A Vision Transformer Framework for Finger Vein Recognition with Regularized Mlp Head

1Ms. B. Vara Lakshmi, 2Ms.Teesa Davis,

Assistant Professor , Department of Computer Science & Engineering,

Geethanjali Institute Of Science And Technology, Nellore, India

Abstract

Vision Transformer (ViT) has drawn the attention of due to its superior performance in many computer vision tasks. However, there is limited research based on ViT models in finger vein recognition. Finger vein recognition is a highly secure biometric method, but its datasets are typically small, posing challenges for deep learning models like Vision Transformers (ViTs), which usually require large-scale data. To address this, we propose FV-ViT, a ViT-based model optimized for finger vein recognition. Instead of altering the ViT backbone, we introduce regMLP, a rigorous regularization technique in the MLP head, improving performance on limited datasets. FV-ViT achieves 0.042% EER on FV-USM and 1.033% EER on SDUMLA-HMT, outperforming existing methods. We also compare pretrained and non-pretrained versions, showing that ViTs can be trained from scratch for finger vein recognition with competitive results (0.068% vs. 0.116% EER on FV-USM and 1.258% vs. 1.022% EER on SDUMLA-HMT). This project explores new insights into the application of Vision Transformers for biometric recognition and highlights the potential of ViTs for high-security authentication systems. Future research may focus on further optimizing ViT architectures for finger vein analysis and exploring. advanced augmentation techniques to enhance performance on limited datasets.
Keywords:

Introduction

FV-ViT, a novel finger vein recognition system based on Vision Transformers (ViTs). Unlike traditional convolutional models, FV-ViT leverages the self-attention mechanism of transformers to extract global and discriminative features from finger vein images. The model is designed to work effectively on small-scale datasets like FV-USM and SDUMLA-HMT, which are typical in biometric research. To address the challenge of data scarcity, this project modify only the MLP classification head of the original ViT architecture—replacing it with a regularized MLP (regMLP)—instead of altering the core ViT encoder. This regMLP incorporates batch normalization, dropout, and GELU activation to improve generalization and reduce overfitting. Additionally, data augmentation techniques are applied to further enhance model performance on limited training data. In the following sections, we'll explore how the model is trained from scratch and achieves state-of-the-art accuracy and low Equal Error Rates (EERs), performing on par with or better than pretrained ViTs fine-tuned on the same tasks. Experimental results demonstrate FV-ViT's robustness, effectiveness, and scalability, making it a compelling alternative to CNN-based methods.
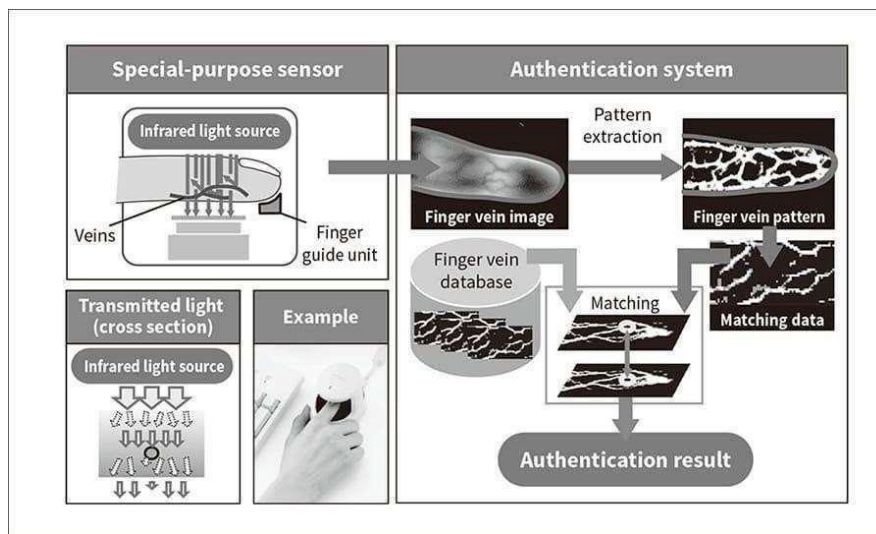
Fig 1. Process of Finger vein Recognition System

