Vision-Language Model-Based Face Verification for Preventing Unauthorized Access

Junwoo Lim
University of Science and Technology (UST)
217 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
wnsdn9005@etri.re.kr

Ho-Sub Yoon
Electronics and Telecommunications Research Institute (ETRI)
218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
yoonhs@etri.re.kr

***Abstract*—Preventing unauthorized intrusion into restricted areas is a critical security concern. While existing methods utilizing CLIP or ViT have demonstrated excellent performance in face verification tasks, they often struggle with significant performance degradation when accessories are worn or facial features undergo substantial changes. To overcome these limitations in recognition accuracy, this study focuses on developing a novel technique leveraging Vision-Language Models (VLMs). Specifically, we propose a robust face verification method that remains resilient to accessory variations and facial feature changes by incorporating face cropping using a detector, an optimized dataset sampling technique for enhanced training efficiency, and LoRA fine-tuning of InternVL3 and Qwen2.5-VL models. Our proposed method achieves an accuracy of 97.32% on the LFW Dataset, thereby demonstrating the significant potential of VLMs in unauthorized access detection.**

***Index Terms*—Vision-Language Model(VLM), Face Verification, Security**

I. Introduction

Despite the active development of intelligent surveillance technologies leveraging CCTV video data for crime prevention and enhanced security, access control through face recognition still faces technical limitations. Previous efforts have proposed methods for authorized and unauthorized access control using face recognition technologies like MTCNN [1] and VGGFace2 [1]-based models, or through approaches such as CLIP-TSA [2] and CLIP [3] to detect both face and anomalous behavior. However, these methods frequently encounter performance degradation due to variations in facial angle, expression changes, or the presence of accessories, posing persistent challenges in accurately distinguishing between authorized and unauthorized individuals.

To overcome these limitations, this study proposes a novel technique for detecting unauthorized entry and anomalous behavior within restricted areas by utilizing Vision-Language Models. Our proposed method first employs a YOLOv8 [4] Face Detector to precisely crop the essential parts of faces for verification. Subsequently, we leverage InternVL3 [5] and Qwen2.5-VL [6] models, applying LoRA fine-tuning [7] to learn feature differences caused by various angles and expressions. For the VLM model training, we utilized the WebFace (Web-Collected Face Dataset) [8], and notably, we enhanced training efficiency by computing cosine similarity between face pairs and sampling the dataset based on specific criteria. Finally, identity verification is conducted by matching the facial features of entrants against that of enrolled authorized individuals, using various prompt formats to guide the Vision-Language Models during inference. Inference results on the LFW (Labeled Faces in the Wild) [9] dataset demonstrated a verification accuracy of up to 97.32%, thereby validating the effectiveness of our proposed approach.

II. Related Works

FaRL [11] (General Facial Representation Learning in a Visual-Linguistic Manner) Zheng et al. proposed a framework that combines image-text contrastive learning and masked image modeling. After pre-training on the LAION-FACE dataset, this approach demonstrated superior transfer performance compared to CLIP-ViT-B/16 across various downstream tasks, including facial parsing, face alignment, and facial attribute recognition.

In contrast, traditional face recognition methods such as ArcFace and CosFace utilize specific loss functions to learn highly discriminative facial features, achieving state-of-the-art performance in face recognition tasks, as seen in models like LVFace [12] that adopt a Vision Transformer (ViT) architecture combined with these traditional loss functions. LatentFace [13] He et al. introduced a generative self-supervised approach using a 3D-aware latent diffusion model to learn facial representations. This model demonstrated a significant improvement of about 3% points in face verification tasks compared to the RAF-DB benchmark.

While previous studies primarily focused on facial representation learning, alignment, and attribute recognition, our research distinguishes itself by concentrating explicitly on face verification for security environments involving accessory usage and varying facial angles. Unlike traditional methods that rely on loss-based feature learning, our study leverages the powerful visual and linguistic reasoning capabilities inherent in VLMs. Specifically, we employ Vision-Language Models (InternVL3 [5] and Qwen2.5-VL [6]) combined with face region cropping, cosine similarity-based dataset sampling, LoRA fine-tuning [7], and prompt optimization, thereby addressing practical verification challenges in real-world security scenarios.

