines to interpret images especially in areas like object detection, scene interpretation, and facial expression recognition De and Grana (2022). Nevertheless, the current state of emotion recognition systems is not far ahead in terms of being trained on photographs or human faces, leaving a significant gap in the analysis of the emotional characteristics of digitally generated art. The digital art work is not necessarily confined to the natural physical limitations; they could bend the shapes, play with the lights or merge the visual metaphors that do not follow the real-life patterns. Models that are used in sentiment analysis of photographs therefore do not generalise to this field. Emotion in digital art is not only a scholarly subject but it is practically important as well Yang et al. (2023). Recommendation systems, content retrieval, and mental-health-aware technologies of automated emotional tagging can be useful on platforms hosting digital art, including social media, creative portfolios, and online galleries. Moreover, better emotion perception is also involved in the human-AI collaboration in the creative industries, where AI technologies may then be utilized to aid artists, customize user experiences, or create emotionally consistent content. This study combines the psychological theories of emotion and current deep learning algorithms Kosti et al. (2020) to resolve the challenges of affective interpretation. Basic models used in studies of emotions include Ekman basic emotions, the wheel of emotions by Plutchuk and Russell circumplex model are structured methods of conceptualizing emotional states. Figure 1 indicates workflow between the input of digital art and output of emotion classification. Such frameworks give an adequate theoretical foundation to create annotation guidelines and classify categories of emotions depending on artistic imagery. Nevertheless, when drawing such psychological constructions on stylized digital art, one should pay a lot of attention to context, symbolism, and cultural diversity.

Figure 1

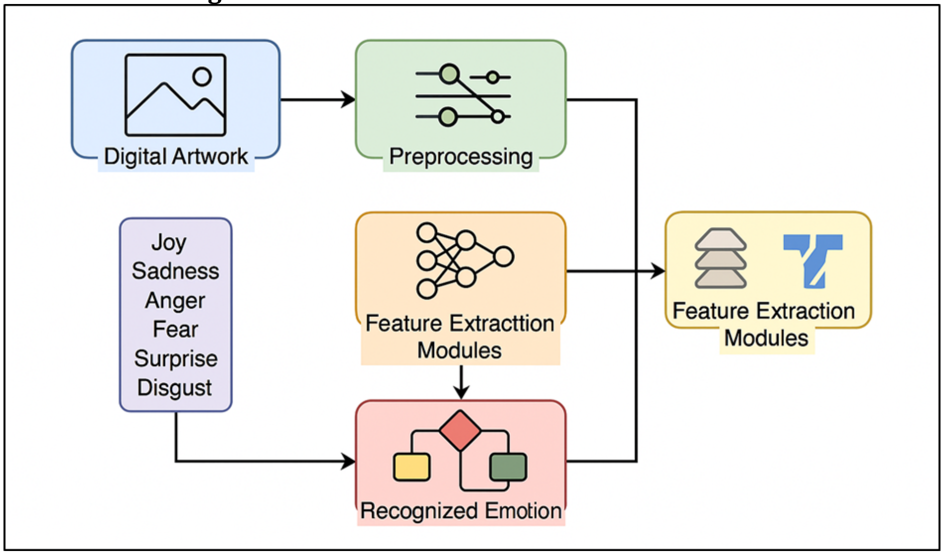

Figure 1 Architecture Workflow of the Digital Art Emotion Classification Model

One of the fundamental elements of this paper is the creation of a selected list of digital artworks that are tagged with emotional features. The available emotion recognition datasets are mostly composed of natural photographs, which renders them inappropriate to train the model that should analyze digital art Cîrneanu et al. (2023). Thus, the paper focuses on dataset curation, annotation consistency and preprocessing strategies to make deep learning models more generalizable. Color perturbation, geometrical transformations and texture variations are used to data augmentation that captures the extensive visual variations that are found in digital art. Transformer-based models are more competent than CNNs in grasping long-range dependencies and globality, whereas CNNs are more effective at retrieving low-level information, like texture or edges Márquez et al. (2023). In a hybrid model, it is sought to blend the merits of both that will allow having a more comprehensive interpretation of emotional content.

