Baghdad Science Journal

Volume 23 | Issue 5 Article 17

5-26-2026

HS-AMM: Hierarchical Spatial Attention Mechanism Model for Facial Expression Recognition based on Deep Learning

Ahmed Oday
Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia AND Department of Bioinformatics, College of Biomedical Informatics, University of Information Technology & Communications, 10001 Baghdad, Iraq, p109064@siswa.ukm.edu.my

Azizi Abdullah
Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia, azizia@ukm.edu.my

Shahnorbanun Sahran
Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia, shahnorbanun@ukm.edu.my

How to Cite this Article

Oday, Ahmed; Abdullah, Azizi; and Sahran, Shahnorbanun (2026) "HS-AMM: Hierarchical Spatial Attention Mechanism Model for Facial Expression Recognition based on Deep Learning," *Baghdad Science Journal*: Vol. 23: Iss. 5, Article 17.
DOI: https://doi.org/10.21123/2411-7986.5301

BAGHDAD SCIENCE JOURNAL 2026;23(5):1795–1807

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

HS-AMM: Hierarchical Spatial Attention Mechanism Model for Facial Expression Recognition based on Deep Learning

Ahmed Oday [1,2,*], **Azizi Abdullah** [1], **Shahnorbanun Sahran** [1]

[1] Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Malaysia
[2] Department of Bioinformatics, College of Biomedical Informatics, University of Information Technology & Communications, 10001 Baghdad, Iraq

ABSTRACT

Facial expression recognition (FER) is a significant challenge due to the complexity of multi-class categorization and the subtle differences between emotional factors, with existing deep learning models often struggling to leverage hierarchical local features for accurate distinction. To find a solution, this study proposes a two-method approach, aiming to enhance FER accuracy by refining classification through the arousal-valence dimension and improving feature extraction via an adaptive attention mechanism to resolve the difficulty of differentiating semantically close expressions. The proposed methodology combines (1) a two-level classification strategy, first, performing fine-grained classification, and second, multi-class categorization by grouping expressions into neutral, positive, and negative classes based on emotional dimensions. (2) a hierarchical attention model that mitigates the challenge of fine-grained expression recognition by stacking global, regional, and local attention in a redesigned spatial hierarchy, first learning a robust semantic grouping (neutral, positive, and negative) of expressions. This approach reduces high-level confusion and provides a clearer pathway for subsequent fine-grained classification. This module extract emotion feature from facial images, all implemented on an enhanced VGG-16 backbone. The experiments on CK+, FER+, AffectNet, and RAF-basic datasets demonstrate that the proposed model achieves superior accuracy and robustness compared to existing methods, effectively capturing multi-level features and leveraging hierarchical subdivision for improved performance. The results confirm that our approach, through its fusion of multi-class grouping and hierarchical spatial attention, offers a more reliable and sophisticated solution for real-world emotion recognition applications.

Keywords: Facial expression recognition, Facial classification, Hierarchical attention, Multi-class, Spatial attention

Introduction

Facial expression recognition (FER) has gained significant interest in artificial intelligence and computer vision, largely owing to its application in human–computer interaction. FER is important for developing empathetic and adaptive technologies because it enables machines to understand and respond to the emotional states and social interactions of humans through facial cues, gesture and speech. [1] Unlike conventional FER systems that use hand-engineered features along with standard classifiers, the deep-learning revolution completely redefines this domain by providing end-to-end learning of features on multiple levels of abstraction from unprocessed images. [2] CNNs first came into the limelight after their efficacy was highlighted by AlexNet through rectified linear unit (ReLU) activations and dropout, a new feature at the time. The prominence of small convolutional filters was highlighted by deep yet simple

Received 19 April 2025; revised 17 December 2025; accepted 2 January 2026.
Available online 26 May 2026

* Corresponding author.
E-mail addresses: p109064@siswa.ukm.edu.my (A. Oday), azizia@ukm.edu.my (A. Abdullah), shahnorbanun@ukm.edu.my (S. Sahran).

https://doi.org/10.21123/2411-7986.5301

VGG-16 networks. These models demonstrate how modern architecture improves performance, effectiveness and complexity. The performance of FER models has significantly improved with the addition of new attention-based features as aids to the hierarchical structure. These features help the network concentrate on important facial areas, such as the eyes, mouth and eyebrows, which aid in emotion recognition.[3] The spatial attention mechanism redistributes weights across locations in the image, allowing the model to focus on key regions. This process involves identifying critical areas, such as objects or landmarks, to improve the prediction accuracy for these areas.[4]

Hierarchical models capture the local dependencies in more than one spatial region, increasing the robustness to pose and occlusion.[5] In emotion recognition, these hierarchical attention mechanisms significantly improve the speed and accuracy of FER systems by combining multi-scale spatial attention with temporal dynamics, thereby enabling state-of-the-art real-time applications. One method uses the emotional dimensions of arousal and valence models to express emotions.[6,7] Arousal refers to the amount of autonomic activation elicited by a stimulus, which can vary from low to high. Valence, in contrast, measures the amount of pleasure an experience offers and is defined on a spectrum ranging from negative to positive. These two measures can be fused to create a multidimensional space in which any given point can represent a different type of emotion.[8] Fig. 1 illustrates this concept by combining arousal and valence to provide a dual-axis representation of various emotions. The facial expressions into psychologically rooted categories extend beyond a methodological choice, offering significant practical advantages for downstream affective computing applications. Firstly, this structure introduces a crucial mechanism for graceful interaction in real-world human-computer interaction. In ambiguous conditions where distinguishing between specific, visually similar emotions like fear and surprise is challenging, the system can default to the reliable coarse-level classification (e.g., High-Arousal) to initiate an appropriate general response, thereby avoiding critical errors and enhancing user trust. Secondly, the output provides a more actionable and interpretable signal for adaptive systems. A platform can first react to a negative valence by offering support and then leverage the fine-grained prediction to respond, such as providing simplifying material for confusion. Finally, this grouping is invaluable for data-efficient modelling in emotion-aware recommendation systems. It allows for the generalization of user preferences across emotionally similar states (e.g., pooling 'Sad' and 'Disgust' under Negative), effectively mitigating the data sparsity problem for rare emotions and providing a robust prior for building more effective personalization models.[8]

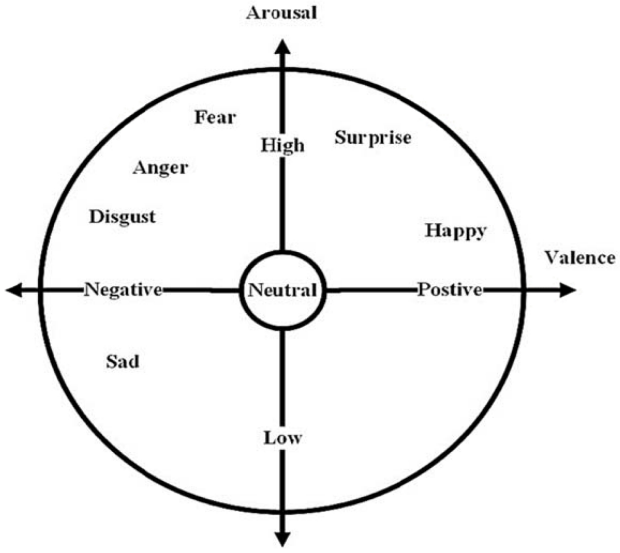

Fig. 1. Arousal and valence dimensions.[8] Arousal refers to the level of autonomic activation triggered by a certain situation, which can be low or high. A certain experience can produce valence, which is rated on a negative to positive scale. These two metrics can create a high-dimensional space, where each point represents a different emotion. This figure illustrates a dual-axis representation of various emotions by combining arousal and valence.

