AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-1| 2026 Published: |30-01-2026|

A METHOD FOR ENHANCING ACCURACY IN FACE RECOGNITION SYSTEMS USING HYBRID NEURAL NETWORKS

https://doi.org/10.5281/zenodo.18211267

Nurjanov Furqatbek Reyimberganovich
PhD, associate professor,
University of Military Security and Defense of the Republic of Uzbekistan

Abstract

This article investigates a hybrid approach based on convolutional neural networks and Vision Transformer architectures to improve the accuracy of face recognition systems. The primary objective of the study is to enhance the quality and robustness of identification under challenging conditions, such as partial face occlusion (e.g., masks, glasses), low illumination, or images captured from angled viewpoints. The proposed hybrid model extracts local features through CNN layers and models global dependencies using Transformer blocks. By integrating local and global features, the approach aims to improve face recognition accuracy and overall system robustness.

Keywords

face recognition, hybrid neural network, CNN, Vision Transformer, identification, accuracy.

Introduction

In recent years, biometric identification systems, particularly face recognition technologies, have gained significant importance in ensuring public safety and enabling the use of e-government and digital services. Although scientific advances in artificial intelligence and deep learning have substantially improved the accuracy and efficiency of face recognition algorithms, achieving high robustness in complex environmental conditions and real-time operation remains a challenging and relevant problem.

In particular, the sharp increase in demand for biometric systems during 2024–2026 has made ensuring an optimal balance between image processing speed and algorithmic robustness one of the key scientific challenges.

Under particularly challenging conditions—such as partial facial occlusion (masks, glasses), low illumination, or images captured from angled viewpoints—the quality of identification may degrade[1]. Therefore, the primary objective of this study is to improve the accuracy and robustness of face recognition systems under

such conditions. To achieve this goal, the proposed model must perform the following tasks:

Local feature extraction. Convolutional neural networks (CNNs) identify facial regions and micro-level details, enabling the preservation of important visual cues even in partially occluded faces.

Modeling global dependencies. Vision Transformer (ViT) blocks analyze long-range relationships among features extracted by the CNN and ensure robustness by capturing the global context of facial structures.

Integration of local and global features. The information extracted by the CNN and ViT is combined through a feature fusion mechanism [2]. This process ensures the complementary interaction of local and global features, maximizing both identification accuracy and system robustness.

Adaptability to real-time systems. The hybrid architecture optimally allocates computational resources, enabling deployment in modern biometric and security systems with real-time operation.

Currently, widely used convolutional neural networks (CNNs) demonstrate high performance in extracting local features from images. However, they have certain limitations in fully capturing the global context and the complex geometric relationships within facial structures. These shortcomings become particularly evident when parts of the face are occluded (masks, glasses, headwear), under low-light conditions, or when images are captured from inconvenient angles.

Another significant challenge facing modern security systems is the increasing number of attempts to deceive them using “deepfake” and artificially synthesized images[3]. Effectively addressing these issues cannot rely on a single architecture alone. This creates the need to develop and deploy hybrid neural networks that combine the advantages of different architectures, such as CNNs and Vision Transformers.

Material methods

The evolution of face recognition technologies can be analyzed in three main stages:

1. Traditional statistical methods,
2. Deep convolutional networks,
3. Modern models based on Transformer mechanisms.

Traditional Methods. In the early stages of face recognition, geometric and statistical approaches were primarily used. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were widely applied for dimensionality reduction of facial images and extraction of key features. Additionally, algorithms such as Local Binary Patterns (LBP) and Haar-like features proved effective in

analyzing facial texture information. However, these methods exhibited low robustness to changes in illumination, external noise, and natural variations in a person's facial appearance.

Convolutional Neural Networks (CNNs). Since 2012, the development of deep learning technologies has brought a fundamental shift in the field of face recognition. Architectures such as AlexNet, VGGNet, ResNet, and Inception have enabled the achievement of high accuracy. The main advantage of CNNs lies in their ability to automatically and hierarchically learn local features from images.

Notably, the FaceNet model achieved over 99% accuracy by representing facial images in a compact Euclidean space using the "triplet loss" function. However, the limited receptive field of CNNs makes it difficult to fully capture long-range global dependencies.

Vision Transformers. Since 2020, the introduction of Transformer architectures into the field of computer vision has opened new scientific directions. Vision Transformer (ViT) models learn the global context more effectively by dividing images into smaller patches[4]. Studies conducted in 2024–2025 have shown that ViT models exhibit higher robustness than CNNs in cases of partial facial occlusion. However, they require substantial computational resources and large-scale training datasets.