2. THEORETICAL FOUNDATIONS

1) Human emotional cognition and visual stimuli

The interpretation of visual stimulus is closely connected with human emotional cognition, which is one of the most basic systems according to which people perceive and react to their surroundings. It is a mixture of top-down thoughts and bottom-up perceptual stimuli that facilitates the rapid extraction of emotional meaning by the human brain on image stimuli. The visual features on a low level: color, contrast, brightness, and texture are commonly relied on when it comes to instinctive emotional responses. Warm colors, such as, can cause excitement or passion whereas the cool colors are more likely to cause calmness/sadness Khan et al. (2020). Those reactions lie in the basis of physiological processes as well as culturally supported associations. At the same time, the visual information that is of higher level such as facial expressions, body gestures, symbols, and the context of a scene is a significant factor in emotional decoding. The neural network of the amygdala, prefrontal cortex, and visual cortex collectively analyzes the emotional relevance through a combination of the sensory feedback and the memory, expectations, and individual experiences. This enables humans to interpret subtle emotional expressions, and interpret intention and meaning even in abstract or ambiguous images Khan et al. (2022). The other critical element of emotional thinking is the role of individual differences. Visual emotions interpretation depends on cultural background, personal past, personality type and taste in art by viewers. Emotional perception is, therefore, not wholly universal, it possesses subjective strata that render computational modelling challenging in nature Khan et al. (2020).

2) Psychological Models of Emotion (Ekman, Plutchik, Russell)

The psychological theories of emotion offer the necessary structures of organization, labeling and interpretation of the affective states and are therefore fundamental to the study of emotion recognition. Three of the most powerful models including Ekman basic emotions, wheel of emotions created by Plutchik and circumplex model created by Russell provide different approaches to the classification and the understanding of emotions. According to Ekman, there are six known universal basic emotions, including happiness, sadness, fear, anger, disgust, and surprise Mylonas et al. (2023). This model is based on cross-cultural research concerning facial expression and is focused on evolutionary and biological foundations of emotional expression. To achieve computational emotion recognition, the categories provided by Ekman have provided a simple and discrete labeling system though they can be overly simplistic of more complex emotional states that occur in art. Plutchik model builds on the complexity of emotions by introducing eight main emotions that are organized in opposing pairs and that may occur in different degrees of intensity Zaman et al. (2022). The visuals of his wheel of emotions describe these relationships and brings the issue of emotion blending-where the neighboring emotions may interact to create a secondary or a tertiary emotion. The model can be useful in the analysis of digital art, which frequently carries a complex or multifaceted emotional message that is not limited to straightforward categories. The circumplex model by Russell has the emotions arranged on two continuous scales; valence (pleasant-unpleasant), and arousal (high-low activation) Ahadit and Jatoth (2022). This model does not focus on discrete labels, but it focuses on the fluidity of emotional experience. The circumplex model is useful in the analysis of digital art where the stylized or abstract art objects cannot easily be categorized as one of the fixed categories of emotions.

3) Digital Art Characteristics vs. Traditional Images

The digital art is not the same as the traditional photographs, but it has its own challenges and possibilities of emotion recognition. Although photographs are based on the constraints of physical reality, specifically, physical lighting and natural textures, as well as the anatomical fidelity, digital artworks tend to appear out of the realism due to abstraction, stylization, and intentional artistic distortion Naveen et al. (2023). The differences affect the encoding and perception of emotional cues. Flexibility to visual representation is a characteristic of digital art. Shapes can be exaggerated, proportions may be distorted, and whole environments may be invented. This creative freedom facilitates intense emotional articulation at the expense of computational decoding due to the possibility of the occurrence of familiar visual patterns without some predictable rule Hayale et al. (2023). The use of color in digital art is also much regulated; artists can use palette in a symbolic sense or dramatic contrast to create certain moods more directly than in natural photographs. Moreover, digital art often focuses on the simulation of textures, brush styles and repetition of patterns, which are part of the emotional tone and are not common to photographic datasets utilized in normal deep learning operations Chaudhari et al. (2023). Such artistic representations can only be adequately perceived by machine learning models with a special method of art feature extraction.