Deep learning simplifies these problems using convolutional neural networks (CNNs), which extract features hierarchically. AlexNet,[9] ResNet-18,[10] and VGG-16[11] are architectural benchmarks that modify CNNs, each performing something new and significant for deep learning. Furthermore, FER models tend to experience problems with generalisation due to overfitting in trained networks; they require more realistic and inclusive benchmarks from which to train. Conventional CNNs, such as VGG16, depend on convolutional layers with uniform kernel sizes, leading to restricted receptive fields. This local focus hinders the network's ability to capture global contextual information and high-level semantic features, resulting in inadequate feature extraction and reduced learning efficiency. Uniform stacking of layers fails to establish long-range dependencies between different regions of the image, which is critical for accurately recognising subtle and complex facial expressions. Additionally, existing attention mechanisms, such as spatial attention, often focus on local regions and struggle to model distant dependencies. To address these issues, this study proposes HS-AMM, a hierarchical spatial attention mechanism integrated into VGG-16. Unlike traditional methods that process features uniformly,[11] HS-AMM stacking combines global,

regional, and local attention to expand the receptive field and prioritize emotion-salient facial regions. This multi-scale approach enriches feature diversity by capturing emotion expressions and subtle expressions, while the reinforced VGG-16 backbone ensures robust hierarchical feature fusion. As a result, HS-AMM improves emotion classification accuracy on the Dataset used, particularly for challenging cases like low-intensity.

The main contributions of this study are as follows:

1. Proposing a hierarchical spatial attention mechanism model (HS-AMM) for multi-class FER that captures and aggregates local, regional and global hierarchical attention features and dependencies of the spatial attention mechanism architecture.
2. Designing multi-level classification approaches: in Level 1, each of the classes is processed at the same level, and in Level 2, the three labels—negative, positive, and neutral according to the valence axis—are processed at a multi-level using hierarchical attention. Experimentally establishing the importance of hierarchical modelling in improving FER performance. Comprehensively evaluating the proposed method (HS-AMM) using various datasets, such as CK+, FER+, AffectNet and RAF-DB, and compared it with the standard CNN (VGG16) baseline model. Experiments demonstrate that the proposed method outperforms other methods in various classifications.

Related work

FER has witnessed a surge in research interest, driven by advances in machine learning and computer vision. Related works can be categorised into deep-learning-based models and hierarchical attention mechanisms. This section analyses these categories by focusing on their contributions, limitations and current state-of-the-art models.

Deep-learning-based models

Deep learning has revolutionised FER by introducing end-to-end trainable models. CNNs have become the backbone of FER systems, as they can automatically learn hierarchical feature representations from raw image data. Notable architectures include VGG16, ResNet and Inception, which have been adapted for FER tasks.[12] The integration of transfer learning further enhances the FER performance; models pre-trained on large-scale datasets, such as ImageNet, provide a strong initialisation, enabling fine-tuning on smaller FER-specific datasets. Transfer learning mitigates the challenges posed by limited labelled data and accelerates convergence.[13] Despite their success, standard CNNs face challenges in capturing global dependencies and contextual relationships. To address these issues, attention mechanisms have been introduced that enable models to focus on critical facial regions, such as the eyes, mouth and eyebrows.[1]

The hierarchical architecture,[3] characterised by sequential 3×3 convolutional blocks with batch normalisation and ReLU, has been widely adopted for facial feature extraction owing to its capacity to capture intricate spatial patterns. Recent studies have integrated channel attention mechanisms with VGG-16 to address local receptive field limitations, enhance global inter-channel dependencies, and prioritise discriminative emotion-specific features (e.g. eyes and mouth) while suppressing noise, thereby improving the recognition accuracy. LeCun[14] constructed the LeNet architecture using CNNs by implementing the backpropagation method to train the network for feature extraction and recognition. Krizhevsky et al.[15] developed the AlexNet model, building upon LeNet but increasing the depth of the network to improve the output features. Thereafter, Simonyan and Zisserman[16] presented the VGG model, which improved upon the AlexNet architecture; it replaced the 11×11 convolution kernels with 3×3 ones, and the 5×5 kernels with 3×3 ones, to boost the network performance. Although CNNs, such as AlexNet and VGG, improve feature extraction via hierarchical layers and smaller kernels, their reliance on local receptive fields limits the global context capture, necessitating deep stacking for distant dependencies.

Hierarchical attention mechanisms

Advances in robotic manipulation have been driven by reinforcement learning and attention-based frameworks.[17] Traditional approaches depend on predefined object models, limiting flexibility in dynamic environments. Learning manipulation and modern reinforcement-learning techniques, including deep Q-networks and actor-critic methods, have enabled learning directly from visual input, improving adaptability and decision-making in complex tasks. Learning manipulation attention mechanisms have further enhanced these models by focusing on task-relevant features, effectively reducing computational complexity in high-dimensional action spaces. Therefore, the hierarchical spatial attention framework proposed by Liu et al.[18] extends these advances by integrating multi-level attention hierarchies, thereby enabling efficient and generalisable learning for

challenging robotic manipulation tasks. However, it often struggles with large observation-action spaces, leading to sampling inefficiency. Attention-based approaches address this by focusing on task-relevant regions and improving learning efficiency and generalisability. Building on this, structured attention constraints have been introduced to significantly reduce sample complexity and improve effectiveness in real-world scenarios.[19] Hierarchical attention mechanisms represent a significant advance in FER, enabling models to emulate human visual attention and focus on discriminative regions while ignoring irrelevant background information. Feng et al.[20] combined multi-scale spatial attention with temporal dynamics; the hierarchical attention systems achieved state-of-the-art performance but at a high computational cost due to the transformer architecture. The generalisability to real-world low-resolution LiDAR scenarios, diverse environments and adverse conditions remains unverified. Real-time performance and robustness in dynamic, noisy and resource-constrained settings have not been thoroughly explored. The hierarchical channel attention-CNN improves FER by employing hierarchical attention layers over the basic VGG16 architecture. This combination extends the receptive field and enriches feature representation, ensuring the effective capture of fine-grained details. The ability of the model to integrate local and global features proved effective in handling variations in pose and occlusion, improving the accuracy on benchmark dataset CK+.[3] Another notable contribution is the hierarchical cross-attention model that integrates cross-modal information from speech and facial expressions. The model aligns audio and visual data by employing co-attention layers, enabling robust emotion recognition, even in noisy environments.[5] FER faces the following persistent challenges: The scarcity of diverse annotated datasets limits the generalisation of FER models. Most datasets are biased towards specific demographics, leading to reduced performance in real-world applications.[21] Moreover, achieving real-time performance while maintaining accuracy is critical for deploying FER in applications such as surveillance and human–computer interaction.[22] Furthermore, the development of lightweight architectures optimised for edge devices can enable real-time FER in resource-constrained environments.[23]