Formation of hybrid approaches. In recent years, interest in hybrid neural networks that combine CNN and ViT architectures has grown significantly. In such approaches, CNN layers are responsible for extracting low-level visual features from images, including edges, textures, and local patterns. Vision Transformer blocks, in turn, analyze long-range relationships among these features to capture global structure and context.

Several architectural variants of hybrid models have been proposed in the scientific literature. In some studies, CNN networks are used in the feature extraction stage, and the resulting feature maps are then passed to Transformer encoders. Other works propose using CNN and ViT blocks in parallel and combining their outputs through feature fusion methods. Such architectures not only improve face recognition accuracy but also enhance the model's generalization capability.

Studies show that hybrid approaches are particularly effective in challenging real-world conditions, such as partially occluded faces, low-light environments, or images captured from angled viewpoints. The CNN's ability to accurately extract local features and the Transformer's capability to model global dependencies complement each other, resulting in enhanced performance.

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-1| 2026 Published: |30-01-2026|

Problem solution. Based on the above analysis, it has been determined that optimizing computational resources is also crucial when developing hybrid neural networks[5]. Compared to full Vision Transformer (ViT) models, CNN–ViT hybrid architectures have fewer parameters, providing higher computational efficiency during both training and inference. This feature expands their applicability in real-time security and identification systems. The key aspects include:

Data preparation. The study utilizes publicly available facial image datasets. All images are standardized to a uniform format, the facial regions are detected and normalized, and the images are resized according to the requirements of the model architecture [6]. Additionally, to enhance the generalization capability of the dataset, data augmentation techniques are applied, including rotation, brightness adjustment, scaling, and horizontal flipping.

Hybrid model architecture.

The proposed architecture consists of two main components:

1.**CNN-based feature extraction module**, which is responsible for extracting low- and mid-level visual features from the image.

2.**Vision transformer (ViT) module**, which analyzes long-range dependencies among the features extracted by the CNN.

The CNN blocks consist of convolution layers, batch normalization, and activation functions, forming feature maps. These feature maps are resized appropriately and passed to the Transformer encoder blocks. The Transformer module comprises multi-head self-attention and feed-forward networks, effectively modeling the global context.

Feature fusion. In the hybrid model, features extracted by the CNN and ViT are combined through a dedicated fusion layer. The feature fusion process is performed using either concatenation or weighted summation and is then passed to the subsequent classification stage. This approach ensures the complementary interaction of local and global information.

Training strategy. The model is trained using a supervised learning approach. For the loss function, classification loss (soft max cross-entropy) is employed, along with metric learning-based loss functions (triplet loss or arc face Loss) to enhance identification accuracy. Optimization is performed using Adam or SGD optimizers, and a learning rate scheduling strategy is applied during training.

Evaluation metrics. The performance of the proposed model is assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, to evaluate the model's robustness under challenging conditions, images with occluded faces and low-light environments are analyzed in separate test scenarios.

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-1| 2026 Published: |30-01-2026|

Furthermore, to conduct a deeper analysis of the model's robustness and reliability, separate experiments were carried out under challenging scenarios. These included images with partially occluded faces (masks, glasses), low-light conditions, and images captured from various angled viewpoints [7]. The results obtained in these scenarios provided an assessment of the hybrid model's ability to operate effectively in real-world conditions.

Experimental results and discussion

The performance of the proposed CNN–Vision Transformer (CNN–ViT) hybrid neural network model is comprehensively evaluated through experiments conducted on various publicly available facial image datasets. The primary aim of these experiments is to analyze the identification accuracy and robustness of the hybrid approach, as well as to compare it with models based solely on traditional convolutional neural networks and individual Vision Transformer architectures. During the experiments, the models' performance was measured using standard evaluation metrics, and their ability to operate under different environmental conditions was analyzed separately.

Specifically, the model's performance was evaluated under challenging scenarios, such as partially occluded faces, low-light conditions, and images captured from various angles. The results demonstrate that the hybrid CNN–ViT architecture achieves high accuracy by effectively combining local and global features, while also maintaining robust performance in real-world conditions.

The experiments were conducted on the following widely used publicly available facial image datasets:

- **Labeled Faces in the Wild (LFW)** — facial images captured in real-world conditions, featuring variations in lighting and angles;
- **VGGFace2** — a highly diverse dataset containing faces with various poses and expressions;
- **CASIA-WebFace** — a widely used dataset for face identification tasks.

The dataset was divided into training (70%), validation (15%), and test (15%) sets. Additionally, special test scenarios were created for images with occluded faces and low-light conditions.