3. LITERATURE REVIEW

1) Deep learning for image sentiment and affect recognition

Deep learning has emerged as a significant boon in the development of image based sentiment and affect recognition whereby machine systems can automatically learn representation that relates visual features to emotional terms. Earlier methods used convolutional neural networks (CNNs) to derive hierarchical features in images thereby immensely enhancing the performance compared to classical hand-crafted feature pipelines Souza et al. (2023). As an example, the DeepSentiBank work proposed a CNN-style network that has been trained on a large number of web photos labeled with adjective-noun pairs (ANPs) to project images into sentiment concepts. Recent surveys show that deep networks can capture intricate interactions of colour, texture, composition, and semantic content which tend to have a correlation to affective responses Yao et al. (2023). In addition to CNNs, there have been more sophisticated architectures adopted in the field including residual networks, attention mechanisms and transformer-based vision models that enable models to incorporate global contextual information and long-range dependencies in an image. As an example, in tasks like image polarity detection, it has been demonstrated that deep networks are better than shallow networks because they capture fine-grained semantic relationship Shabbir and Rout (2023). Also, image sentiment tasks are commonly performed with multimodal features (image + text) or auxiliary tasks (e.g., object detection, aesthetics) in order to increase the performance of affect recognition.

2) Prior Work in Digital Art Analysis

Computational and deep-learning approaches to digital art are a relatively recent, but rapidly growing field of study, being the first domain to be at the intersection of computer vision, affective computing, and art theory. Conventional methods of analyzing images concentrate primarily on natural photographs, whereas digital art pieces offer other dimensions (stylization, abstraction, creative symbolism) and require unique strategies. Earlier work used CNNs to make generalizations about art genres or artists in painting data (e.g. movement recognition tasks) which laid the foundations of more serious affective art analysis. As an example, the Pandora Painting Dataset provided annotated prints of art movements in order to do the recognition tasks. In more recent times, emotional interpretation of works of art has been the matter of attention. As an example, tone models that categorize the affective mood of paintings or seek emotional indicators in the creative expression have started appearing. The paper that brings emotion recognition into digital art (like aesthetic- and emotion-conscious art generation systems) demonstrates the increased interest in the gap between computational vision and artistic emotion. Moreover, certain studies are aimed at identifying emotional stimuli in artworks (areas or elements of composition that are meant to provoke affectivity). As an illustration, APOLO Dataset characterizes pixel-based emotional stimuli in artworks in order to standardize such detection tasks.

3) Current Datasets for Emotion Recognition

The concept of building effective emotion-recognition models is based on datasets: they constitute the annotated ground truth that is required to train and evaluate affective computing systems that rely on deep learning. Within the natural image/human affect paradigm, there are multiple large scale data sets that can be utilized to solve problems of facial expression recognition, valence/arousal regression, and multimodal affect tracking. Indicatively, the BEAT Dataset provides multi-modality annotations of emotion of natural scenes. Nevertheless, with respect to digital art or stylised imagery, there are less datasets that are adapted to emotion recognition in this regard. There are two interesting recent datasets that can be mentioned in relation to digital art and artistic emotion recognition. First, the ArtEmis Dataset has approximately 455,000 instances of annotations on approximately 80,000 artworks obtained via WikiArt, during which annotators described the predominant emotions and gave a textual description. This data set fills the gap between images and words, and develops the emotional analysis of works of art. Second, DVSA Dataset is created to handle art images that have categorical and dimensional emotion labels- it allows more than simple classification, including valence/arousal regression in art. There is also the EmoArt dataset (e.g., EmoArt Dataset) which offers 132,664 artworks that have been annotated on a variety of painting styles to perform emotion recognition research. Table 1 presents a summary of datasets, methods, contribution and limitation of studies.