Global attention

The global-attention mechanism (GAM) is a new method integrated into the CNN systems to enhance performance through global interaction between the channels and spatial parts of the image. Compared with earlier approaches, GAM does not use pooling in the spatial attention component, which diminishes the amount of information captured, and uses group convolution for feature extraction. The channel attention module uses a three-dimensional permutation with a multi-layer perceptron to handle multi-dimensional data and construct dependencies among different dimensions. GAM outperformed the existing models in terms of accuracy on the CIFAR-100 and ImageNet-1K benchmarks. Its design guarantees strong global feature representation, while offering parameter efficiency owing to group convolutions, enabling the architecture to scale across both complex and lightweight structures, as well as capturing a wide range of contextual relations.[24]

Regional attention

The regional attention network employs a novel and efficient attention mechanism for gesture- and action-recognition tasks. It assigns weights to regional features with self- and co-attention, emphasising important regions and suppressing irrelevant ones. In contrast with conventional techniques that depend on explicit markings, the regional attention network uses squeeze-and-excitation layers integrated with skip connections to improve feature representation. The model was tested on 10 datasets, and it outperformed current state-of-the-art methods in head pose estimation, driver state recognition and human action classification tasks, focusing on mid-to-fine details.[25]

Local attention

The local attention mechanism (LAM) introduces a novel approach for long-sequence time-series forecasting by emphasising the temporal locality. Unlike full attention mechanisms with quadratic complexity, LAM efficiently reduces the computational overhead to $\theta(n \log n)$ using an additive mask to limit the attention to the nearby inputs, where the mean theta and n are inputs. LAM focuses on local dependencies to capture short-term patterns while preserving computational efficiency. The experimental results demonstrate that LAM, integrated with the vanilla transformer architecture, outperforms state-of-the-art models such as Informer and Log Sparse in multivariate predictions, particularly for long prediction horizons. Its superior efficiency and accuracy make LAM a robust solution to complex time-series forecasting challenges.[26]

Despite advances in hierarchical emotion classification,[19,20] there is an unresolved issue in the fine-grained distinction of semantically close expressions within emotion groups (e.g., fear vs.

anger in 'high-arousal negative'). Current VGG-based models[11] lack specialized attention to differentiate subtle cues; anger typically contracts eyebrows vertically, while fear raises them horizontally,[8] leading to misclassification rates exceeding 35% for these pairs on RAF-DB. This propagates errors to the grouping stage, degrading system-level reliability. Our HS-AMM tackles this by including VGG16, which employs a deep architecture with 13 convolutional layers, small 3 × 3 filters, and pooling layers to extract hierarchical spatial attention features to improve parameter utilisation.

In contrast to previous research that relies on flat emotion classification using conventional CNN architectures, this work integrates an enhanced attention module into the VGG backbone and restructures the output space into a hierarchical form. The proposed attention mechanism aims to improve the discriminative capability of feature maps at multiple abstraction levels, allowing the model to focus adaptively on the most relevant facial regions. Moreover, the hierarchical output design groups emotions according to the dimensional emotion model (arousal and valence dimensions), enabling the system to produce not only discrete emotion labels but also a more comprehensive representation of the underlying affective state. This hybrid design—combining a modified attention mechanism, a hierarchical structure, and emotion dimensional grouping—provides both improved recognition accuracy and deeper interpretability, representing the main novelty and contribution of this study.

Proposed method

This section presents a new approach for improving the accuracy and efficiency of FER using deep-learning techniques. The proposed Hierarchical Spatial Attention Mechanism Model (HS-AMM) is used to analyse facial features with high accuracy. The proposed methodology was designed to ensure reliable results and fast performance. The approach combines three parts: first, CNN-VGG-16 for feature extraction; second, the proposed model for feature classification; and third, output levels as a hierarchical combination of two class levels as in Fig. 2.

First, feature maps are extracted from the input facial images using the CNN-VGG-16 backbone network, which extracts features to be ready for hierarchical features, and subsequently processed through the HS-AMM framework, which consists of global, regional, and local attention modules.

Second, this architecture enables the adaptive processing of expression characteristics at different granularities, improved feature discriminability through spatial attention refinement, and robust performance across stacking varying attention mechanism intensities. The attention maps are concatenated to fuse multi-level features sequentially, starting with average- and max-pooling operations to preserve both discriminative and salient spatial information. The multi-level attention maps retain the complementary information. Pooling reduces dimensionality while preserving spatial hierarchy attention. The model prioritises critical visual patterns across scales,

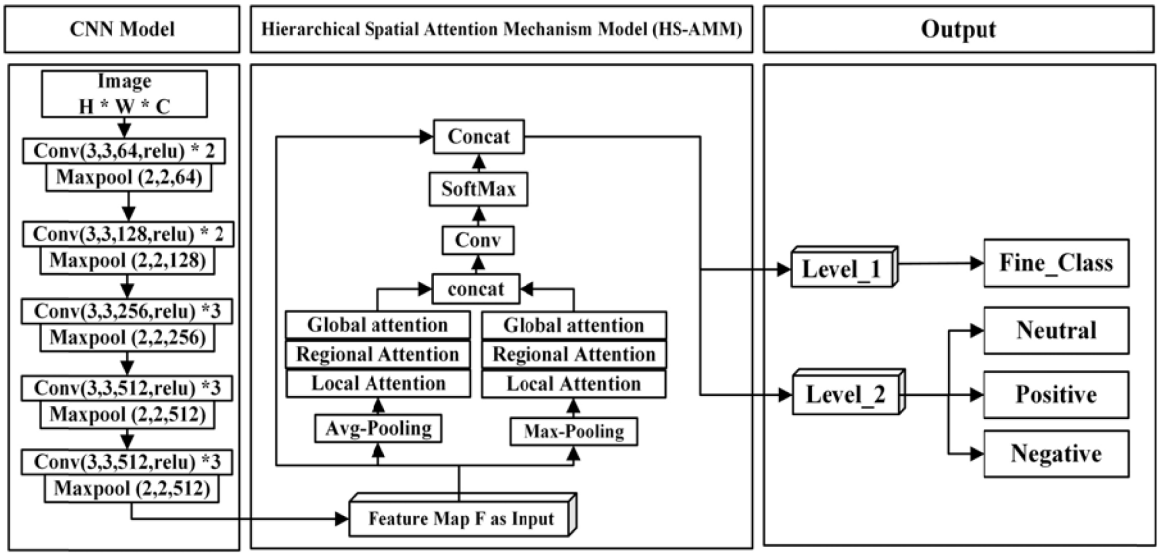

Fig. 2. Proposed hierarchical spatial attention mechanism model (HS-AMM). CNN model: A deep convolutional neural network constructed on a VGG-16 backbone architecture designed for image classification tasks. Dimensional images were used to extract multi-scale feature maps. HS-AMM: The feature maps are hierarchically refined using global, regional, and local-attention modules to emphasise spatially relevant regions at different granularities. Output levels: Level 1 fine classifies the output into seven classes, according to the input dataset. Level 2 focuses on arousal and valence dimension classification by integrating the multi-class.