III. Unauthorized Access Anomaly Detection System

The process of detecting and classifying unauthorized individuals within restricted areas is meticulously illustrated in "Fig. 1". Initially, upon the appearance of a pedestrian in the CCTV footage, the YOLOv8 [4] Face Detector is employed to accurately detect the pedestrian's face. The detected face image (with non-facial regions removed) and a pre-registered authorized person's face image (Input Image 2 in "Fig. 1") are then supplied as input to a Vision-Language Model. This model infers whether the two face images belong to the same individual based on a prompt, subsequently classifying the person as either authorized or unauthorized according to the inference result.

The accuracy of the face verification is measured as follows: A dataset comprising image pairs of the same individual and image pairs of different individuals is provided as input to the inference model. The model then infers whether the input image pairs are of the same person or different people, as depicted in "Fig. 1". Finally, the accuracy of the system is calculated by determining the ratio of results classified as 'true' through prompt-based classification relative to the total number of face pairs.

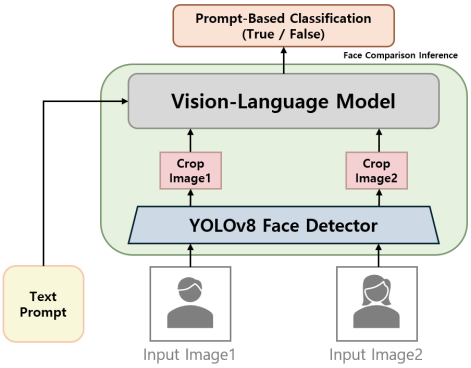

Fig. 1. Verification Process of the Proposed System

A. Face Detection and Cropping Images

For reliable face comparison, cropping input images to only the face region is crucial. Extraneous information can cause Vision-Language Models to focus on irrelevant features, degrading verification accuracy. Therefore, we used the YOLOv8 Face Detector to crop only the facial region, ensuring the VLM focuses solely on relevant features. Equation (1) illustrates the impact of such extraneous points.

$$E_{\text{feature}} = f(F_{\text{face}}) - \sum_{i=1}^{N} w_i \cdot F_{\text{other,i}} \tag{1}$$

Here, E_{feature} denotes extraction efficiency, $f(F_{\text{face}})$ is pure facial feature information, N is the number of non-facial features, w_i is their weight, and $F_{\text{other,i}}$ is their information amount. Increased non-facial features lower extraction efficiency.

B. Dataset Sampling for Fine-Tuning

In this study, we used the WebFace [8] dataset, which contains about 500,000 face images from over 10,000 identities with diverse angles and expressions, making it well-suited for training face verification models. "Fig. 2" shows examples from WebFace.

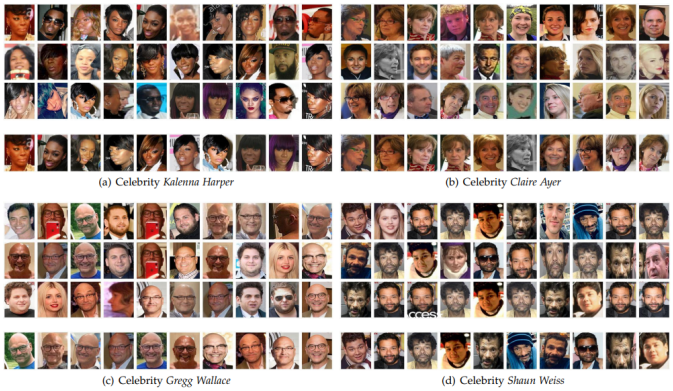

Fig. 2. Visualization of the WebFace [8] data.

For efficient training, we constructed 10,000 image pairs from the WebFace [8] dataset, splitting 62.5% for training and 37.5% for testing. Using cosine similarity, we sampled 5,000 same-person pairs (similarity 0.5–0.8) and 5,000 different-person pairs (similarity ≥ 0.1) to improve verification performance. This strategy is expressed in "(2)".

$$\begin{aligned} D_{\text{fine-tuning}} = \; & \{I_{\text{same},i} \mid 0.5 \leq \cos(F_{i,A}, F_{i,B}) \leq 0.8\}_{i=1}^{n} \\ & \cup \; \{I_{\text{diff},j} \mid \cos(F_{j,A}, F_{j,B}) \leq 0.1\}_{j=1}^{n} \end{aligned} \tag{2}$$

Here, $D_{\text{fine-tuning}}$ denotes the dataset used for fine-tuning, I_{same} and I_{diff} represent image pairs belonging to the same and different individuals, respectively, and $F(\cdot)$ denotes the facial feature extraction function. n represents the number of facial images. "TABLE I" provides a summary of the Fine-Tuning Dataset sampling strategy.