MOTIVATION

The main motivation for FV-ViT stems from the limitations of traditional finger vein recognition methods and the underutilization of Vision Transformers (ViTs) in this field. While CNN-based models have shown promising results, they mainly focus on local features and often struggle with capturing long-range dependencies, which are crucial for accurately recognizing complex vein patterns. Although ViTs have demonstrated powerful feature extraction abilities in many computer vision tasks, they are rarely applied to finger vein recognition because of a key challenge: ViTs typically require large-scale datasets, whereas finger vein databases are small and contain only a few samples per class. This project was motivated by this gap and sought to demonstrate that ViTs can still perform exceptionally well on small datasets without needing architectural changes or heavy pretrained models. By simply enhancing the MLP classification head with regularization techniques (batch norm, dropout, GELU) and applying data augmentation, they aimed to unlock the potential of ViTs in biometric recognition while addressing overfitting and improving generalization. Ultimately, the project aims to encourage further exploration of transformerbased models in biometric applications and show that, with minimal tweaks, ViTs can match or outperform pretrained or CNN-based approaches even in data-constrained environments.

Literature Review

In this chapter will review some papers to get knowledge and understanding on the techniques had been proposed. All those techniques have the same aim which is to recognize the finger vein patterns. As Archimedes once said, "Man has always learned from the past. After all, you can't learn history in reverse!" it is essential for man to learn from history. Thus, considering all past researches, the most relevant research glimpses have been picked to be explained in detail. The overview shall discuss relevant aspects contributing to our research.

2.1 How to train your ViT? Data Augmentation and Regularization
in Vision Transformers:

"Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman,
Jakob Uszkoreit, Lucas Beyer" This research explores "How to Train Your ViT?" provides a thorough empirical study on training Vision Transformers (ViTs), showing that data augmentation and regularization (AugReg) can significantly boost performance, often matching the effect of having 10 times more training data. It emphasizes that transfer learning—fine-tuning ViTs pre-trained on large datasets like ImageNet- 21k—is generally more effective and compute-efficient than training from scratch, especially for small to mid-sized datasets. The study finds that augmentation (e.g., Mixup, Rand Augment) is more beneficial than regularization techniques like

dropout, particularly for smaller datasets or shorter training runs. Additionally, it shows that larger pre- training datasets lead to more generalizable models, and recommends choosing pre-trained models based on their upstream validation performance.

2.2 ViT-Cap: A Novel Vision Transformer-Based Capsule Network

Model for Finger Vein Recognition:

"Yupeng Li, Huimin Lu, Yifan Wang, Ruoran gao and Chengcheng Zhao" This research introduces ViT-Cap, a new model for recognizing finger vein patterns, which are used as a secure and reliable biometric identification method. This model combines two powerful AI techniques—Vision Transformers, which focus on understanding the whole image using attention mechanisms, and Capsule Networks, which are good at detecting detailed patterns and relationships in small datasets. Together, they help the model identify important vein features even when image quality is poor or data is limited. Tested on several public finger vein datasets, ViT-Cap and outperformed many existing recognition methods. It's especially effective for real-world scenarios where image conditions aren't ideal and large amounts of training data aren't available.

2.3 Vision Transformers for Vein Biometric Recognition:

"Raul Garcia-Martin1 and Raul Sanchez-Reillo"

This research explores the use of Vision Transformers (ViTs) for recognizing people based on their vein patterns, which is known as Vascular Biometric Recognition (VBR). Traditionally, CNNs (Convolutional Neural Networks) were used for this task, but the researchers showed that ViTs, when pre-trained on large image datasets (ImageNet) and then fine-tuned on smaller vein datasets, can perform even better. They tested four types of vein images — finger, palm, back of the hand (dorsal), and wrist — using 14 different datasets. They also introduced a new dataset called UC3M-CV3, which contains contactless wrist vein images collected using smartphones. Overall, the study proves that Vision Transformers are highly effective for vein- based identification, especially when combined with transfer learning.