enhancing interpretability and robustness in complex image-recognition scenarios. Typically, it consists of a combined feature representation derived from various stacked attention mechanisms, such as local, regional, and global attention. Each of these mechanisms processes the input data differently to capture unique aspects. Base feature extraction is shown in Eq. (1), where the input facial image X is first processed by the VGG-16 network to generate a foundational feature map F, which encapsulates the high-level spatial features essential for expression recognition.

$$F\left[b, c, i, j\right] = \text{VGG16}\,(X)\left[b, c, i, j\right] \text{ for } b \in \text{Batch}, \quad c \in \{1, \ldots, 512\}, i, j \in \{1, \ldots, 14\} \tag{1}$$

Attention mechanisms are executed in the model via local, regional, and global attention modules, which focus on higher levels of feature extraction. For certain regions, local attention feature extraction allocates weights to a few features, provided that the detailed features are important, as shown in Eqs. (2) and (3). The feature map F. For each spatial location i, a learnable weight α is computed to signify its importance, allowing the model to emphasize critical details while suppressing irrelevant information.

$$\alpha\left[b, i\right] = \text{softmax}\left(W_{\text{local}} \times F\left[b, i, :\right] + b_{\text{local}}\right), \tag{2}$$

where $\alpha[b, i]$ is the learned attention weight, W_{local} and b_{local} are learnable parameters.

$$\text{Local}\left[b, d\right] = \sum_{i=0}^{195} \alpha\left[b, i\right] \times F\left[b, i, d\right], \tag{3}$$

where F[b, i, d] is the learnable weight matrix. Regional attention extends this to broader spatial regions by computing the relationships between features using a query-key-value mechanism, as shown in Eq. (4). We employ a self-attention mechanism. This stage calculates the interdependencies between different feature regions, allowing the model to contextually group local features. This is implemented via a query-key-value (Q, K, V) mechanism applied to the local features.

$$\text{Regional Attention}\left[b, 1, d\right] = \sum_{h=0}^{511} \text{Softmax}\left(\frac{Q\left[b, 1, h\right] \times K[b, 1, h]}{\sqrt{512}}\right) \times V\left[b, 1, d\right] \tag{4}$$

where $Q = W_Q\text{Local}$, $K = W_K\text{Local}$ and $V = W_V\text{Local}$, which allows the network to capture the interdependencies between different regions. Global attention operates across the entire input space, aggregating information to capture the global patterns. The model employs a convolutional layer for spatial scaling and feature transformation. This layer refines the combined features from attention mechanisms and prepares them as prediction layers. Convolution is expressed as Eqs. (5) to (7). The final stage integrates all regional information to form a coherent global context. This is achieved through multi-head attention, which allows the model to attend to information from different representation subspaces simultaneously. The output is a comprehensive feature vector that encapsulates the entire emotional context. This involves a multi-perspective feature representation, which concatenates the max-pooled and average-pooled versions of the global context. This operation captures both the most active features and the overarching emotional signals.

$$\text{Global Attention}\left[b, d\right] = \frac{1}{4}\sum_{h=1}^{4} \text{MultiHead}_{\,h}\left(\text{Regional Attention}\right)\left[b, d\right] \tag{5}$$

$$Z\left[b, d\right] = [\text{Global Attention}_{\,\max} \,||\, \text{Global Attention}_{\,\text{avg}}] \times [b, d] \;\text{(concatenation)}, \tag{6}$$

$$\text{ConvOut}\left[b, c, i, j\right] = \text{LeakyReLU}\left(\sum_{m,n} W_{\text{conv}}\,[c, m, n] \times Z\left[b, 1, i + m, j + n\right]\right), \tag{7}$$

where W and b are learnable weights and biases, respectively.

The integration of these spatial layers into the hierarchical facial expression model provides several advantages. Pooling layers reduce the size of feature maps and improve the computational efficiency of the model while preserving critical information. Focused attention mechanisms prioritise important regions, thereby enhancing the ability of the model to capture nuanced expressions. The combination of local, regional, and global attention enables multi-scale feature analysis. Convolutional layers, in conjunction with pooling and attention layers, ensure that the model generalises well to varying input conditions. These spatial layers collectively enable the model to achieve high accuracy in emotion detection by leveraging hierarchical and adaptive feature extraction.

Third, structured methods assist in the classification by decreasing the errors typically made in highly

Table 1. Proposed Layer of hierarchical spatial attention mechanism model (HS-AMM).

Module	Output Shape (approx.)	Trainable Params
VGG16 (features)	[B, 512, 14, 14]	14,714,688
Local Attention (m)	[B, 512]	513
Regional Attention (m)	[B, 1, 512]	787,968
Global Attention (m)	[B, 512]	1,050,624
Local Attention (a)	[B, 512]	513
Regional Attention (a)	[B, 1, 512]	787,968
Global Attention (a)	[B, 512]	1,050,624
Conv After Combined	[B, 8, 32, 32]	88
Level 1 Linear	[B, 3]	24,579
Level 2 ModuleDict	[B, variable per branch]	57,351
ChannelPoolmax	[B, 1, . . .]	0
ChannelPoolavg	[B, 1, . . .]	0
Spatial Conv	[B, 1, H, W]	11
Total (reported)		18,474,927

detailed distinctions. This approach allows Level 1 to handle specifics, whereas Level 2 classifies the data into coarse categories. Such a structure can recognise facial expressions because the classes are closely related to each other at the lower level, but are separate at a higher level. This prediction structure resembles how people classify these features, making the model predictions more explanatory and accurate. The input and output processes of levels 1 and 2 separately render the system highly effective for handling complex classification tasks. The output of Level 2 has 'negative', 'positive' and 'neutral' as the predefined high-level categories. These categories are determined based on arousal and valence measures within the input data. The output is generated by a fully connected layer with the Softmax activation function, expressed as Eq. (8)

$$Y_{level1} = \text{Softmax}\ (W_1 \times \text{ConvOut} + b_1), \tag{8}$$

where W_1 and b_1 are trainable weights and biases of the Level 1 layer, respectively.

Level 2 branches out from the input but divides it into multiple sub-classifications based on a pre-defined branch corresponding to a high-level category from Level 1, and is further split into subcategories. For example, 'Happy' or 'Surprised' predicted at Level 1 would be further classified into 'Positive'. The output for a specific branch k in Level 2 is given by Eq. (9)

$$Y_{level2,k} = \text{Softmax}\left(W_{2,k} \times \text{ConvOut} + b_{2,k}\right), \tag{9}$$

where $W_{2,k}$ and $b_{2,k}$ are weights and biases specific to the k-th branch, respectively. Table 1 explains the layers of the proposed model, the parameters for each layer, the sequence of connections, and zero Non-trainable Params layers.