TABLE I
Face Pair Configuration for Fine-Tuning

Category	Dataset	Pairs	Cosine Sim.	Purpose
Same Person	WebFace [8]	5,000	0.5 – 0.8	Intra-class enhancement
Diff. Person	WebFace [8]	5,000	≤ 0.1	Inter-class enhancement
Total		**10,000**		

C. Prompt Formats

It is no exaggeration to say that the prompt is one of the most significant factors influencing the performance of Vision-Language Models [10]. In fact, even for the same task, the way a prompt is constructed can drastically alter model performance. In this study, we experimented with various prompt formats to identify the optimal prompt that yields

the best performance, considering both the Vision-Language Model and its parameter size. Below are the different prompt formats applied in our experiments:

- Basic Plain Text Prompt: Simple, direct text prompt (e.g., "Compare if faces are the same person.").
- Persona-Augmented Prompt: Incorporates a persona into the VLM prompt to assign task responsibility.
- Low-Detail Descriptive Prompt: Adds instructions for VLM to focus on a few key facial features (eyes, nose, mouth, face shape).
- High-Detail Descriptive Prompt: Provides explicit, sequential guidelines for VLM inference, comparing broad to fine facial details.
- YAML Format Prompt: Loads detailed prompt content from a YAML file.
- JSON Format Prompt: Loads detailed prompt content from a JSON file.

IV. Experiments

We evaluated our model using the LFW [9] dataset, which consists of 6,000 face image pairs (3,000 same-person and 3,000 different-person pairs). Based on this dataset, we analyzed performance differences across Zero-Shot settings, fine-tuning dataset sampling strategies, and various prompt formats.

A. Zero-Shot Evaluation of Each Model

To verify the face verification performance of Vision-Language Models without additional training, we conducted Zero-Shot tests on both InternVL3 [5] and Qwen2.5-VL [6] models. For InternVL3 [5], we utilized the 2B and 8B parameter models, while for Qwen2.5-VL [6], we used the 3B-Instruct and 7B-Instruct models. We measured the accuracy by providing cropped images along with text prompts as input to the Vision-Language Models. The results are as follows: the InternVL3-2B model achieved 89.73% accuracy, the InternVL3-8B model achieved 95.22%, and the Qwen2.5-VL-3B-Instruct model achieved 90.95%. Notably, the Qwen2.5-VL-7B-Instruct model demonstrated the best accuracy in Zero-Shot testing with an accuracy of 95.42%. "TABLE II" below visualizes the results of each experiment.

TABLE II
Zero-Shot Accuracy of Vision-Language Models

Model	Parameters	Accuracy (%)
InternVL3 [5]	2B	89.73
InternVL3 [5]	8B	95.22
Qwen2.5-VL-Instruct [6]	3B	90.95
Qwen2.5-VL-Instruct [6]	**7B**	**95.42**

B. LoRA Training Details

We fine-tuned the Vision-Language Models using the LoRA method on a powerful computing infrastructure to ensure stable and efficient training. The process was executed on eight NVIDIA A6000 GPUs, a configuration chosen to manage the significant computational demands of fine-tuning large-scale models. This setup allowed us to maintain a consistent training environment while applying the following hyperparameters:

- LoRA Rank : 8
- LoRA Alpha : 16
- LoRA Dropout: 0.1
- Learning Rate: 1e-4
- Optimizer: AdamW
- Epochs: 3.0

C. Evaluation of Fine-Tuning by Sampling Strategy

To effectively compare faces with varying angles and feature differences, efficiently structuring the Fine-Tuning dataset is critically important. In this study, instead of simply randomly extracting face image pairs from the Webface [8] dataset used for Fine-Tuning, we calculated the Cosine Similarity between image pairs and divided them into specific ranges to conduct separate tests. For these tests, we utilized the Qwen2.5-VL-3B-Instruct model, which demonstrated commendable performance even with relatively fewer parameters.