Proposed Model

To overcome the limitations of traditional finger vein recognition methods, this project introduces FV-ViT (Finger Vein Vision Transformer), a deep learning-based approach designed to enhance the accuracy, robustness, and efficiency of biometric authentication. The core of this system is the Vision Transformer (ViT) architecture, which offers a powerful alternative to convolutional neural networks by capturing both local and global features within an image. Unlike CNNs that rely on small receptive fields and local convolutions, ViTs divide the image into fixed-size patches and apply a self-attention mechanism, enabling the model to understand long-range dependencies and complex spatial relationships in the vein patterns. FV-ViT is a Transformer-based finger vein recognition system that preprocesses images and splits them into non-overlapping patches. These are converted into linear embeddings with positional encoding and passed through a Transformer encoder to extract discriminative features. A regularized MLP head improves generalization, especially on small or varied datasets. Trained on benchmark datasets like SDUMLA- HMT and FV-USM, FVViT learns subtle vein patterns under varying conditions. The end- to-end framework eliminates manual feature engineering and ensures high accuracy, robustness to noise and variations, and computational efficiency for both real-time and offline use.

ADVANTAGES OF PROPOSED SYSTEM:
1. Improved feature extraction using self-attention
2. Higher recognition accuracy
3. Robustness to variability in image quality and environmental conditions
4. Better generalization across diverse datasets
5. Enhanced computational efficiency, making it suitable for real-time use

This study is carried out to evaluate the economic and practical implications that the proposed FV-ViT system will have on the organization or institution deploying it. Given the limited budget typically allocated to research and development in biometric systems, it was essential to ensure that the expenses were justified and aligned with available funding . The proposed system was developed using primarily open-source tools and technologies, such as Python, PyTorch, and publicly available datasets (e.g., SDUMLA-HMT, FV-USM). This significantly reduced the cost of implementation. High-performance hardware was used where necessary, but wherever possible, existing infrastructure was utilized, making the overall solution cost-effective and accessing.

System Architecture
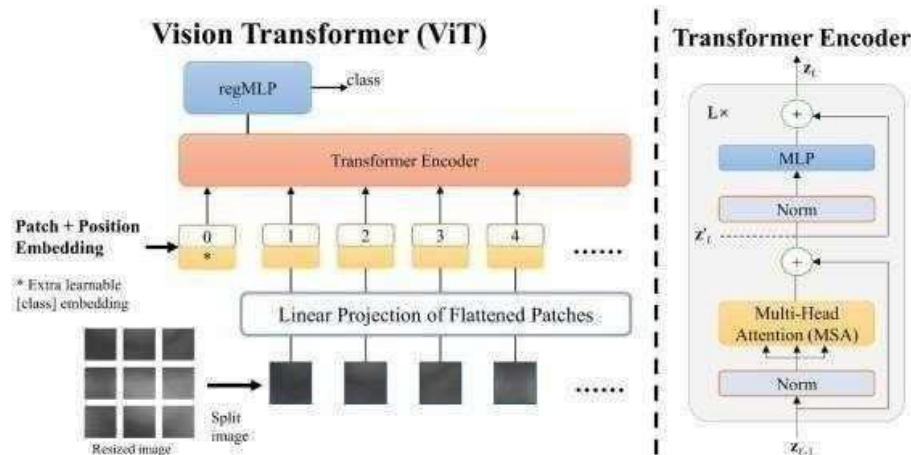
Fig.1. Architecture of ViT

1. Image Patches

The input image is split into small fixed-size patches (like tiny tiles). Each patch is flattened into a vector for processing.

2. Linear Projection of Flattened Patches

Each flattened patch is passed through a linear layer to turn it into a feature embedding (vector representation).

3. Patch + Position Embedding

Each patch embedding gets position information added, so the model knows where the patch came from in the image.

A special [class] token is also added, which will eventually carry the full image's representation.

4. Transformer Encoder

A stack of Transformer blocks, each with:

1. Multi-Head Self-Attention (MSA): Lets patches "see" each other.
2. MLP (Feedforward Network): Further processes each patch.
3. Layer Normalization and Residual Connections: Help stabilize and improve learning.

5. regMLP (Classification Head)

After transformer processing, the output of the [class] token is passed to an MLP (multi-layer perceptron). This produces the final class prediction (e.g., what object is in the image).