Materials and dataset

Metrics

The proposed model's performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. These metrics were chosen because they provide a balanced assessment of both classification correctness and error distribution, which is crucial for evaluating facial expression recognition.[27]

Dataset

The CK+[28] is often cited in the literature as a benchmark for facial emotion-recognition tasks and consists of 593 image sequences collected from 123 subjects. The emotions are anger, contempt, disgust, fear, happiness, sadness, and surprise. CK+ consists of high-quality annotated facial expressions, making it ideal for the training and evaluation of emotion-recognition models in a controlled environment. CK+ is a benchmark dataset with posed expressions. Performance has saturated in the literature, so our analysis focuses on the more challenging in-the-wild datasets (FER, AffectNet, and RAF-DB).

The FER+[29] builds upon the original FER dataset by re-labelling certain images based on the emotion categories through crowdsourcing. It consists of 35,887 facial images grouped into eight emotions: anger, disgust, fear, happiness, neutrality, sadness, surprise and contempt. In FER+, the quality of emotion-recognition tasks is improved by correcting the labelling errors.

AffectNet[30] is a comprehensive resource for FER that contains over one million images collected from the web. It includes annotations of eight emotions: happiness, sadness, anger, fear, disgust, surprise, contempt and neutrality. AffectNet supports large-scale training and evaluation, thereby enabling robust emotion recognition in diverse real-world scenarios. Because the test set of AffectNet is not yet available, we allocated 30% of the training samples for testing.

The RAF-DB-basic[31] initially consists of 15,339 facial images from real scenes. The dataset used for this project is highly diverse, featuring facial expressions from individuals of different ages, genders and races, in various postures and settings. This extensive dataset offers excellent generalisability and robustness. It includes the following single expression labels: surprise, 1619; fear, 355; disgust, 877; happiness, 5957; sadness, 2460; anger, 867; and neutral, 3204. Fig. 3 illustrates examples for each class in the dataset used.

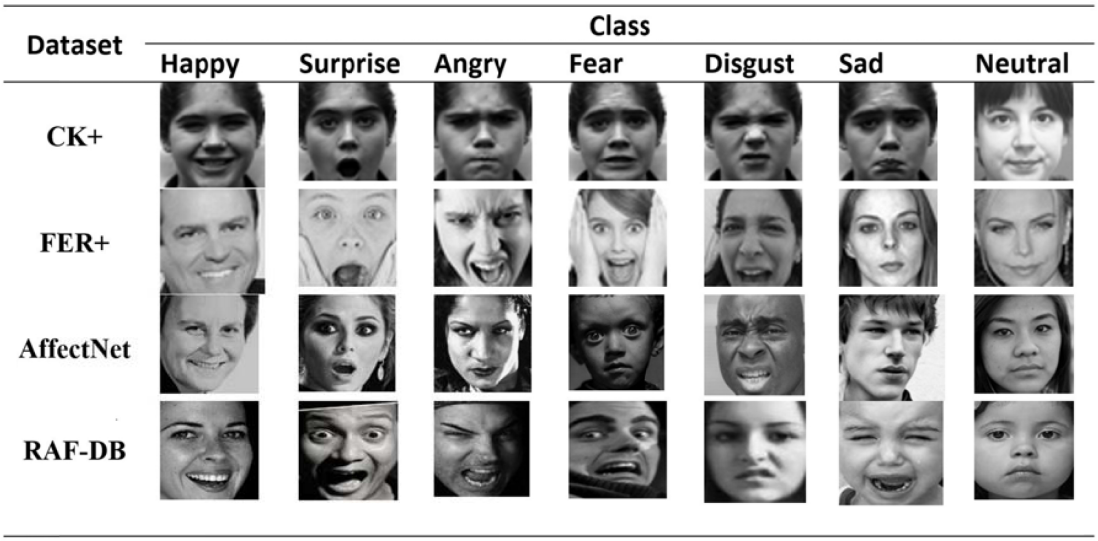

Fig. 3. Shows the example for each class of facial expression.[28–31]

Model training environment and experimental setup

An NVIDIA GeForce 4080 RTX GPU with 12 GB of memory and 64 GB of computer memory was used for training. The hyperparameters for our model were selected based on established conventions in deep learning for computer vision and in FER.[11,15] A batch size of 32 represents a common trade-off between computational efficiency and stable gradient estimation. The learning rate of 0.0001 was chosen to facilitate stable convergence with the Adam optimizer, as larger rates (e.g., 0.001) often lead to divergence in fine-tuned architectures, while smaller rates (e.g., 0.00001). The inclusion of weight decay (1e-4) serves as a regularizer to prevent overfitting. The model was trained for 200 epochs using early stopping to determine the optimal number of epochs. The HS-AMM approach addresses imbalance in three ways: First, Data-level through extensive augmentation (random horizontal flipping, rotation, and color jitter) to increase the effective diversity of minority classes; Second, prediction-level by employing class-weighted loss functions, calculating weights inversely proportional to class frequencies for both coarse-level and fine-level predictions to penalize rare emotions more heavily; Third, the Evaluation-level prioritizes the F1-score over accuracy to ensure a fair assessment across all classes. This rigorous approach ensures the proposed model learns robust and unbiased features across all emotion categories.

Results and discussion

The results in Table 2 show the performance of different CNN models on the CK+ dataset for emotion recognition. AlexNet, ResNet-18 and VGG-16 achieved accuracies of 93%, 86.3% and 98%, respectively. The proposed HS-AMM model achieved an impressive accuracy of 99.5% and F1 score of 99.49%, making it the most effective model in this study. Among the models considered, it achieved the best accuracy and reliability in emotion recognition.

Table 3 shows the performance of various CNN models on the FER+ dataset, with AlexNet achieving an accuracy of 65.7% and ResNet-18 achieving an accuracy of 64.8%. VGG-16 performed better, attaining an accuracy of 71% and an F1 score of 87.88%. The proposed HS-AMM model outperformed all baseline models, achieving the highest accuracy of 78.53%, demonstrating its superior emotion-recognition capability on this dataset.

Table 4 presents the performance of CNN models on the AffectNet dataset. AlexNet achieved an accuracy

Table 2. Performance comparison of CNN models and proposed HS-AMM on CK+ dataset.

CNN models	Accuracy %	F1-score %
AlexNet[32]	93.00	90.80
ResNet-18[33]	86.30	85.10
VGG-16[28,34]	98.00	97.30
HS-AMM (proposed)	99.50	99.49

Table 3. Performance comparison of CNN models and the proposed HS-AMM on the FER+ dataset.

CNN models	Accuracy %	F1-score %
AlexNet[35]	65.70	65.00
ResNet-18[33]	64.80	63.10
VGG-16[36]	71.00	69.40
HS-AMM (proposed)	78.53	87.80

Table 4. Performance comparison of CNN models and proposed HS-AMM on the AffectNet dataset.

CNN models	Accuracy %	F1-score %
AlexNet[30]	72.00	69.40
ResNet-18	74.48	72.12
VGG-16	79.56	78.25
HS-AMM (proposed)	84.59	91.22

Table 5. Performance comparison of the CNN models and the proposed HS-AMM on the RAF-basic dataset.