The proposed hybrid model was compared with the following baseline models (Table 1);

- CNN-based model (ResNet-50)
- **Vision Transformer (ViT-Base)**
- **Proposed CNN–ViT hybrid model**

The table below presents the average results of the models on the test dataset:

Table 1. Average test results of different models

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-1| 2026 Published: |30-01-2026|

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN (ResNet-50)	96.1	95.8	95.5	95.6
VisionTransformer (ViT-Base)	96.8	96.5	96.2	96.3
Proposed CNN-Vit Hybrid	98.2	98.0	97.9	97.9

These scenarios reflect the most challenging conditions that face recognition systems encounter in real-world applications. In cases of partial facial occlusion (wearing masks or glasses), important visual features of the face are hidden, complicating the identification process. Under low-light conditions, image quality deteriorates and noise levels increase, limiting the model's ability to accurately extract features [8]. Images captured from angled viewpoints can introduce changes in facial geometry and perspective effects, which may negatively affect identification accuracy.

The experimental results obtained under these challenging scenarios allowed for a comparison of model robustness. The analysis shows that the proposed CNN–Vision Transformer hybrid model demonstrated higher accuracy across all scenarios compared to both the traditional CNN and standalone Vision Transformer models. This superior performance can be attributed to the complementary synergy between the CNN's effective local feature extraction and the Transformer's ability to model global dependencies (table 2).

Table 2. Model perfomance under challenging scenarios

Сценарий	CNN (%)	ViT (%)	CNN–ViT (%)
Қисман ёпилган юз	90.4	92.1	**95.6**
Паст ёруғлик	88.7	90.5	**94.2**
Бурчак остида	89.9	91.3	**95.0**

The hybrid model demonstrated a significant advantage, particularly in cases of partially occluded faces and low-light conditions. This can be explained by the

complementary interaction between the CNN's ability to accurately extract local features and the Transformer's capability to model global dependencies.

The obtained results confirm the effectiveness of the hybrid approach. The CNN–ViT architecture not only improves identification accuracy but also ensures robust performance [9]. Furthermore, from a computational perspective, the hybrid model is more efficient than a full Vision Transformer, making it suitable for deployment in real-time security systems.

Conclusion

This paper proposes a hybrid neural network model based on convolutional neural networks and Vision Transformer architectures to improve the accuracy and robustness of face recognition systems. Within the study, the advantages and limitations of traditional CNNs and standalone Vision Transformer models were analyzed, and by combining their strengths, an architecture was developed that can operate effectively under challenging real-world conditions.

The experimental results demonstrated that the proposed CNN–ViT hybrid model outperforms traditional models in terms of identification accuracy, robustness, and generalization capability. In particular, the hybrid approach showed a significant advantage in scenarios involving partially occluded faces, low-light conditions, and images captured from various angles. This performance is attributed to the complementary interaction between the CNN's ability to accurately extract local features and the Transformer's capacity to model global dependencies.

The research results confirm that the proposed approach can be effectively applied in modern biometric identification and security systems, including real-time applications. Future research will focus on testing the model on large-scale and highly diverse datasets, as well as implementing mechanisms to enhance robustness against deepfake and artificially synthesized images.

REFERENCES:

1. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1701–1708.

2. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823.

3. Nurjanov F.R. Method and algorithm for identifying the parameters of the image face person // International Journal of Science and Research (IJSR). Volume

8 Issue 6, June 2019, India 08/06/2019. 7 page. Research Gate Impact Factor (2018):0.28/ SJIF(2018)7.426

4. Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR).

5. Круглов В. В., Борисов В. В. Искусственные нейронные сети. Теория и практика. — 1-е. — М.: Горячая линия - Телеком, 2001. - С. 382. -ISBN 5-93517-031-0.

6. R. Gonzalez, R. Woods, S. Eddins Digital image processing in MATLAB. M .: Tekhnosfera, 2006, -616 p.

7. Nurjanov F.R. Method and algorithm for identifying the parameters of the image face person // International Journal of Science and Research (IJSR). Volume 8 Issue 6, June 2019, India 08/06/2019. 7 page. Research Gate Impact Factor (2018):0.28/ SJIF(2018)7.426

8. Russell, SJ, & Norvig, P. (2016). Sun'iy intellekt: zamonaviy yondashuv. Pearson Education Limited.

9. Абламейко С.В., Лагуновский Д.М. Обработка изображений: технология, методы, применение. Минск: Ин-т техн, кибернетики НАН Беларуси, 2009. 300с.