Figure 1 represents the overall architecture of the FV-ViT (Finger Vein Vision Transformer) model, which is an adaptation of the standard Vision Transformer (ViT) specifically designed for finger vein image analysis. The process starts with a resized finger vein image, which is divided into smaller fixed-size patches. Each patch is then flattened into a one-dimensional vector and passed through a linear projection layer, converting the raw pixel data into a dense feature representation (embedding). To enable the model to understand the spatial arrangement of patches, position embeddings are added to each patch embedding. Additionally, a special

learnable classification token (often denoted as [cls]) is prepended to the sequence of patch embeddings. This token is designed to aggregate information from all patches and is later used to represent the entire image for classification purposes.

The resulting sequence (which includes the position-augmented patch embeddings and the [cls] token) is then fed into a Transformer Encoder. This encoder consists of multiple stacked layers. Each layer contains two main components: a Multi-Head Self-Attention (MSA) mechanism, which allows the model to focus on different parts of the image simultaneously, and a regularized Multi-Layer Perceptron (regMLP), which is a modified version of the standard MLP to improve performance on the finger vein task. Both the MSA and regMLP blocks are followed by Layer Normalization (Norm) and include residual (skip) connections to maintain stable training and preserve gradient flow. After passing through the Transformer Encoder, the output corresponding to the [cls] token is extracted. This token now contains a global representation of the input image. Finally, it is passed through a regMLP head, which serves as the classifier, outputting the final class prediction—such as the identity of the person based on their finger vein pattern. This architecture keeps the core structure of the original Vision Transformer but introduces regMLP in place of the standard MLP to enhance the model's ability to extract features from finger vein images.

SYSTEM IMPLEMENTATION
SYSTEM MODULES
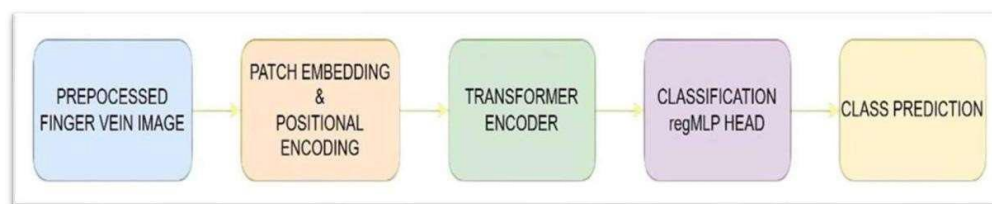
Pre-processed Finger Vein Image

The system begins with a pre-processed finger vein image. Preprocessing includes steps like noise reduction, contrast enhancement, and normalization to highlight the vein patterns and prepare the image for feature extraction. This ensures consistent input quality across the dataset.

Patch Embedding & Positional Encoding

The pre-processed image is divided into small patches, each of which is flattened and linearly embedded into a vector. Positional encoding is added to retain the spatial arrangement of the patches, allowing the transformer to understand the image structure despite its sequential input format.

Transformer Encoder

The transformer encoder processes the embedded patches using self-attention to capture both local and global features. This enables the model to focus on key patterns and relationships within the image, generating a rich representation of the finger vein structure.

Classification regMLP Head

The encoded features are passed through a regularized MLP head, which refines and condenses the information. Regularization techniques like dropout or normalization may be applied to enhance generalization and prepare the features for final classification. Class Prediction In the final stage, the model outputs a class label that corresponds to an individual's identity. Based on the extracted features, the system determines the most likely match, completing the finger vein recognition process.