In experiments on Fine-Tuning dataset sampling criteria, randomly sampling 5,000 same-person pairs and 5,000 different-person pairs yielded 91.58% accuracy but raised reliability concerns. The Accuracy decreased to 61.70% when sampling 5,000 pairs each of same-person pairs with cosine similarity ≤ 0.6 and different-person pairs ≥ 0.1, primarily due to training on same-person pairs with significant feature variations.

To improve this, sampling 5,000 pairs each with same-person cosine similarity ≤ 0.7 and different-person ≥ 0.1 increased accuracy to 93.37%. After testing various cosine similarity ranges, the highest accuracy of 95.90% was achieved by sampling 5,000 pairs each with same-person cosine similarity of 0.5-0.8 and different-person pairs ≥ 0.1. Detailed results are presented in "TABLE III".

TABLE III
Accuracy Based on Cosine Similarity Sampling

Same-Person Range	Diff.-Person Range	Accuracy (%)
Random (5K pairs)	Random (5K pairs)	91.58
≤ 0.6	≥ 0.1	61.70
≤ 0.7	≥ 0.1	93.37
0.3–0.5	≥ 0.15	90.95
0.4–0.7	≥ 0.1	92.33
0.5–0.8	$\geq$ **0.1**	**95.90**

"Fig. 3" illustrates the inference process involving the application of the Fine-Tuned adapter to the Vision-Language Model. Each image pair is fed into the YOLOv8 [4] Face Detector to crop only the facial region. The cropped images, along with a text prompt, are then input into the Vision-Language Model, and the accuracy is measured by integrating the inference results.

D. Prompt Formats

The performance of Vision-Language Models is heavily dependent on the text prompt provided as input. [10] Selecting the correct prompt format tailored to the task and

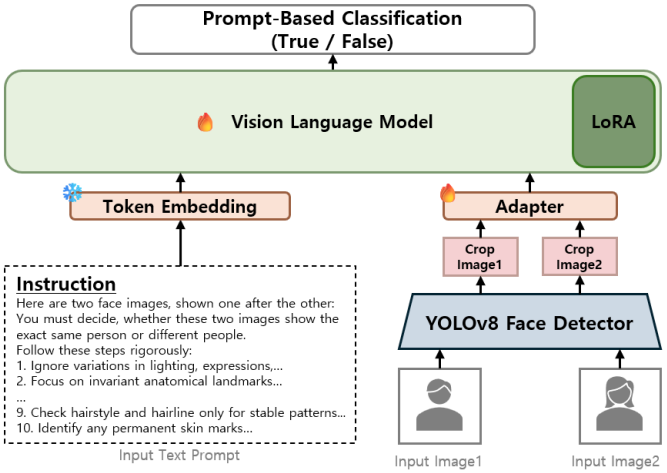

Fig. 3. VLM's inference process with Fine-Tuned adapter

model is crucial, as it significantly impacts performance variations. In this study, we conducted tests on Qwen2.5-VL-Instruct(3B, 7B) and InternVL3(2B, 8B) models using various prompt formats, including Basic Plain Text Prompt, Persona-Augmented Prompt, Low-Detail Descriptive Prompt, High-Detail Descriptive Prompt, YAML Format Prompt, and JSON Format Prompt. All tests were performed using models trained with image pairs sampled based on a cosine similarity of 0.5-0.8 for same-person pairs and ≥ 0.1 for different-person pairs.

The experimental results show that the Qwen2.5-VL-3B-Instruct model achieved its best accuracy with 95.90% accuracy using the Basic Plain Text Prompt, while the Qwen2.5-VL-7B-Instruct model yielded its highest accuracy of 97.32% accuracy with the High-Detail Descriptive Prompt. Similarly, the InternVL3-2B model performed best with 94.00% accuracy using the Basic Plain Text Prompt, and the InternVL3-8B model achieved its top accuracy of 96.03% accuracy with the High-Detail Descriptive Prompt. "TABLE IV" presents a verification of the accuracy based on each model's parameters and prompt format.