CNN models	Accuracy %	F1-score %
AlexNet	54.00	52.10
ResNet-18	77.60	76.30
VGG-16[36]	70.00	68.40
HS-AMM (proposed)	82.74	79.70

of 72% with an F1 score of 57%. ResNet-18 exhibited a low accuracy of 40.03% and a low F1 score of 10.90%. In contrast, VGG-16 and the proposed HS-AMM model demonstrated better performance, with accuracies of 84.59% and 91.22%, respectively. HS-AMM achieved the highest F1 score of 91.22%, highlighting its effectiveness in handling complex emotion recognition on this dataset.

Table 5 lists the performance of CNN models on the RAF-basic dataset. AlexNet achieved an accuracy of 70%, and ResNet-18 showed better performance on this set, achieving an accuracy of 82.74%. The proposed HS-AMM model achieved the best results, with an accuracy of 91.22%. These findings indicate that HS-AMM is highly effective for emotion recognition on the RAF-basic dataset, outperforming standard CNN baseline models in terms of accuracy. To situate our contribution within recent advances, we compare HS-AMM against contemporary state-of-the-art methods in Tables 2 to 5; this demonstrates that HS-AMM is not only stronger than conventional architectures but also achieves performance at the level of recent specialized FER methods. As shown in Fig. 4, it provides a comprehensive comparison of model performance across four benchmark datasets (CK+, FER+, AffectNet, and RAF-DB), evaluating both accuracy and F1-score. Our HS-AMM model demonstrates consistent superiority over all baseline architectures (AlexNet, ResNet-18, and VGG-16) on every dataset and both metrics. Notably, HS-AMM achieves particularly substantial gains on the more challenging in-the-wild datasets. This indicates that our hierarchical attention mechanism provides robust generalization across both controlled laboratory settings and complex real-world environments.

Ablation experiments

The performance of the five VGG-based schemes in emotion recognition was evaluated.

- Scheme 1 used a standard VGG model as the baseline.
- Scheme 2 added normal spatial attention classification.
- Scheme 3 introduced the proposed level-2 multi-class method.
- Scheme 4 incorporated spatial attention without level-2 processing (level-1 fine class only).
- Scheme 5, the proposed HS-AMM model, combines multi-class classification.

These schemes progressively enhance the ability of the model to capture the features corresponding to various emotions, with Scheme 5 being the most comprehensive approach for emotion recognition. Table 6 shows the ablation of the five schemes on four emotion-recognition datasets (CK+, FER+, AffectNet and RAF) in terms of accuracy and F1 score.

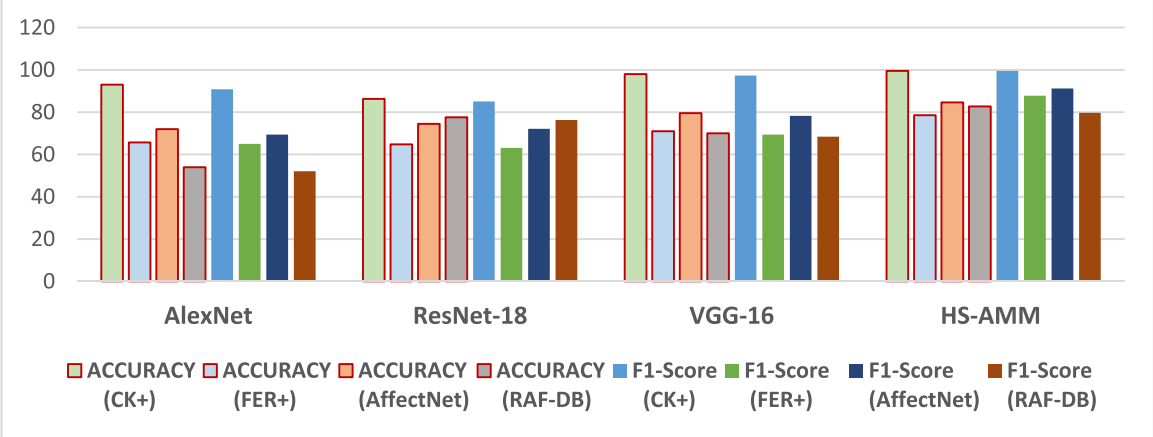

Fig. 4. Compares the Model with the Dataset, with Accuracy and F1-Score.

Table 6. Ablation experiment results of five schemes.

	Accuracy (%)				F1-score (%)			
Schemes	CK+	FER+	AffectNet	RAF	CK+	FER+	AffectNet	RAF
1	98	62.42	40.03	70.00	100	64.84	10.90	78.62
2	96	65.00	26.00	76.00	86.64	64.84	10.90	79.39
3	95.43	73.53	84.59	76.98	98.70	62.74	77.53	86.51
4	20.81	60.34	46.93	72.20	70.17	59.77	29.98	74.48
5	99.5	78.53	84.59	82.74	99.49	87.88	77.53	91.22

Table 7. Statistical Comparison of Basile VGG-16 model VS HS-AMM model using the independent sample T-Test.

		VGG-16		HS-AMM			
Dataset	Round	Mean (%)	STD (%)	Mean (%)	STD (%)	t-value	p-value
CK+	8	97.71	0.78	98.52	0.72	1.86	0.43
FER+	6	71.34	0.65	77.70	0.51	15.17	0.001
AffectNet	8	80.01	1.40	82.49	2.19	2.29	0.02
RAF-DB	5	75.66	2.92	79.82	2.16	2.56	0.03

Scheme 5 was the top performer, achieving the highest accuracy and F1 scores for CK+, FER+ and RAF. In addition, it achieved the best results, together with Scheme 3, on AffectNet. Scheme 3 also displayed strong performance, particularly on AffectNet and RAF, whereas schemes 1 and 2 achieved moderate results. Scheme 4 lagged behind, with significantly lower metrics across all datasets. These findings highlight Scheme 5 as the most effective approach for emotion recognition, offering superior accuracy and reliability.

Analysis experiments

The following presents the quantitative results of our comprehensive comparative study. The proposed HS-AMM model demonstrates superior performance, achieving the accuracy and F1-scores across all four benchmark datasets. Notably, on the challenging in-the-wild datasets, HS-AMM attains an accuracy of 78.53% on FER+ and 82.74% on RAF-DB, outperforming all compared baseline models, including VGG-16 (71.0%, 70.0%) and ResNet-18 (64.8%, 65.6%). This significant margin of improvement, particularly on naturalistic data, underscores the efficacy of our hierarchical attention mechanism in generalizing beyond controlled laboratory settings. The most profound improvement is observed in the Macro F1-scores, which are a critical metric for imbalanced datasets: HS-AMM's F1-score of 91.22% on RAF-DB and 87.88% on FER+ demonstrates a substantial increase over previous methods. This demonstrates that our model is not merely leveraging majority classes but is exceptionally proficient at accurately recognizing the full spectrum of emotions, including rare and difficult categories. To assess the statistical significance of the performance improvements achieved by the proposed HS-AM module, independent samples t-tests were conducted across the four datasets. The difference is statistically significant in Table 7. Finally, its performance is validated only on existing benchmarks, leaving its efficacy across diverse data unknown.