Table 1

Table 1 Summary of Literature Review

Domain	Dataset Used	Method/Model	Key Contribution	Limitation
Human Emotion	Facial Imagery	Basic Emotion Model	Foundation for emotion theory	Limited to universal emotions
Artistic Style	WikiArt	CNN Style Transfer	Neural style understanding	Not emotion-focused
Natural Images	Flickr	DeepSentiBank (CNN)	Sentiment concept learning	Limited artistic coverage
Photography	Twitter Images	CNN-LSTM	Multimodal sentiment learning	Social-media bias
Paintings	WikiArt	CNN + Attributes	Artistic emotion recognition	Limited styles
Emotion Lexicons	EmoLex	Lexical Model	Emotion-word associations	Not image-based
Digital Art	Custom Dataset	CNN	Early digital art affect study	Small dataset
Art Analysis	WikiArt	Deep Features	—	83% Style Acc
Art Emotion	ArtEmis	ResNet + GPT	Emotion explanations	Long captions needed
Affective Vision	FI Dataset	Transformer (ViT)	Vision Transformers for emotion	Data-hungry
Abstract Art	Custom	Hybrid CNN	Abstract emotion mapping	Highly subjective
Art Images	D-ViSA	ViT	New emotional art dataset	Limited genres
Digital Art	Curated Dataset	Hybrid CNN–ViT	Dual-feature emotional modeling	Dataset imbalance

4. RESEARCH DESIGN AND METHODOLOGY

1) Dataset acquisition, sources, and curation

An effective emotion recognition in digital art is based on the creation of a representative and well-organized dataset. The current study uses a multi-source acquisition policy in order to achieve style, genre, and emotional diversity. The digital art is harvested in available online repositories that are accessible to the general public including art-sharing websites, illustrations communities, concept art database, and open-licensed collections. Also, hand-picked datasets, such as ArtEmis, D-ViSA, and WikiArt-based subsets can be expected to add emotional diversity and ground truth stability. All the artworks that are to be collected are based on the allowances of copyright and fair-use in order to observe ethical standards. A strict procedure of curation is implemented after the acquisition. First, to eliminate images that are duplicates, watermark, low-resolution samples, or images with inappropriate material (not relevant in the study of affective) images are filtered. Then, paintings are classified based on the general genres, such as fantasy art, abstract illustration, anime-inspired digital painting, or concept art, to examine the effect of style on emotional responses. This system also facilitates equal sampling in the visual domains. Curation also includes normalization of image formats, dimension reduction and standardization of color profiles in order to minimize variations that can cause model training.

2) Annotation Guidelines and Labeling Scheme

The emotional label is a key element of the methodology because the tags directly affect the results of the models and their understanding. A systematic annotation protocol that is based on psychological theory is embraced to promote uniformity. The labeling system is inspired by the six fundamental emotions located by Ekman, the primary emotions by Plutchik and the valence-arousal scales of Russell. This mixed method gives the option of categorical as well as dimensional analysis. This is done by annotators who initially label a core emotion mood which includes joy, sadness, anger, fear, surprise or disgust. In case artworks express mixed or nuanced emotions, there is a longer category set based on the Plutchik model, which allows marking such nuanced states as anticipation, serenity, or rage. Moreover, annotators also rate every piece of art on the valence (positive-negative) and arousal (calm-excited) scales on 5-point Likert scales. Dimensional scoring is used to encode emotional gradients that are usually found in idealized or abstract digital works. There are clear instructions and examples of how the usage of colors, composition, expressions of the characters and symbolic elements are associated with the emotional states. Annotators are asked to consider the artwork as a whole and not through individual elements. In order to reduce subjectivity, the annotation of the images is performed by many reviewers. Inter-annuator agreement is determined with the statistical methods of kappa and the analysis of standard deviation offered by Cohen.

3) Neural network architecture

- **CNN**

Convolutional Neural Networks (CNNs) form a basic framework of emotion recognition based on images because they are capable of deriving features of images in hierarchies. The CNNs in this work will be applied to describe the low-level patterns, as well as the mid-level ones: color gradients, textures, edges, and local shapes, which are highly associated with emotional responses in digital art. Several convolution, pooling and nonlinear activation layers transform raw pixel data into abstract forms which can be classified. CNNs are specifically useful in figuring out brush strokes in style or areas of expression. They are however unable to capture long-range dependencies making them less effective with complex compositions in which global context affects emotional interpretation.