Results & Analysis

The execution process of the project is shown below

Step -1: Setting up the environment

Step 2: Running the application and accessing the Web Interface using the URL

Commands that are used for running the project (in cmd
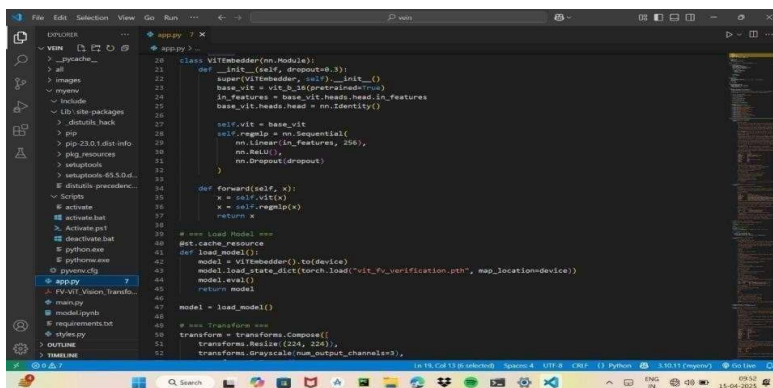prompt): To Activate virtual environment:
Myenv\scripts\activate
To Run the application:
Streamlit run
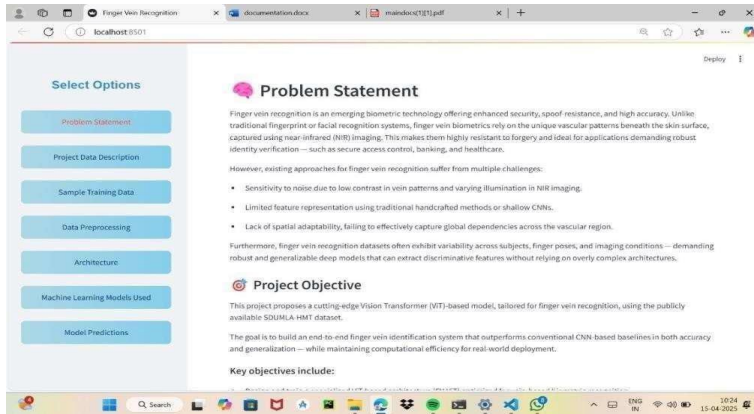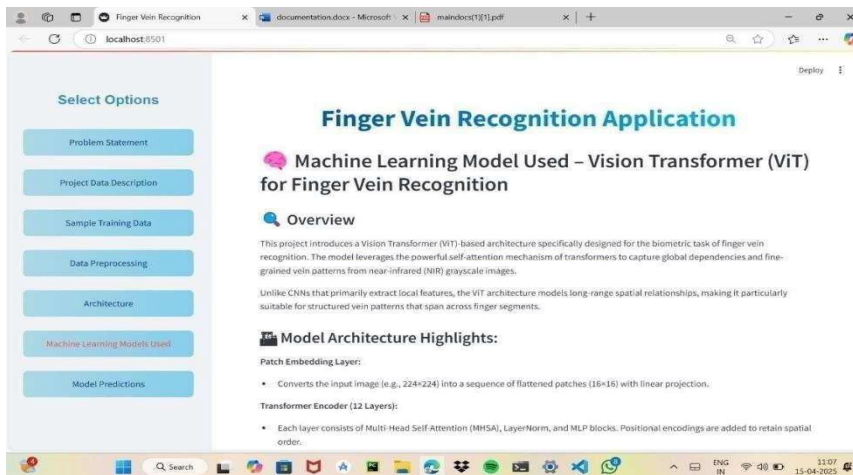app.py To Stop the application:
Ctrl + C