Conclusions

This study has introduced the Hierarchical Spatial Attention Mechanism Model (HS-AMM), a framework that advances facial expression recognition (FER) by a structured process of coarse-to-fine emotion perception. By integrating a psychologically-grounded hierarchical categorization with multi-scale spatial attention mechanisms, the HS-AMM represents a step toward context-aware emotion recognition systems. This work includes: First, providing validation for hierarchical models of emotion processing, demonstrating that a computational framework of human cognitive strategies can achieve superior performance; Second, establishing an adaptive architectural module that combines spatial attention mechanisms as an HS-AMM framework offers substantial advantages for real-world applications. The hierarchical output enables graceful degradation in ambiguous scenarios, allowing systems to default to robust valence-based predictions when fine-grained classification is uncertain. The computational complexity of the proposed three-level attention architecture may present deployment challenges for resource-constrained environments. While the model achieves performance on curated datasets (CK+: 99.49%, RAF-DB: 91.22 % F1-score), its relatively lower performance on AffectNet (77.53% F1-score) indicates persistent challenges in handling noisy, in-the-wild data. Building upon this research, we identify re-

search directions by exploring adaptive, learnable hierarchy structures that can accommodate differences in emotion expression, thereby increasing the applicability of FER systems. These directions address the current limitations while leveraging the strengths of our hierarchical approach to advance toward more human-like emotion understanding.

Acknowledgment

This work has been supported by the Ministry of Higher Education Malaysia Grant: (FRGS/1/2024/ICT02/UKM/02/5) and the University Kebangsaan Malaysia Research University Grant (GUP-2021-063).

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University Kebangsaan Malaysia.

Authors' contributions statement

A.O. conceived and designed the study, performed the experiments, and analyzed the data. A. A. contributed to data interpretation, literature review, and drafting the manuscript. S.S. assisted in experimental work, data validation, and critical revision of the manuscript. All authors contributed significantly to the work reported and read and approved the final version of the manuscript before submission.

References

1. Li K, Jin Y, Akram MW, Han R, Chen J. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput. 2020;36(2):391–404. https://doi.org/10.1007/s00371-019-01627-4.
2. Fernandez PDM, Pena FAG, Ren TI, Cunha A. FERAtt: Facial expression recognition with attention net. IEEE Computer society conference on computer vision and pattern recognition workshops, long beach, CA, USA; 2019. p.837–846. https://doi.org/10.1109/CVPRW.2019.00112.
3. Bao Z, Zhao Y. Facial emotion recognition based on hierarchical attention mechanism. ACM Int Conf Proceedings S. 2024;287–292. https://doi.org/10.1145/3641584.3641627.
4. Gan C, Xiao J, Wang Z, Zhang Z, Zhu Q. Facial expression recognition using densely connected convolutional neural network and hierarchical spatial attention. Image Vis Comput. 2022;117:104342. https://doi.org/10.1016/J.IMAVIS.2021.104342.
5. Dutta S, Ganapathy S. HCAM – Hierarchical cross attention model for multi-modal emotion recognition. Arxiv. 2024;1–11. https://doi.org/10.48550/arXiv.2304.06910.
6. Handrich S, Dinges L, Al-Hamadi A, Werner P, Al Aghbari Z. Simultaneous prediction of valence/arousal and emotions on affectnet, Aff-wild and AFEW-VA. Procedia Comput Sci. 2020;170:679–84. https://doi.org/10.1016/j.procs.2020.03.134.
7. Xiang Z, Radzi NHM, Hashim H. Research on emotion classification based on multi-modal fusion. Baghdad Sci J. 2024;21(2):Article 26. https://doi.org/10.21123/bsj.2024.9454.
8. Costanzi M, Cianfanelli B, Saraulli D, Lasaponara S, Doricchi F, Cestari V, *et al.* The effect of emotional valence and arousal on visuo-spatial working memory: Incidental emotional learning and memory for object-location. Front Psychol. 2019;10:2587. https://doi.org/10.3389/fpsyg.2019.02587.
9. Shin DH, Chung K, Park RC. Detection of emotion using multi-block deep learning in a self-management interview app. Appl Sci. 2019;9(22). https://doi.org/10.3390/app9224830.
10. B. Hasani, M.H. Mahoor. Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Comput Soc Conf Comput vis pattern recognit workshops. 2017;2017:2278–2288. https://doi.org/10.1109/CVPRW.2017.282.
11. Guo D, Mu J, Xu F, Li M. Research on facial expression recognition based on wide attention and multiscale fusion mechanism. J Ambient Intell Smart Environ. 2025;0(0). https://doi.org/10.1177/18761364241296439.
12. Gheorghe C, Duguleana M, Boboc RG, Postelnicu CC. Analyzing real-time object detection with YOLO algorithm in automotive applications: A review. *CMES* Comput Model Eng Sci. 2024;141(3):1939–81. https://doi.org/10.32604/cmes.2024.054735.
13. Febrian R, Halim BM, Christina M, Ramdhan D, Chowanda A. Facial expression recognition using bidirectional LSTM - CNN. Procedia Comput Sci. 2023;216:39–47. https://doi.org/10.1016/j.procs.2022.12.109.
14. K. Li, Y. Jin, M.W. Akram, R. Han, J. Chen. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput. 2020;36(2):391–404. https://doi.org/10.1007/s00371-019-01627-4.
15. Do LN, Yang HJ, Nguyen HD, Kim SH, Lee GS, Na IS. Deep neural network-based fusion model for emotion recognition using visual data. J Supercomput. 2021;77:1077390. https://doi.org/10.1007/s11227-021-03690-y.
16. Liu Y, Qiao X, Liu Z, Xia X, Zhang Y, Cui J. Deep multi-negative supervised hashing for large-scale image retrieval. Expert Syst Appl. 2025;264:125795. https://doi.org/10.1016/j.eswa.2024.125795.
17. Ouyang S, Wang H, Niu Z, Bai Z, Xie S, Xu Y, *et al.* HSVLT: Hierarchical scale-aware vision-language transformer for multi-label image classification. MM 2023, Proceedings of the 31st ACM international conference on multimedia, ACM, New York, USA; 2023. p.4768–4777. https://doi.org/10.1145/3581783.3612159.
18. Liu ZG, Fu YM, Pan Q, Zhang ZW. Orientational distribution learning with hierarchical spatial attention for open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2023;45(7):8757–72. https://doi.org/10.1109/TPAMI.2022.3227913.