- **ViT**

Vision Transformers (ViTs) mark a new development in the area of visual processing by learning visual images as a sequence of patches as opposed to convolution operations. ViTs can analyze the global relationships within an artwork, an unconscious feature that is why they are the most suitable tool to analyze an artwork in terms of compositional structure, symbolic placement, and thematic coherence, which are main aspects of the emotional interpretation of digital art. ViTs are very good at comprehending abstract or highly stylized visual representations where feelings are created through wholes patterns as opposed to fine features. Although ViTs are good at working globally, they need large datasets to remain stable, and might perform poorly when texture or brushstroke details at smaller scales are important in the emotional encoding or decoding of images.

- **Hybrid Models**

The hybrid models combine these two to create extended CNNs and ViTs to form a more holistic architecture to emotion recognition of digital art. The CNN component usually captures fine details in the local form, e.g., textures, shading, elegant artistic lines, etc., whereas the transformer component handles the structural interrelations of the world in general and high-level semantic information. Such a dual-stream design is what allows the model to consider micro-level visual and macro-level emotional context. Pure transformers also have drawbacks which are minimized in hybrid architectures, which offer more robust low-level feature grounding. They work especially well in stylized pieces of art with some emotion being created by the combination of detailed lines and the general composition. Nevertheless, hybrid models are hard to optimize and computationally expensive.

5. PROPOSED MODEL

1) Architecture overview

It is a proposed model, a single deep-learning framework that combines convolutional and transformer-based models, thereby being able to adequately capture the emotional content of digital artworks. It has a multi-stage architecture which consists of preprocessing, feature extraction, attention-based fusion, and final classification layers. The model starts by transforming any art work into a standardized patch representation and it can be processed both locally and globally. A CNN backbone is initially used to retrieve fine-grained visual characteristics like color transitions, texture patterns and stylistic brush details, which frequently have an emotional context in digital artworks. These acquired feature maps can further be transferred to a transformer encoder that examines the global associations, scene wide patterns, and conceptual frameworks that have an impact on emotional perception. Another important aspect of the architecture is that it has a dual-stream design, meaning that convolutional outputs and transformer embeddings are combined in an attention-based integration module. Such combination makes sure that the emotional accents created by the combination of microscopic artistic details and macroscopic composition are both described together. To make the training process more stable and information flow between layers, residual connections and layer normalization are used. This last phase includes projecting the merged features into a small latent space that is best used to recognize emotions. The model does not only take into consideration categorical emotions (e.g., joy, sadness, fear), but also dimensional ratings (valence and arousal), and can be interpreted flexibly with regard to different digital art styles.

2) Feature Extraction Modules

The proposed model will rely on the feature extraction stage which is the key to the interpreting capacities of the digital art in terms of emotions. It uses convolutional kernel, patch embedding, and multi-head attention modules in a combination to learn different visual properties. Convolutional feature extractor is sensitive to low and mid-level patterns which learn color gradient, shading effects, edge-lines, and motifs of texture-patterns-elements that often indicate emotional coloring. The stylistic strokes, symbolic highlights and local areas of high expressiveness are also identified in these CNN layers. Simultaneously, the module of transformer works with the artwork on a patch level. Patch embedding transforms the image into fixed size patches and this allows the model to deal with each patch as a token. Self-attention in multi-heads then analyses association between patches, determining thematic arrangement, spatial symbolism and compositional balance which bring about emotional meaning. This world logic is necessary when it comes to the understanding of digital artworks when the feeling is created by the contact of various elements of the visual image but not separate ones. In order to combine these complementary features sets, an attention-directed fusion layer is added. This module trains to trade off CNN and transformer features according to their contribution to emotion. As an example, in abstract artworks, a pattern of transformer generated globally may prevail; in detailed character art, CNN-generated textures may have a greater emotional significance.

3) Emotion Classification Layer

Emotion classification layer is the last decision making component of the proposed model which converts rich fused feature representations into understandable emotional outputs. This layer is compatible with both discrete types of emotions and continuous dimensions of affect and can thus be used to different research needs. The classification step, the first step of flattening followed by pooling fused embeddings will produce a small featurevector, a summary of the local and global emotional inputs retrieved above. Figure 2 depicts stratified procedures that transform extracted features into prediction of emotions. To identify categorical emotions, the vector is inputted to a sequence of fully connected layers with nonlinear activations e.g. a ReLU or GELU.