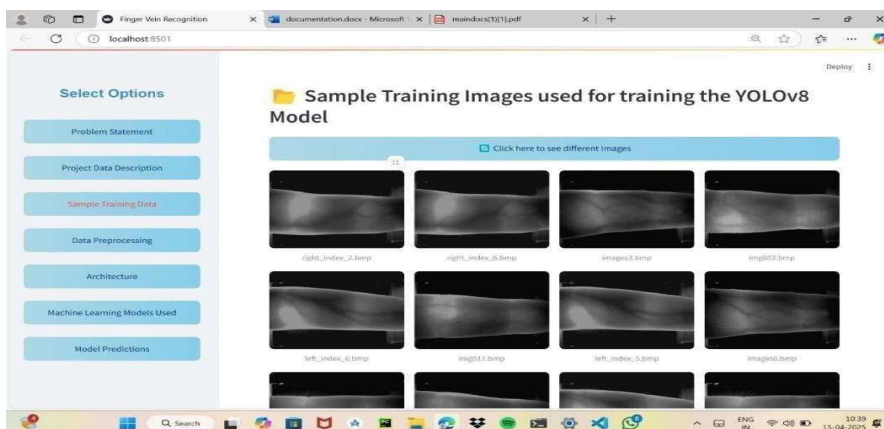Step-3: Run stream lit

Step -4: Home page of the project

Step -5: problem statement
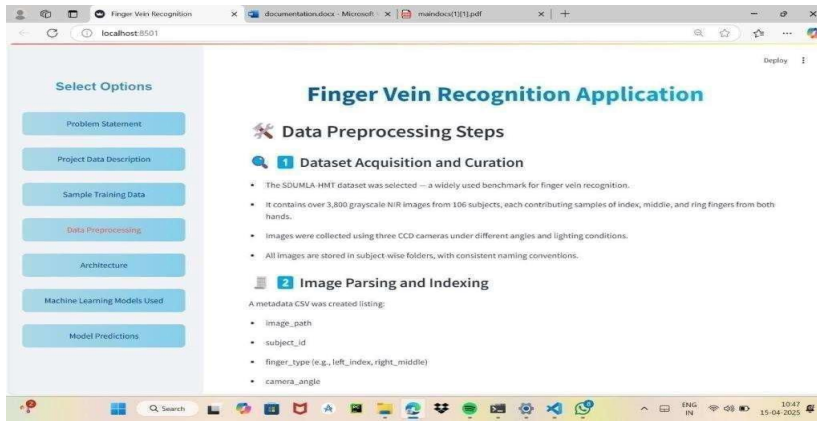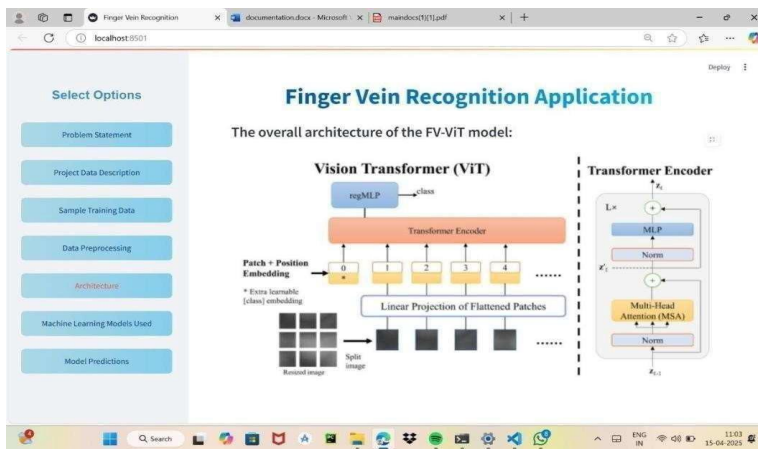

Step 6: project data description

Step -7 : Sample traning data

Step - 8: Data Preprocessing Steps
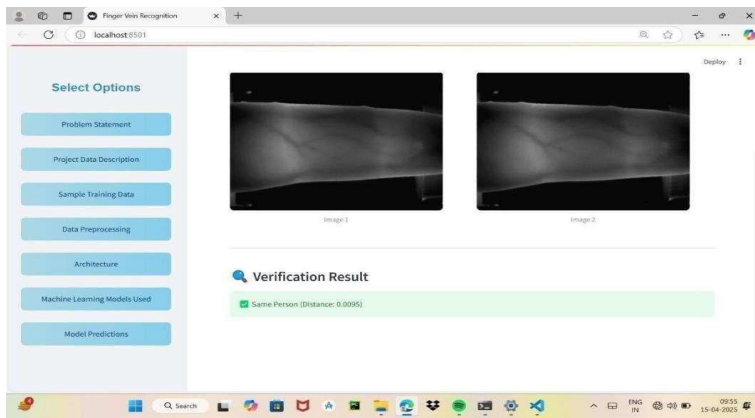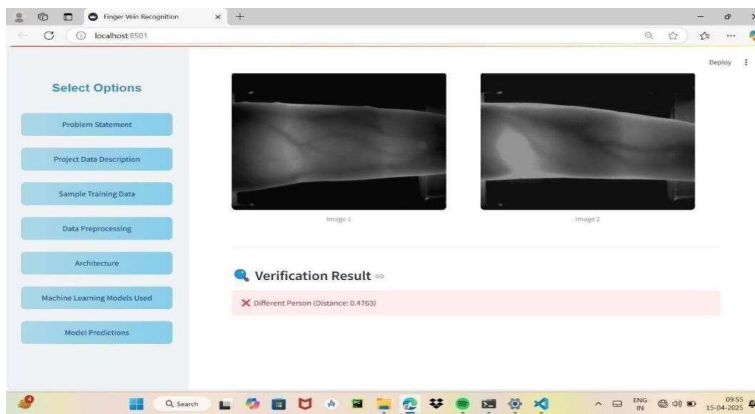
Step -9 : Architecture of vision Transformer