TABLE IV
ACCURACY OF MODELS BY PROMPT FORMATS AND PARAMETERS

Model	Params	Basic	Persona	Low Det.
Qwen2.5-VL [6]	3B	**95.90**	95.80	94.03
Qwen2.5-VL [6]	7B	86.62	86.45	95.25
InternVL3 [5]	2B	**94.00**	93.25	88.68
InternVL3 [5]	8B	63.22	62.90	89.43
Model	**Params**	**High Det.**	**YAML**	**JSON**
Qwen2.5-VL [6]	3B	94.75	94.85	95.12
Qwen2.5-VL [6]	7B	**97.32**	95.17	96.12
InternVL3 [5]	2B	91.10	88.35	88.62
InternVL3 [5]	8B	**96.03**	63.53	64.65

These results collectively indicate a trend where simpler prompts tend to yield better performance for models with smaller parameter sizes, whereas more complex and detailed prompts demonstrate superior performance for models with larger parameter sizes. This can be attributed to the ability of larger parameter models to leverage in-context learning from examples or detailed explanations within the prompt, allowing them to comprehend and respond to complex prompts effectively, and to perform stepwise reasoning through Chain-of-Thought (CoT).

V. CONCLUSION

In this study, we proposed a novel Vision-Language Model-based face verification system for unauthorized access detection within restricted areas. This research focuses on exploring the potential of VLMs for face verification using a web-based image dataset. We maximized the system's performance through facial region cropping, cosine similarity-based Fine-Tuning dataset sampling, and prompt formatting optimization strategies. As a result, we achieved a peak face verification accuracy of 97.32%, demonstrating considerable accuracy in determining the identity of two faces with significant feature differences, such as varying facial angles or accessory usage. This approach leverages the powerful feature extraction and inference capabilities of VLMs, showcasing high potential in the field of unauthorized access detection.

This research serves as a preliminary study exploring the potential of VLMs for face verification. While the achieved accuracy is promising, we acknowledge that the evaluation is not yet comprehensive enough to reflect real-world access control scenarios. In future work, we plan to address these limitations. We will conduct further evaluations using standard face verification metrics such as ROC, DET, AUC, and EER to provide a more robust assessment of the system's security and risk management capabilities. We also plan to verify the model's robustness to changes in pose, age, and occlusion using more diverse real-world datasets, such as CPLFW, CALFW, and CFP-FP. To properly contextualize the performance of our VLM-based approach, we will also conduct an objective performance comparison against strong existing face recognition baselines, such as ArcFace, CosFace, and other state-of-the-art models. Through these planned efforts, we aim to validate the competitiveness of VLM-based solutions for real-world security applications.

ACKNOWLEDGMENT

This work was supported by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korean government (MOTIE) (RS-2024-00442120, Development of AI technology capable of robustly recognizing abnormal and dangerous situations and behaviors during night and bad weather conditions.

REFERENCES

[1] Said, Mayur Ishwar, "Detection of Unauthorised Person in a Restricted Place using Deep Learning Algorithms." Diss. Dublin, National College of Ireland, 2023.
[2] Abdalla, Moshira, et al., "Video Anomaly Detection in 10 Years: A Survey and Outlook." arXiv preprint arXiv:2405.19387, 2024.
[3] Luu, Nhan T, "CLIP Unreasonable Potential in Single-Shot Face Recognition." arXiv preprint arXiv:2411.12319, 2024.
[4] Reis, Dillon, et al., "Real-time flying object detection with YOLOv8." arXiv preprint arXiv:2305.09972, 2023.

[5] Zhu, Jinguo, et al., “Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.“ arXiv preprint arXiv:2504.10479, 2025.
[6] Bai, Shuai, et al., “Qwen2. 5-vl technical report.“ arXiv preprint arXiv:2502.13923, 2025.
[7] Hu, Edward J., et al., “Lora: Low-rank adaptation of large language models.“ ICLR 1.2, 2022.
[8] Zhu, Zheng, et al., “Webface260m: A benchmark unveiling the power of million-scale deep face recognition.“ Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition., 2021.
[9] Huang, Gary B., et al., “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.“ Workshop on faces in’Real-Life’Images: detection, alignment, and recognition., 2008.
[10] He, Jia, et al., “Does Prompt Formatting Have Any Impact on LLM Performance?.“ arXiv preprint arXiv:2411.10541, 2024.
[11] Zheng, Yinglin, et al., “General facial representation learning in a visual-linguistic manner.“ Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.
[12] You, Jinghan, et al., “LVFace: Large Vision model for Face Recogniton.” arXiv preprint arXiv:2501.13420, 2025.
[13] He, Ruian, et al., “A Generative Framework for Self-Supervised Facial Representation Learning.” arXiv preprint arXiv:2309.08273, 2023.