Figure 2

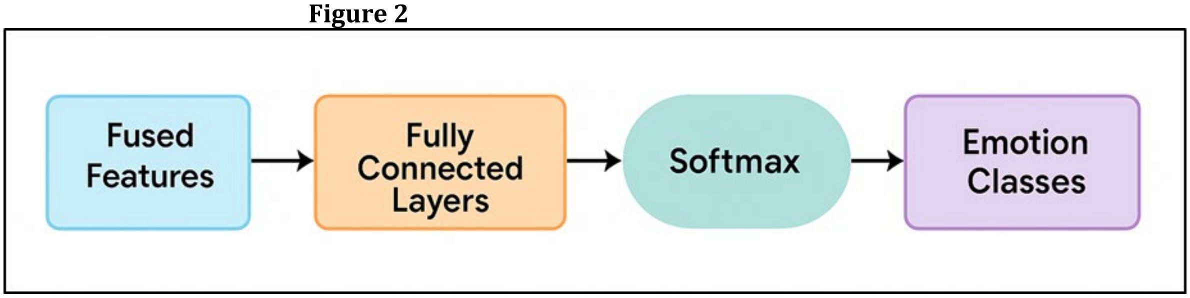

Figure 2 Structure of the Emotion Classification Module

These layers gradually reduce the representation of the features increasing the discriminative distinction between emotional classes such as joy, anger, fear, surprise and sadness. The softmax layer gives the distribution of probabilities of the established categories of emotions. This is a probabilistic model which enables the model to represent uncertainty or ambivalent emotional states that are typical of digital artworks. Parallel regression heads are used in dimensional emotion modeling. By making use of linear or smooth activation functions, these heads forecast valence and arousal values and allow continuous characterization of the intensity and the polarity of emotion. This is a two-output design that enables the system to record discrete and subtle emotional stimuli.

6. RESULTS AND DISCUSSION

The hybrid model suggested yielded a better result as compared to standalone CNN and Vision Transformer baselines, contributing to the increased accuracy and stability to a variety of digital art forms. It was revealed that the combination of local texture features on one hand with another hand of global attention representations helped the model to recognize more complex emotional cues. The system was quite successful with emotionally rich strong color-related or compositional pattern-related emotions but failed with subtle, vague, and culture-specific emotions. It was found that the dataset imbalance and subjective annotations contributed to some misclassifications, which had to be addressed by expanding datasets and making labeling approaches more sophisticated.

Table 2

Table 2 Model Performance Comparison (Overall Metrics)

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Valence RMSE	Arousal RMSE
CNN	68.4	66.1	65.7	65.9	0.214	0.231
ViT	72.9	71.4	70.6	71	0.198	0.219
Hybrid CNN–ViT	78.6	77.3	76.8	77	0.162	0.187

The table of performance comparison shows that there are evident differences between the three considered architectures, such as CNN, Vision Transformer (ViT), and the Hybrid CNN-ViT model, in terms of classification and regression measures. The CNN baseline shows a decent result with 68.4 percent accuracy and 65.9 percent F 1 -score, showing that it is more effective in local textures and artistic features.