Step -10 : Finger Vein Recognition Application

It shows same person when both finger vein images are matched

It shows different person when both finger vein images are not matched

Conclusion

This project presents a practical and resource-efficient solution for finger vein recognition by leveraging the Vision Transformer architecture with minimal modifications. The proposed FV-ViT model introduces a regularized MLP classification head (regMLP) while keeping the original ViT backbone intact, enabling it to achieve high accuracy and low error rates on small-scale datasets without requiring extensive pretraining. Through the combination of strategic regularization and effective data augmentation, the model demonstrates robust performance, outperforming or matching state-of-the-art methods. Its modularity and simplicity not only reduce computational overhead but also enhance adaptability across various biometric recognition tasks. Overall, FV-ViT exemplifies how lightweight, attention-based models can be optimized for niche applications like finger vein recognition.

FUTURE SCOPE

This project lays a strong foundation for advanced research in biometric authentication using Vision Transformers. It can be extended to multi-modal biometric systems by integrating additional traits like face or iris recognition. Future work can focus on optimizing the model for real-time deployment on edge devices and exploring more efficient transformer architectures. With access to larger and more diverse datasets, the system's accuracy and robustness can be further enhanced. This project only focuses on closed protocol, but the open protocol is mainstream in real life. It needs further enhancement to improve the performance of the models in the open protocol. We add regularization to the MLP head, but constructing a kind of Transformer Encoder that performs well on small databases without sacrificing scalability is still worth exploring. The finger vein images

used in the paper are simply processed, we believe that using advanced pre- processing techniques can achieve better performance.

References

1. R. Garcia-Martin and R. Sanchez-Reillo, "Vision transformers for vein biometric recognition," IEEE Access, vol. 11, pp. 22060–22080, 2023, doi: 10.1109/ACCESS.2023.3252009

2. I. Boucherit, M. O. Zmirli, H. Hentabli, and B. A. Rosdi, "Finger vein identification using deeply-fused convolutional neural network," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 3, pp. 646–656, Mar. 2022, doi: 10.1016/j.jksuci.2020.04.002

3. K. Shaheed et al., "DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition," Expert Systems with Applications, vol. 191, Apr. 2022, Art. no. 116288, doi: 10.1016/j.eswa.2021.116288

4. J. Huang et al., "FVT: Finger vein transformer for authentication," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2022.3173276

5. Y. Li et al., "ViT-cap: A novel vision transformer-based capsule network model for finger vein recognition," Applied Sciences, vol. 12, no. 20, p. 10364, Oct. 2022, doi: 10.3390/app122010364

6. H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," arXiv preprint, arXiv:2106.08254, 2021.

7. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv preprint, arXiv:2103.14030, 2021.

8. H. Lu et al., "A novel finger-vein recognition approach based on vision transformer," in Proc. Int. Conf. Frontiers Electron., Inf. Comput. Technol., May 2021, pp. 1–5.

9. Z. Tao et al., "DGLFV: Deep generalized label algorithm for finger-vein recognition," IEEE Access, vol. 9, pp. 78594–78606, 2021, doi: 10.1109/ACCESS.2021.3084037

10. B. Hou and R. Yan, "ArcVein-arccosine center loss for finger vein verification," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–11, 2021, doi: 10.1109/TIM.2021.3062164