19. Yang X, Luo Y, Zhang Z, Tang D, Zhou D, Tang H. AMHFN: Aggregation multi-hierarchical feature network for hyperspectral image classification. Remote Sens. (Basel) 2024;16(18):3412. https://doi.org/10.3390/rs16183412.
20. Feng C, Wang X, Zhang Y, Zhao C, Song M. CASwin transformer: A hierarchical cross attention transformer for depth completion. IEEE Conf. Intell. Transp. Syst. Proc. ITSC. 2022:2836–2841. https://doi.org/10.1109/ITSC55140.2022.9922273.
21. Rehman A, Mujahid M, Elyassih A, AlGhofaily B, Bahaj SAO. Comprehensive review and analysis on facial emotion recognition: performance insights into deep and traditional learning with current updates and challenges. Comput Mater Continua. 2025;82(1):41–72. https://doi.org/10.32604/cmc.2024.058036.
22. Zhao Z, Yang X, Zhou Y, Sun Q, Ge Z, Liu D. Real-time detection of particleboard surface defects based on improved YOLOV5 target detection. Sci Rep. 2021;11(1):21777. https://doi.org/10.1038/s41598-021-01084-x.
23. Mollahosseini A, Hasani B, Mahoor MH, AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2019;10(1):18–31. https://doi.org/10.1109/TAFFC.2017.2740923.
24. Liu Y, Shao Z, Hoffmann N. Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv.org. 2021. https://doi.org/10.48550/arXiv.2112.05561.
25. Behera A, Wharton Z, Liu Y, Ghahremani M, Kumar S, Bessis N. Regional attention network (RAN) for head pose and fine-grained gesture recognition. IEEE Trans Affect Comput. 2023;14(1):549–562. https://doi.org/10.1109/TAFFC.2020.3031841.
26. Aguilera-Martos I, Herrera-Poyatos A, Luengo J, Herrera F. Local attention: enhancing the transformer architecture for efficient time series forecasting. International Joint Conference on Neural Networks (IJCNN), 2024; Yokohama, Japan; 2024.p.1–8. https://doi.org/10.1109/IJCNN60899.2024.10650762.
27. Kinasih FMTR, Machbub C, Yulianti L, Rohman AS. Two-stage multiple object detection using CNN and correlative filter for accuracy improvement. Heliyon. 2023;9(1):e12716. https://doi.org/10.1016/j.heliyon.2022.e12716.
28. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. IEEE Computer Society Conference on computer Vision and pattern recognition – workshops, CVPRW 2010; San Fransicso, CA, USA; 2010,p.94–101. https://doi.org/10.1109/CVPRW.2010.5543262.
29. Schindler S, Bublatzky F. Attention and emotion: An integrative review of emotional face processing as a function of attention. Cortex. 2020;130:362–86. https://doi.org/10.1016/j.cortex.2020.06.010.
30. Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2019;10(1):18–31. https://doi.org/10.1109/TAFFC.2017.2740923.
31. Ajlouni N, Özyavaş A, Ajlouni F, Takaoğlu F, Takaoğlu M. Enhanced hybrid facial emotion detection & classification. Franklin Open. 2025;10:100200. https://doi.org/10.1016/j.fraope.2024.100200.
32. Sukhavasi SB, Sukhavasi SB, Elleithy K, El-Sayed A, Elleithy A. Deep neural network approach for pose, illumination, and occlusion invariant driver emotion detection. Int J Environ Res Public Health. 2022;19(4):2352. https://doi.org/10.3390/IJERPH19042352.
33. Shen T, Xu H. Facial expression recognition based on multi-channel attention residual network. Comput. Model Eng Sci. 2023;135(1):539–560. https://doi.org/10.32604/CMES.2022.022312.
34. Abdullah A, Wong WS, Albashish D. EB-CNN: Ensemble of branch convolutional neural network for image classification. Pattern Recognit Lett. 2025;189:1–7. https://doi.org/10.1016/J.PATREC.2024.12.017.
35. Agrawal A, Mittal N. Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. Vis Comput. 2020;36(2):405–412. https://doi.org/10.1007/S00371-019-01630-9/METRICS.
36. Ge H, Zhu Z, Dai Y, Wang B, Wu X. Facial expression recognition based on deep learning. Comput Methods Programs Biomed. 2022;215:106621. https://doi.org/10.1016/j.cmpb.2022.106621.

HS-AMM : نموذج آلية الانتباه المكاني الهرمي للتعرف على تعابير الوجه بالاعتماد على التعلم العميق

أحمد عدي[1,2]، عزيز عبد الله[1]، شه نوربانون سحران[1]

[1]مركز بحوث التعرف على الأنماط، مركز تقنيات الذكاء الاصطناعي، كلية علوم وتكنولوجيا المعلومات، جامعة الوطنية ماليزيا، 43600 بانجي، ماليزيا.
[2]قسم المعلوماتية الحيوية، كلية المعلوماتية الطبية الحيوية، جامعة تكنولوجيا المعلومات والاتصالات، 10001 بغداد، العراق.

الخلاصة

يُعدّ التعرّف على تعابير الوجه تحديًا كبيرًا نظرًا لتعقيد تصنيفها إلى فئات متعددة والاختلافات الدقيقة بين العوامل العاطفية، حيث غالبًا ما تواجه نماذج التعلّم العميق الحالية صعوبة في الاستفادة من السمات المحلية الهرمية لتحقيق تمييز دقيق. ولإيجاد حل، تقترح هذه الدراسة منهجًا ثنائيًا يهدف إلى تحسين دقة التعرّف على تعابير الوجه من خلال تحسين التصنيف عبر بُعد الإثارة والتكافؤ، وتحسين استخلاص السمات عبر آلية انتباه تكيفية للتغلب على صعوبة التمييز بين التعابير المتقاربة دلاليًا. تجمع المنهجية المقترحة بين: (1) استراتيجية تصنيف ثنائية المستوى، حيث يتم أولًا إجراء تصنيف دقيق، وثانيًا تصنيف متعدد الفئات من خلال تجميع التعابير في فئات محايدة وإيجابية وسلبية بناءً على الأبعاد العاطفية. (2) نموذج انتباه هرمي يُخفف من صعوبة التعرّف الدقيق على التعابير من خلال دمج الانتباه العالمي والإقليمي والمحلي في تسلسل هرمي مكاني مُعاد تصميمه، حيث يتم أولًا تعلّم تجميع دلالي قوي (محايد، إيجابي، سلبي) للتعابير. يُقلل هذا النهج من التشويش على مستوى عالٍ، ويُوفر مسارًا أوضح للتصنيف الدقيق اللاحق. تستخلص هذه الوحدة ميزات المشاعر من صور الوجه، وقد تم تنفيذها بالكامل على بنية VGG-16 المُحسّنة. تُظهر التجارب على مجموعات بيانات CK+ وFER+ وAffectNet وRAF-basic أن النموذج المقترح يحقق دقةً ومتانةً فائقتين مقارنةً بالأساليب الحالية، حيث يلتقط بفعالية ميزات متعددة المستويات، ويستفيد من التقسيم الهرمي لتحسين الأداء. تؤكد النتائج أن نهجنا، من خلال دمجه بين التجميع متعدد الفئات والانتباه المكاني الهرمي، يُقدم حلاً أكثر موثوقيةً وتطورًا لتطبيقات التعرف على المشاعر في العالم الحقيقي.

الكلمات المفتاحية: التعرف على تعابير الوجه، تصنيف الوجه، الانتباه الهرمي، متعدد الفئات، الانتباه المكاني.