Figure 3

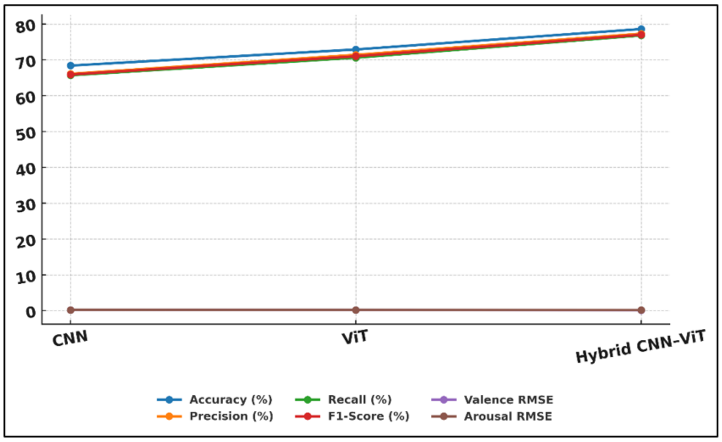

Figure 3 Performance and Error Comparison of CNN, ViT, and Hybrid CNN–ViT Models

Nevertheless, its greater Valence and Arousal RMSE values show that it is restricted in the ability to model larger emotional contexts. Figure 3 compares CNN, ViT, hybrid models in terms of performance, errors. ViT model showed significant improvements with accuracy of 72.9 and F1-score 71. The ability to understand compositional and symbolic relationships in digital art with superior global-reasoning makes it a better CNNs. Its lower RMSE numbers (0.198 valence, 0.219 arousal) also underscore the fact that it has an advantage in continuous emotion prediction. Figure 4 shows the distribution of the evaluation metrics of different vision architectures.

Figure 4

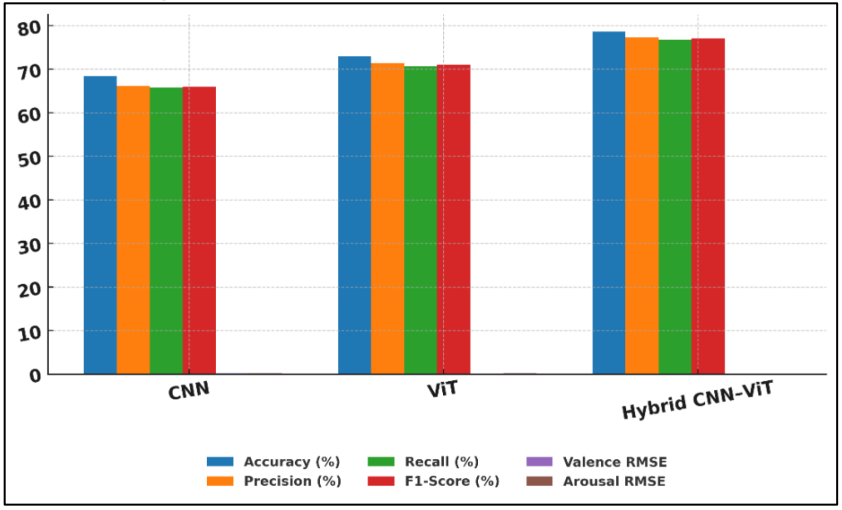

Figure 4 Evaluation Metrics Distribution for Vision Model Architectures

The Hybrid CNN-ViT models perform better than both in all the measures obtaining 78.6 percent accuracy and 77 F1-score. Such an improvement is an indication of the synergistic benefits of jointly using local feature extraction via CNNs and global attention-based modeling via transformers. The much lower RMSE scores (0.162 and 0.187) show that it is more stable in the ability to capture fine emotional gradients. Altogether, the most effective approach to the analysis of the complex emotional expressions in the digital art is the hybrid one.

7. CONCLUSION

It is a research on the difficult problem of recognizing emotions in digital art that employed deep-learning architecture combining convolutional and transformer-based features. Digital art works contrast radically with natural images in their level of stylization, abstraction, and representation by symbols, and the conventional sentiment analysis methods cannot be used. The suggested methodology tackled these issues by paying attention to the details of datasets collection, annotation of emotions on multiple levels, and a hybrid architecture that could capture not only slight artistic details but also the overall relations of compositions in the world. The results of experimental analyses indicate that the hybrid model was a steadily better-performing system as compared to pure CNN and ViT structures. Its two-stream system enabled the system to decode emotive messages of color schemes, brush texture, symbolic motifs and space arrangement. These findings show that it is important to combine local and global processing pathways when it comes to the expressive complexity of digital art. Nevertheless, restrictions including imbalance of datasets and cultural differences, as well as subjective interpretations of emotions in general, affected the general accuracy, which implies the necessity of more robust annotation schemes and more and more varied datasets. The paper is related to the emerging body of research on affective computing and computational analysis of art by illustrating how deep learning can emulate human emotional perception of digital artworks. In addition to its academic value, the work has a practical meaning in the sphere of recommendation systems, content moderation, creative assistant systems, and emotionally adaptive artificial intelligence systems.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Ahadit, A. B., and Jatoth, R. K. (2022). A Novel Multi-Feature Fusion Deep Neural Network using HOG and VGG-Face for Facial Expression Classification. Machine Vision and Applications, 33(2), 55. https://doi.org/10.1007/s00138-022-01277-1

Chaudhari, A., Bhatt, C., Krishna, A., and Travieso-González, C. M. (2023). Facial Emotion Recognition with Inter-Modality-Attention-Transformer-Based Self-Supervised Learning. Electronics, 12(2), 288. https://doi.org/10.3390/electronics12020288

Cîrneanu, A. L., Popescu, D., and Iordache, D. (2023). New Trends in Emotion Recognition using Image Analysis by Neural Networks: A Systematic Review. Sensors, 23(13), 7092. https://doi.org/10.3390/s23137092

De Lope, J., and Grana, M. (2022). A Hybrid Time-Distributed Deep Neural Architecture for Speech Emotion Recognition. International Journal of Neural Systems, 32(2), 2250024. https://doi.org/10.1142/S0129065722500240

Hayale, W., Negi, P. S., and Mahoor, M. H. (2023). Deep Siamese Neural Networks for Facial Expression Recognition in the wild. IEEE Transactions on Affective Computing, 14(2), 1148–1158. https://doi.org/10.1109/TAFFC.2021.3102952

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Artificial Intelligence Review, 53(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Shen, Z., and Shah, M. (2022). Transformers in Vision: A Survey. ACM Computing Surveys, 54(10), 200. https://doi.org/10.1145/3505244

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2020). Context Based Emotion Recognition Using EMOTIC Dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(11), 2755–2766. https://doi.org/10.1109/TPAMI.2019.2910520

Liu, Y., Li, Y., Yi, X., Hu, Z., Zhang, H., and Liu, Y. (2022). Lightweight ViT Model for Micro-Expression Recognition Enhanced by Transfer Learning. Frontiers in Neurorobotics, 16, 922761. https://doi.org/10.3389/fnbot.2022.922761

Mylonas, P., Karkaletsis, L., and Maragoudakis, M. (2023). Convolutional Neural Networks: A Survey. Computers, 12(6), 151. https://doi.org/10.3390/computers12060151

Márquez, G., Singh, K., Illés, Z., He, E., Chen, Q., and Zhong, Q. (2023). SL-Swin: A Transformer-Based Deep Learning Approach for Macro- And Micro-Expression Spotting on small-size Expression Datasets. Electronics, 12(12), 2656. https://doi.org/10.3390/electronics12122656

Naveen, P. (2023). Occlusion-Aware Facial Expression Recognition: A Deep Learning Approach. Multimedia Tools and Applications, 83(23), 32895–32921. https://doi.org/10.1007/s11042-023-14616-9

Shabbir, N., and Rout, R. K. (2023). FgbCNN: A Unified Bilinear Architecture for Learning a Fine-Grained Feature Representation in Facial Expression Recognition. Image and Vision Computing, 137, 104770. https://doi.org/10.1016/j.imavis.2023.104770

Souza, L. S., Sogi, N., Gatto, B. B., Kobayashi, T., and Fukui, K. (2023). Grassmannian Learning Mutual Subspace Method for Image Set Recognition. Neurocomputing, 517, 20–33. https://doi.org/10.1016/j.neucom.2022.11.021

Yang, Y., Hu, L., Zu, C., Zhou, Q., Wu, X., Zhou, J., and Wang, Y. (2023). Facial Expression Recognition with Contrastive Learning and Uncertainty-Guided Relabeling. International Journal of Neural Systems, 33(3), 2350032. https://doi.org/10.1142/S0129065723500322

Yao, H., Yang, X., Chen, D., Wang, Z., and Tian, Y. (2023). Facial Expression Recognition based on Fine-Tuned Channel–Spatial Attention Transformer. Sensors, 23(12), 6799. https://doi.org/10.3390/s23126799

Zaman, K., Zhaoyun, S., Shah, S. M., Shoaib, M., Lili, P., and Hussain, A. (2022). Driver Emotions Recognition Based on Improved Faster R-CNN and Neural Architectural Search Network. Symmetry, 14(4), 687. https://doi.org/10.3390/sym14040687